IV Conferencia Científica Internacional
Universidad de las Ciencias Informáticas

Temática: Bioinformatics and its applications.

Análisis preliminar de la potencialidad de las diferentes subregiones genómicas de SARS-Cov-2 para su uso como marcadores filogenéticos

Preliminary analysis of the potentiality of the different genomic subregions of SARS-Cov-2 for their use as phylogenetic markers)

Jorge Alejandro Jiménez Garí[1], Camila Castro Martínez[2], Kamila Alejandra Ramaya Soler[3], Antonio De Jesús Oliva Gregorio[4], Pablo Enmanuel Ramos Bermudez[5], Yasniel Yoan Rodriguez Cruz[6], Lienny Morffi Hidalgo[7], Cecilia De La Caridad González González[8], Mario Pupo Meriño[9*]

[1] Departamento de Bioinformática, Universidad de las Ciencias Informáticas. Carretera San Antonio Km 2 ½, Reparto Torrens, La Lisa, La Habana, jorgeajg@estudiantes.uci.cu

[2] Departamento de Bioinformática, Universidad de las Ciencias Informáticas. Carretera San Antonio Km 2 ½, Reparto Torrens, La Lisa, La Habana, camilacm@estudiantes.uci.cu

[3] Departamento de Bioinformática, Universidad de las Ciencias Informáticas. Carretera San Antonio Km 2 ½, Reparto Torrens, La Lisa, La Habana, kamilaars@estudiantes.uci.cu

[4] Departamento de Bioinformática, Universidad de las Ciencias Informáticas. Carretera San Antonio Km 2 ½, Reparto Torrens, La Lisa, La Habana, antoniodjog@estudiantes.uci.cu

[5] Departamento de Bioinformática, Universidad de las Ciencias Informáticas. Carretera San Antonio Km 2 ½, Reparto Torrens, La Lisa, La Habana, pabloerb@estudiantes.uci.cu

[6]Departamento de Bioinformática, Universidad de las Ciencias Informáticas. Carretera San Antonio Km 2 ½, Reparto Torrens, La Lisa, La Habana, yasnielyrc@estudiantes.uci.cu

[7] Departamento de Bioinformática, Universidad de las Ciencias Informáticas. Carretera San Antonio Km 2 ½, Reparto Torrens, La Lisa, La Habana, liennymh@estudiantes.uci.cu

[8] Departamento de Bioinformática, Universidad de las Ciencias Informáticas. Carretera San Antonio Km 2 ½, Reparto Torrens, La Lisa, La Habana, ceciliadlcgg@estudiantes.uci.cu

[9] Departamento de Bioinformática, Universidad de las Ciencias Informáticas. Carretera San Antonio Km 2 ½, Reparto Torrens, La Lisa, La Habana, mpupom@uci.cu

[*] Autor para correspondencia: mpupom@uci.cu

Resumen

El uso de herramientas filogenéticas pudiera ser clave en la toma de decisiones en el manejo de las epidemias. La reconstrucción filogenética requiere de marcadores apropiados, que contengan la información necesaria y permitan reconstruir la historia evolutiva del patógeno. Para el estudio el SARS-Cov-2 a nivel internacional se han secuenciado múltiples genomas completos del virus partiendo de aislados de varios países. Esta información, disponible en bases de datos internacionales, ha facilitado la realización de estudios filogenómicos. En el caso de Cuba, las capacidades tecnológicas no permiten secuenciar genomas completos, lo que obliga a evaluar las diferentes regiones genómicas del SARS-Cov-2 para su potencial uso como fuentes de información. En este trabajo se describe un análisis, realizado a inicios de la pandemia, de las regiones genómicas del SARS-Cov-2, para evaluar su posible uso como marcadores filogenéticos. Para ello se emplearon secuencias y herramientas públicas, teniendo en cuenta su variabilidad, tendencia a la saturación y presencia de ruido, además de evaluar su capacidad para reconstruir las mismas relaciones filogenéticas que las obtenidas con el análisis de todo el genoma. Debido a la relativamente baja tasa evolutiva del virus, y al poco tiempo transcurrido desde el comienzo de la transmisión del SARS-Cov-2 en humanos en el momento del estudio, se observa que la variabilidad en las regiones genómicas individuales no aporta el mismo nivel de información, que el genoma completo, y que la longitud del segmento seleccionado y el muestreo taxonómico son determinantes en la capacidad resolutiva de los métodos filogenéticos empleados.

Palabras clave: SARS-Cov-2, filogenética, señal filogenética, marcador filogenético, saturación

Abstract

The use of phylogenetic tools could be key in making decisions in the management of epidemics. Phylogenetic reconstruction requires appropriate markers, which contain the necessary information and allow reconstructing the evolutionary history of the pathogen. For the study of SARS-Cov-2 at the international level, multiple complete genomes of the virus have been sequenced starting from isolates from several countries. This information, available in international databases, has facilitated phylogenomic studies. In the case of Cuba, technological capabilities do not allow complete genomes to be sequenced, which makes it necessary to evaluate the different SARS-Cov-2 genomic regions for their potential use as sources of information. This work describes an analysis, carried out at the beginning of the pandemic, of the genomic regions of SARS-Cov-2, to evaluate their possible use as phylogenetic markers. For this, sequences and public tools were used, taking into account their variability, tendency to saturation and presence of noise, in addition to evaluating their ability to reconstruct the same phylogenetic relationships as those obtained with the analysis of the entire genome. Due to the relatively low evolutionary rate of the virus, and the short time that has elapsed since the beginning of the transmission of SARS-Cov-2 in humans at the time of the study, it is observed that the variability in the individual genomic regions does not contribute the same level of information, that the complete genome, and that the length of the selected segment and the taxonomic sampling are decisive in the resolution capacity of the phylogenetic methods used.

Keywords: *SARS-Cov-2, phylogenetics, phylogenetic signal, phylogenetic marker, saturation*

Introducción

La epidemiología molecular en enfermedades infecciosas se usa ampliamente para definir la fuente de infección, así como las relaciones ancestrales entre los patógenos aislados de los individuos muestreados de una población. La teoría coalescente y el análisis filogeográfico han tenido aplicación científica en varios eventos recientes de pandemia y brotes nosocomiales (Ciccozzi, Lai et al. 2019). Los datos sobre secuencias aisladas del microorganismo son esenciales para aplicar las herramientas filogenéticas y la investigación en el campo de la filodinámica de las enfermedades infecciosas está creciendo.

La propagación pandémica en curso de un nuevo coronavirus humano, el SARS-COV-2, asociado con la enfermedad de neumonía grave (COVID-19), ha dado como resultado la generación de millones de secuencias de genomas de virus, con una tasa de generación de genomas sin precedentes (Rambaut, Holmes et al. 2020). En la base de datos EpiCov™ asociada a la iniciativa GISAID https://www.gisaid.org/ (Elbe and Buckland-Merrett 2017; Shu and McCauley 2017), y disponible en https://www.epicov.org/ bajo subscripción, se puede encontrar el principal reservorios de secuencias genómicas, así como los metadatos asociados a ellas.

Esta alta disponibilidad de secuencias genómicas ha favorecido la realización de estudios filogenéticos, filodinámicos y filogeográficos (en realidad filogenómicos, al emplear todo el genoma) que han aumentado considerablemente el conocimiento sobre el virus, permitiendo inferir su posible origen, su tasa evolutiva, las vías de distribución geográfica y el establecimiento de los primeros esfuerzos en su clasificación (Lu, Zhao et al. 2020; Lv, Li et al. 2020; Volz, Fu et al. 2020; Zhou, Yang et al. 2020). En un inicio, estos esfuerzos estuvieron encaminados a la clasificación taxonómica del nuevo coronavirus, y varios trabajos, entre ellos los de (Lu, Zhao et al. 2020; Lv, Li et al. 2020), lo ubicaron en un clado del género *Betacoronavirus*, con características más cercanas a los coronavirus de murciélagos que a los presentes en humanos, indicando su origen zoonótico. Posteriormente, el Grupo de Estudio Coronaviridae del Comité Internacional de Taxonomía de Virus, presentó una evaluación de la relación genética del coronavirus humano recientemente identificado con los coronavirus conocidos, y detallaron las bases para (re) nombrar este virus como SARS-Cov-2, ubicándolo en el subgénero *Sarbecovirus* del género *Betacoronavirus* (of the International 2020). La sub- tipificación, por otro lado, es un trabajo en progreso, y se pueden encontrar en la literatura varias propuestas (Lin, Chern et al. 2001; Jia, Yang et al. 2020; Lu, Zhao et al. 2020; Lv, Li et al. 2020; Wang, Hozumi et al. 2020; Yin 2020).

Para comprender la evolución y transmisión del SARS-CoV-2, el genotipado de los aislados de virus es de gran importancia (Yin 2020). A medida que el brote evoluciona con el tiempo y hay más datos disponibles, se han

propuesto diferentes sistemas de clasificación de las variantes (Forster, Forster et al. 2020; Hodcroft 2020; Rambaut, Holmes et al. 2020; Yin 2020). En dependencia de cual definición se seleccione, las cepas circulantes pueden ser clasificadas en diferentes clados según las variantes genéticas. Existen tres sistemas de clasificación ampliamente aceptados en las publicaciones recientes sobre el SARS-Cov-2. Dos de ellos se basan en árboles filogenéticos y uno en la construcción de redes filogenéticas.

De los sistemas de clasificación del SARS-Cov-2 basados en árboles, el de (Rambaut, Holmes et al. 2020) propone un sistema dinámico para etiquetar linajes transitorios que tienen importancia epidemiológica local. Su propuesta da como resultado una gran cantidad de etiquetas de corta duración que mantienen cierta información sobre la estructura jerárquica y facilitan la discusión de la dinámica local a corto plazo, pero no pretenden proporcionar una estructura a gran escala. Este es el sistema de clasificación adoptado por GISAID, donde todas las secuencias introducidas se etiquetan por clados usando la herramienta PANGOLIN (disponible en https://pangolin.cog-uk.io/ para su uso en línea, o como repositorio en https://github.com/hCoV-2019/pangolin/).

El otro enfoque basado en árboles filogenéticos es el que ofrece la plataforma Nextstrain (Hadfield, Megill et al. 2018). Esta plataforma, disponible en https://nextstrain.org/, es un proyecto de código abierto diseñado para aprovechar el potencial científico y de salud pública de los datos del genoma de los patógenos, y ha cobrado una enorme importancia en el estudio del SARS-Cov-2. En ella se proporciona una vista continuamente actualizada de los datos disponibles públicamente con análisis y visualizaciones potentes que muestran la evolución de los patógenos y la propagación de epidemias, y su objetivo es proporcionar una instantánea en tiempo real de las poblaciones de patógenos en evolución y proporcionar visualizaciones interactivas de datos a virólogos, epidemiólogos, funcionarios de salud pública y la comunidad científica en general. En cuanto a la tipificación del SARS-Cov-2, el objetivo de los desarrolladores de Nextstrain es introducir un sistema de nombres que facilite la discusión de los patrones de diversidad a gran escala del SARS-CoV-2 y las etiquetas de clados que persisten durante al menos varios meses y tienen una distribución geográfica significativa (Hodcroft 2020). Para proporcionar nombres memorables y pronunciables, proponen nombrar los clados principales de acuerdo al año en que se estima surgieron y una letra, por ejemplo, 19A, 19B, 20A. Estos son los principales grupos genéticos y no están destinados a resolver completamente la diversidad genética. En el enfoque de (Forster, Forster et al. 2020), por su parte, se construyen redes filogenéticas basadas en caracteres para reconstruir los caminos evolutivos y el genoma ancestral en el huésped humano de un conjunto de secuencias de SARS-Cov-2. De acuerdo a los autores, una aplicación práctica de la red filogenética es reconstruir rutas de infección donde son desconocidas y representan un riesgo para la salud pública. Esta aplicación es

aplicable en general a la reconstrucción filogeográfica y a los métodos de sub- tipificación en cuanto exista asociación entre el subtipo y una región geográfica o una fuente de infección determinada. Un análisis rápido que, si bien hay predominio de determinadas variantes, tanto de las temporales (clasificación GISAID) como de las más estables (clasificación Nextstrain) en determinadas regiones geográficas, y en general estas regiones se distinguen por un patrón específico en la distribución de variantes, no es trivial la asignación de una región geográfica solo partiendo del conocimiento del subtipo. Foster y colaboradores (Forster, Forster et al. 2020) encuentran tres variantes centrales que se distinguen por los cambios en determinados aminoácidos, que denominaron A, B y C, siendo A el tipo ancestral según el coronavirus del grupo de murciélagos tomado como *outgroup* (siguiendo el trabajo inicial de (Zhou, Yang et al. 2020) se tomó Bat CoV RaTG13). En el estudio los tipos A y C se encuentran en proporciones significativas países europeos y en EEUU. En contraste, el tipo B es el tipo más común en el este de Asia, y su genoma ancestral parece no haberse extendido fuera del este de Asia sin mutar primero en los tipos B. En el trabajo se reclama que la red rastrea fielmente las rutas de infecciones para casos documentados de enfermedad por coronavirus, lo que indica que las redes filogenéticas también se pueden usar con éxito para ayudar a rastrear fuentes de infección por COVID-19 no documentadas.

En Cuba, donde se ha seguido un enfoque orientado a la supresión de la epidemia con énfasis en intervenciones no farmacéuticas, cobra vital importancia la epidemiología molecular en la vigilancia epidemiológica. Dado que la interrupción de la cadena de transmisión de COVID-19 mediante la prevención, la vigilancia y la detección rápida de casos son estrategias proactivas comprobadas para mitigar esta amenaza (Gorry 2020), y en este proceso la determinación de las fuentes de infección es fundamental, se hace necesario contar con todas las herramientas posibles. La estrategia cubana se ha caracterizado por la vigilancia activa en fronteras y posterior cierre de las mismas, medidas de distanciamiento social y cierre de localidades con eventos activos de transmisión, cierre de escuelas y centros laborales, suspensión del transporte público, acompañado de la pesquisa puerta por puerta en la búsqueda de personas con síntomas, detección activa de casos en grandes grupos de población utilizando pruebas rápidas y pruebas de reacción en cadena de la polimerasa en tiempo real (RT-PCR) en casos sospechosos, seguimiento y aislamiento de sospechosos y sus contactos, además de poner en función del control de la epidemia a un subsistema de atención primaria de salud que ya estaba funcionando bajo protocolos para evaluaciones continuas de salud comunitaria (Aguilar-Guerra and Reed 2020; Gorry 2020; Reed 2020).

En Cuba, donde se ha seguido un enfoque orientado a la supresión de la epidemia con énfasis en intervenciones no farmacéuticas, cobra vital importancia la epidemiología molecular en la vigilancia epidemiológica. Dado que la

interrupción de la cadena de transmisión de COVID-19 mediante la prevención, la vigilancia y la detección rápida de casos son estrategias proactivas comprobadas para mitigar esta amenaza (Gorry 2020), y en este proceso la determinación de las fuentes de infección es fundamental, se hace necesario contar con todas las herramientas posibles. La estrategia cubana se ha caracterizado por la vigilancia activa en fronteras y posterior cierre de las mismas, medidas de distanciamiento social y cierre de localidades con eventos activos de transmisión, cierre de escuelas y centros laborales, suspensión del transporte público, acompañado de la pesquisa puerta por puerta en la búsqueda de personas con síntomas, detección activa de casos en grandes grupos de población utilizando pruebas rápidas y pruebas de reacción en cadena de la polimerasa en tiempo real (RT-PCR) en casos sospechosos, seguimiento y aislamiento de sospechosos y sus contactos, además de poner en función del control de la epidemia a un subsistema de atención primaria de salud que ya estaba funcionando bajo protocolos para evaluaciones continuas de salud comunitaria (Aguilar-Guerra and Reed 2020; Gorry 2020; Reed 2020).

En cuanto a la selección de un marcador filogenético, inicialmente, con el rápido avance de la tecnología de secuenciación de genes, y la creciente disponibilidad de bases de datos genéticas, seleccionar el mejor o los mejores genes y el conjunto más informativo de taxones se convirtió en uno de los principales problemas en la inferencia filogenética (Nabhan and Sarkar 2012). No todos los genes son marcadores filogenéticos adecuados y no todas las moléculas marcadoras son útiles para el análisis de un grupo dado de organismos (Patwardhan, Ray et al. 2014). Desde hace algunos años la asombrosa tasa de acumulación de conjuntos de datos de secuencias genómicas y metagenómicas ha permitido la realización de estudios filogenéticos a nivel de todo el genoma, no obstante, hay muchas razones para restringir los análisis de datos (Wu, Jospin et al. 2013). Más allá de los problemas específicos que se presentan en el estudio de genomas bacterianos, relacionados con la plasticidad que presenta el genoma en estos organismos, hay problemas en la reconstrucción filogenética que son comunes también a la reconstrucción de genomas virales. El análisis combinado de varios pares de bases aumenta teóricamente la señal filogenética, sin embargo, genera preocupación la posible existencia de errores sistemáticos, que son muy difíciles de detectar al ser errores asociados con la medición misma, sobre todo debidos a suposiciones incorrectas del modelo y que a menudo resultan en árboles filogenéticos inconsistentes (Russo, Aguiar et al. 2017). Las señales filogenéticas en conflicto pueden causar falta de resolución y filogenias incongruentes (Nabhan and Sarkar 2012). Por ejemplo, diferentes genes pueden generar árboles filogenéticos conflictivos y aún mostrar altos valores soporte en análisis combinados, haciendo prácticamente imposible detectar conflictos entre genes individuales confiando en un único análisis combinado. (Russo, Aguiar et al. 2017)

Entre las buenas cualidades de un marcador, está capacidad de ser utilizado para producir árboles filogenéticos robustos que reflejan tanto como sea posible la evolución de las especies de las cuales provienen los genes (Wu, Jospin et al. 2013). Para un mismo gen, las señales filogenéticas en los conjuntos de datos dependen del grupo taxonómico específico en estudio. Teniendo en cuenta este problema, puede ser crucial decidir qué conjunto de marcadores genéticos son adecuados para su inclusión en el análisis filogenético. En un análisis filogenético molecular, diferentes marcadores pueden producir topologías contradictorias para el mismo grupo de diversidad. Cuestiones como la longitud y la tasa de evolución jugarán un papel en la idoneidad de un marcador molecular particular para desarrollar las relaciones filogenéticas para un conjunto dado de taxones combinado (Russo, Aguiar et al. 2017). Tal decisión puede depender de estudios de inferencia filogenética previos que describen el comportamiento de marcadores genéticos particulares con respecto a las señales filogenéticas existentes (Nabhan and Sarkar 2012). La selección de secuencias moleculares de acuerdo a su capacidad de resolver relaciones dentro de un grupo particular incluye estudios que evalúan la capacidad de un gen para recuperar relaciones filogenéticas bien establecidas. La tasa de sustitución debe ser óptima para proporcionar suficientes sitios informativos, pues un gen que evoluciona demasiado rápido puede alcanzar un estado de saturación debido a múltiples sustituciones (Patwardhan, Ray et al. 2014)

En el presente trabajo se realiza un estudio de las subregiones genómicas del SARS-Cov-2 para determinar su posible valor como marcadores filogenéticos. Se sigue como premisas que una subregión del genoma será un buen marcador filogenético para el SARS-Cov-2 si la señal filogenética es fuerte y permite reproducir el historial evolutivo de la misma forma en que lo hacen los estudios con el genoma completo. Con este fin se propone una metodología que parte del análisis descriptivo de la variabilidad de las subregiones del genoma, acompañado de un análisis estadístico, continúa con la evaluación de la señal filogenética a partir de la saturación y el ruido, y termina evaluando la capacidad de cada región de reconstruir filogenias o redes equivalentes a las obtenidas con la información del genoma completo.

Materiales y métodos

Para el presente estudio se tuvieron en cuenta las características de las distintas subregiones del genoma del SARS-Cov-2. El genoma de aproximadamente 29,903 nucleótidos (nt) se organiza de forma similar a los virus del género *Betacoronavirus*, siguiendo a dirección 5 'a 3', se encuentran (Wu, Zhao et al. 2020):

- la replicasa ORF1ab (21,291 nt, su producto es una poliproteína que tiene como productos de escisión predichos 16 proteínas no estructurales:
    - nsp1 (proteína líder), nsp2, ADRP, adenosina difosfato-ribosa 1'-fosfatasa; nsp4, nsp6 al nsp11, 3CLpro, cisteína proteinasa de tipo 3C; RdRp, ARN polimerasa dependiente de ARN; Hel, helicasa; ExoN, exonucleasa de 3' a 5'; NendoU, endoribonucleasa nidoviral específica para U (endoRNAsa); OMT, ribosa 2'-O-metiltransferasa dependiente de S-adenosilmetionina),
- S (que codifica la glucoproteína de espiga estructural),
- ORF3a (proteína de ORF3a),
- E (proteína estructural de la envoltura),
- M (glucoproteína estructural de membrana),
- ORF6 (proteína ORF6),
- ORF7a (proteína ORF7a),
- ORF7b (proteína ORF7b),
- ORF8 (proteína ORF8),
- ORF9 (incluye N, fosfoproteína estructural de la nucleocápside) y
- ORF10 (proteína ORF10).

Adicionalmente, tiene secuencias terminales 5 'y 3' que son típicas de los betacoronavirus, con 265 nt en el extremo 5' terminal y 229 nt en el extremo terminal 3'.

Con el objetivo de evaluar su potencial uso como marcadores filogenéticos, se siguió una metodología que se puede resumir en los siguientes pasos:

1. Selección de las subregiones de mayor variabilidad a partir de:
    a. la presencia de varias posiciones con polimorfismos,
    b. la existencia de diferencias entre áreas geográficas en los patrones de variabilidad en dichas subregiones.
2. Evaluación de la señal filogenética en las subregiones seleccionadas, en comparación con la señal filogenética de todo el genoma, teniendo en cuenta:
    a. la existencia de saturación,
    b. la existencia de ruido

3. Evaluación de la capacidad relativa de las subregiones seleccionadas para reproducir las relaciones evolutivas, con respecto a las construidas con la información de todo el genoma. En este caso, se considera:
    a. la capacidad de reconstruir las filogenias, a partir de:
        i. la comparación estructural de las topologías obtenidas con la subregión y con todo el genoma,
        ii. la comparación probabilística de las topologías obtenidas con la subregión y con todo el genoma.

El estudio inicial de la variabilidad por regiones se realizó mediante la herramienta para el análisis de variabilidad disponible en la plataforma Nextstrain (https://nextstrain.org/), en la cual la variabilidad es calculada en función de todas las secuencias genómicas disponibles. La inspección visual de la variabilidad por regiones del genoma se utilizó para realizar un filtrado inicial de las regiones del genoma, partiendo del criterio que un buen marcador filogenético debe ser suficientemente variable como para ser informativo. Este criterio en sí mismo es ambiguo, pero se matiza en el contexto de trabajo en particular: por ejemplo, si el marcador se pretende usar para determinar el origen geográfico de las secuencias, debe tener un patrón de variabilidad que permita discriminar composiciones entre las áreas geográficas en estudio. Esta situación es similar a la que se enfrenta en los sistemas de genotipificación. Si se observa la *Figura. 1* se puede observar que, aunque la comparación no es trivial, se deben esperar diferencias entre los patrones de variantes entre distintas áreas geográficas. Complementando este criterio, se realizó la comparación por la prueba estadística de Friedman (Friedman 1940) seguida de la comparación por pares usando Wilcoxon (Wilcoxon 1992) corregido por el criterio de Holm (Holm 1979) para comparaciones múltiples, de la entropías por sitio de todo el genoma entre el comportamiento global y el de seis áreas geográficas: Asia, Europa, América del Norte, América del Sur, África y Oceanía. El mismo análisis se realizó para cada región del genoma en particular, con el objetivo de evaluar si el patrón de comportamiento es el mismo que a nivel de genoma. Estos análisis fueron realizados en el paquete estadístico R (Core 2017). Con el propósito de ofrecer contexto a los resultados anteriores, se incluye una evaluación de la asociación existente entre las áreas geográficas y las diferentes clasificaciones taxonómicas de las secuencias, para lo cual se evaluó la capacidad de las clasificaciones de Nextstrain y GISAID, tomadas como variables independientes, de predecir el área geográfica de origen, mediante una derivación del método CHAID (Biggs, De Ville et al. 1991), el CHAID exhaustivo, implementado en el programa IBM SPSS v20 (IBM). Estos procesamientos se realizaron empleando los metadatos disponibles para su descarga en:

- https://nextstrain.org/ncov/global?c=region (datos globales),

- https://nextstrain.org/ncov/north-america?c=division&f_region=North%20America&r=division (Norteamérica),
- https://nextstrain.org/ncov/europe?f_region=Europe (Europa),
- https://nextstrain.org/ncov/asia?f_region=Asia (Asia),
- https://nextstrain.org/ncov/south-america?f_region=South%20America (Sudamérica),
- https://nextstrain.org/ncov/africa?f_region=Africa (África) y
- https://nextstrain.org/ncov/oceania?c=division&f_region=Oceania&r=division (Oceanía).

Para los estudios filogenéticos se trabajó con las mismas secuencias empleadas en (Forster, Forster et al. 2020), cuya tabla de agradecimientos (con los códigos de acceso a GISAID, y datos de origen y autores) aparece en el Material Suplementario D. Se empleó este conjunto de datos porque permite evaluar la capacidad de la inferencia filogenética para reconstruir rutas de infección y se toman casos de estudio para los que existía un historial de viajes conocido.

Los genomas completos se alinearon utilizando la herramienta progressiveMauve (Darling, Mau et al. 2010). Se tomaron como potenciales marcadores las regiones codificantes del genoma del SARS-Cov-2. Las diferentes regiones se obtuvieron de las secuencias genómicas utilizando la anotación asociada al ensamblaje GCF_009858895.2 del NCBI, tomando como referencia la secuencia con código NC_045512.2 del GenBank correspondiente al aislado Wuhan-Hu-1, mediante el programa BLAT v 36x5 (Kent 2002). Los conjuntos de secuencias para cada región se alinearon con el programa MAFFT v7 (Katoh and Standley 2013), empleando el marco de trabajo de herramientas de análisis bioinformático del EMBL-EBI (Goujon, McWilliam et al. 2010). En todos los casos los alineamientos se procesaron con el software Gblocks (Castresana 2000) versión 9b, para eliminar las regiones pobremente alineadas que pudieran afectar la calidad de los análisis.

Los efectos de saturación se investigaron trazando el número absoluto de transiciones y transversiones versus distancia genética para todos los CV-A24v, utilizando el software DAMBE v7.2.43 (Xia and Xie 2001) y la prueba de Xia et al. 2003 (Xia, Xie et al. 2003) disponible en el mismo software, en ambos casos usando el modelo evolutivo de Felsenstein (F84) (Felsenstein and Churchill 1996) para el cálculo de las distancias genéticas. Se realizó el estudio en las posiciones 1 + 2 + 3, teniendo en cuenta que, de aparecer saturación, se evaluarían las posiciones 1+2 y 3. El ruido en la señal se evaluó mediante el método de mapeo de verosimilitud (Strimmer and von Haeseler 1996), implementado en el programa Tree-Puzzle v 5.3 (Schmidt, Strimmer et al. 2002). Antes de cada procesamiento, las secuencias idénticas fueron eliminadas del conjunto de datos, empleando el propio programa DAMBE v7.2.43.

Los estudios filogenéticos preliminares se realizaron utilizando el método de máxima verosimilitud (ML) mediante el programa IQ-TREE (Nguyen, Schmidt et al. 2015), mediante el servidor web W-IQ-TREE (Trifinopoulos, Nguyen et al. 2016) disponible en http://iqtree.cibiv.univie.ac.at/. Para cada conjunto de taxones se realizó la reconstrucción filogenética para el genoma completo y para cada región genómica. Posteriormente se emplearon los métodos KS (Kishino, Miyata et al. 1990), SH (Shimodaira and Hasegawa 1999) y AU (Shimodaira 2002) para comparar el soporte de las topologías obtenidas por cada región respecto a la obtenida con todo el genoma, utilizando 10000 réplicas. Los árboles se visualizaron y editaron empleando el programa FigTree v 1.4.3 (Rambaut 2017).

Adicionalmente se realizó la comparación de las topologías obtenidas mediante el programa ETE Toolkit versión 3 (Huerta-Cepas, Serra et al. 2016), para determinar el nivel de similitud entre las topologías obtenidas por regiones y la obtenida con todo el genoma. Se compararon los árboles máximo verosímiles obtenidos por IQ-TREE, empleando la métrica RF (Robinson and Foulds 1981).

Resultados y discusión

El primer paso para la selección de los posibles marcadores filogenéticos en la secuencia de SARS-Cov-2 consistió en el análisis visual de la variabilidad por regiones. En la Figura 1 se la variabilidad por regiones obtenida en la plataforma Nexstrain.

En la figura se observa que, si bien los eventos de mutaciones comprenden toda la extensión del genoma, en general se caracteriza por pocos eventos de mutación (pocos eventos y baja entropía) en la mayoría de las regiones, y solo hay algunos puntos calientes (*hot spots*), en las regiones del ORF1 (a y b) , en la región que codifica a la espiga (S), el ORF3a, ORF8, el gen N de la fosfoproteína de la nucleocápside (ORF9), y el ORF14. El ORF14, al no estar caracterizado, no está incluido en la tabla de anotación del genoma empleada para el estudio, lo cual dificultaba el procesamiento automático de las secuencias, por lo que no fue considerado para el estudio. En el Material Suplementario A se muestran los gráficos ampliados para la entropía de todas las subregiones.

Estos puntos calientes de mutación han sido reportados en varios estudios, incluyendo el de (Pachetti, Marini et al. 2020), en el que se determinó el comportamiento mutacional por áreas geográficas del virus, haciendo especial énfasis en la ARN polimerasa dependiente de ARN (RdRp), una proteína no estructurales producidas como productos de escisión de la poliproteína ORF1b. En (Velazquez-Salinas, Zarate et al. 2020) se reportan sitios de selección positiva en el ORF3a y el ORF8.Es de esperarse que, de existir un buen marcador filogenético, se encuentre en una de estas regiones.

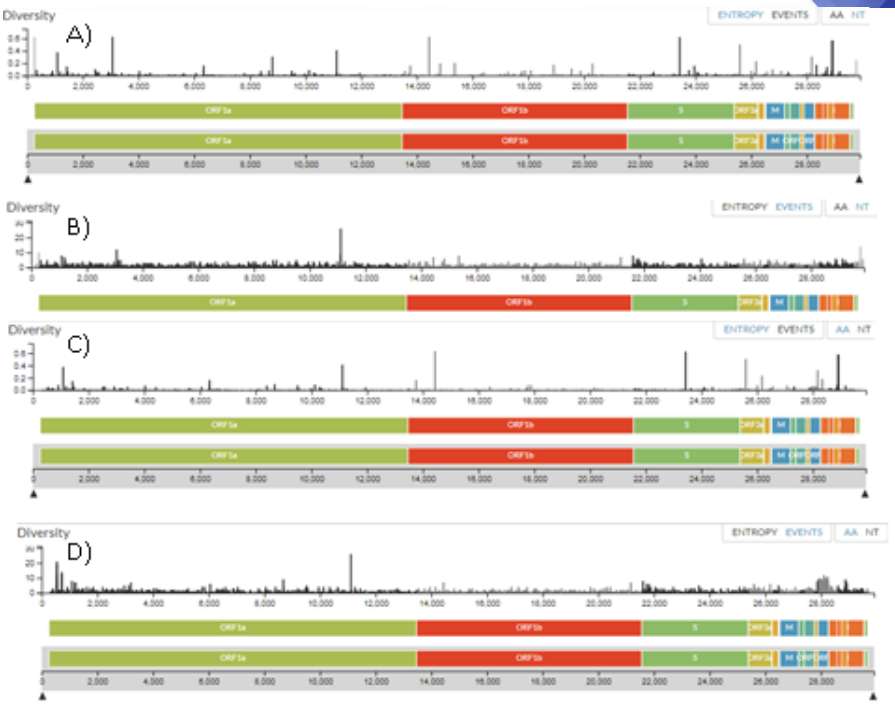


Figura. 2 Gráficos de la variabilidad por regiones en el genoma del SARS-Cov-2 disponible en la plataforma Nexstrain. Se visualiza la diversidad calculada como entropía A) y C) o como cantidad de eventos de mutación B) y D), en función de la composición nucleotídica A) y B) y en función de la composición aminoacídica C) y D). Obtenido a partir de un muestreo de 3092 secuencias obtenidas entre diciembre de 2019 y junio de 2020 https://nextstrain.org/ncov/gisaid/global?c=region&dmax=2020-07-01 .

Esta situación también se ve matizada por la relativamente baja tasa de mutación del SARS-Cov-2 (para un virus de ARN). Entre las herramientas que brinda la plataforma Nextstrain es el estimado de la tasa de variación del virus, que hasta la fecha se estima en 24.884 (esta tasa se actualiza por días) mutaciones por año (ver *Figura. 3*). Si la diversidad genética del SARS-CoV-2 es baja durante el período pandémico temprano, habrá una asociación directa entre la asignación de linaje y la presencia de conjuntos particulares de mutaciones (con respecto a la secuencia raíz) (Rambaut, Holmes et al. 2020). La falta de diversidad puede limitar el poder de resolución de los métodos filogenéticos para asignar un origen ancestral, al reducir la cantidad de sitios filogenéticamente informativos, pero al mismo tiempo hace más relevantes las mutaciones específicas que puedan indicar una ruta evolutiva determinada. El comportamiento en la variabilidad de las secuencias de SARS-Cov-2 en el curso actual de la pandemia, correspondiente a la fase de expansión, implica unas condiciones cercanas a las estipuladas por el modelo de

mutaciones infinitas de Kimura (Kimura 1969), en el cual se asume que hay un número infinito de sitios donde pueden ocurrir mutaciones y cada nueva mutación ocurre en un sitio nuevo (unido a la no ocurrencia de recombinación). La tasa estimada de mutaciones por año en conjunto con la longitud del genoma, cercana a las 30 Kb, hace poco probables las mutaciones en la misma posición o en posiciones cercanas del genoma. La asignación de linajes se volverá más compleja, pero aún manejable, a medida que se acumule la diversidad genética del SARS-CoV-2, lo que aumenta las posibilidades de homoplasias y mutaciones inversas (Rambaut, Holmes et al. 2020).

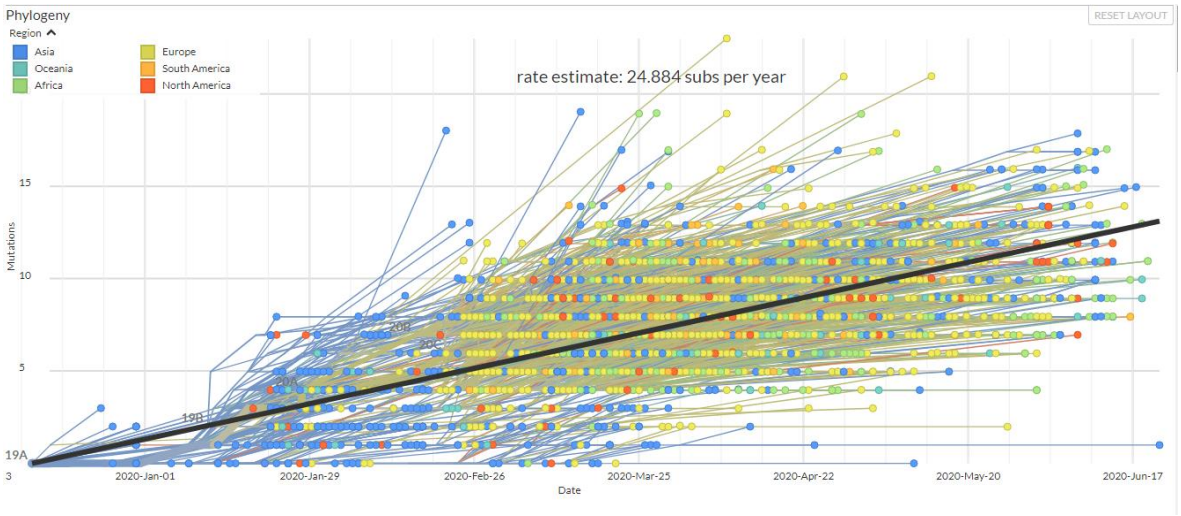


Figura. 3.Representación de la cantidad de mutaciones en función del tiempo de SARS-Cov-2, utilizada para la estimación de la tasa de mutación de acuerdo a la hipótesis del reloj molecular, tomado de https://nextstrain.org/ncov/gisaid/global?c=region&dmax=2020-07-01&l=clock . El árbol se obtuvo de un muestreo de 3092 secuencias muestreadas entre diciembre de 2019 y junio de 2020.

El poder discriminativo (por ejemplo, para determinar el origen geográfico de una secuencia) de la información contenida en el genoma, o en una subregión determinada, estará influenciado por las diferencias en los patrones de variación y en los rasgos característicos, entre grupos de secuencias (secuencias agrupadas por áreas geográficas u otro tipo de agrupamiento). Para determinar si existen diferencias significativas en los patrones de variabilidad entre las diferentes áreas geográficas, tanto a nivel de genoma como a nivel de segmentos, se realizaron las comparaciones estadísticas correspondientes. En el Material Suplementario B se muestran las salidas de las pruebas estadísticas. En todos los casos la prueba de Friedman arroja la existencia de diferencias estadísticamente significativas globales (no confundir con el término global utilizado en los datos para definir las variaciones obtenidas al considerar todas las secuencias en la base de datos, sin separarlas por áreas geográficas) en las entropías por posiciones de secuencias. Es

válido hacer notar que los valores de las medianas de la entropía resultan bajos, y en muchos casos su valor es cero, lo cual tiene impacto en las pruebas estadísticas.

En las comparaciones por pares, cuando se emplea la información de todo el genoma encuentran diferencias estadísticamente significativas entre todas las áreas geográficas, salvo entre Sudamérica y Oceanía, y entre todas estas y el valor global. La no existencia de diferencias entre Sudamérica y Oceanía se pueden explicar por el submuestreo en estas regiones, con lo cual la variabilidad puede estar subestimada, tendiendo a cero en la mayor parte de las posiciones. Esto se hace más evidente cuando se comparan las subregiones del genoma, pues muchas comparaciones no se pueden hacer por tener demasiados ceros en los datos de ambas regiones.

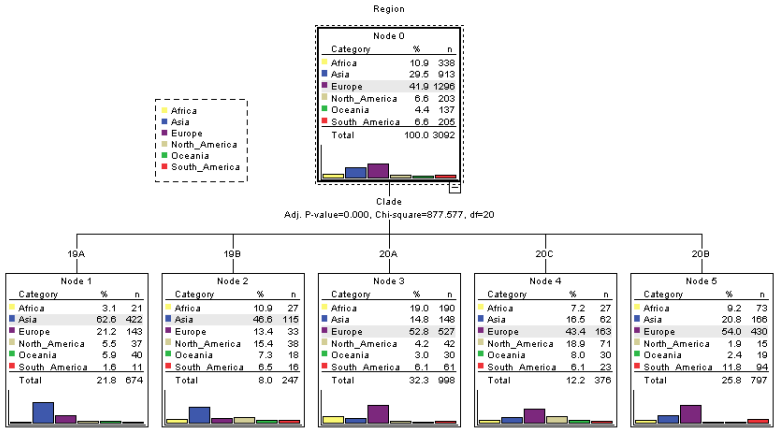
En las comparaciones por subregiones del genoma se puede ver, entre los resultados más notables, que en el ORF1a muestra un comportamiento similar al del genoma completo, no así el resto de las regiones. En el ORF1b, Oceanía, África y Sudamérica no se pudieron comparar por exceso de ceros en los datos. Con respecto a S, sucede lo mismo con Oceanía, África y Sudamérica y adicionalmente no se encuentran diferencias entre Norteamérica y Asia. Todo lo anterior ocurre también con el ORF3a, E, M, ORF7a y ORF8. Adicionalmente, tampoco se observan diferencias en estos últimos entre Norteamérica y Asia con respecto a Europa, comportamiento que se repite en el ORF6 y el ORF7b, con el detalle adicional de que Norteamérica no presenta diferencias ni con el global ni con Sudamérica y Oceanía. En el ORF7b tampoco existen diferencias entre Asia y África. En la subregión correspondiente a N, por su parte, además de la imposibilidad de comparar África, Sudamérica y Oceanía, no se obtienen diferencias entre estas dos últimas y Norteamérica, y esta a su vez no difiere de Asia. En el ORF10 casi todas las regiones muestran una mediana igual a cero en la entropía, por lo que solo se pudo realizar la comparación del global contra las regiones, dando diferencias estadísticamente significativas.

Con respecto a estos resultados hay dos aclaraciones importantes que hacer. Una es que se están comparando los valores numéricos de variabilidad, por lo tanto, que dos áreas no tengan diferencias solo implica que las posiciones del genoma tienden a variar con la misma frecuencia, no que tengan el mismo patrón cualitativo de mutaciones. La otra se relaciona con un sesgo evidente que tiene el estudio que se presenta. El hecho de que el ORF1a presente un comportamiento similar al del genoma completo, puede estar asociado al hecho de haberlo tomado en toda su extensión (alrededor de 13 Kb, cerca del 43% la longitud del genoma) y no en las subsecuencias representativas de los productos de escisión. Este sesgo se observará, por supuesto, en los análisis subsiguientes.

Siguiendo esta línea de pensamiento, se evaluó si existen patrones de mutación específicos de regiones. Para ello, se evaluó la capacidad predictiva de las clasificaciones Nextsatrain y GISAID de las variantes de SARS-Cov-2, como

variables predictivas del área geográfica. En la Figura. 4 se puede observar el resultado del procedimiento de clasificación, donde se evidencia que, en primer lugar, el sistema de clasificación ofrecido por Nextstrain es el más infuyente en la predicción de del área geográfica (es la única variable que aparece en el árbol resultante). Incluso esta variable es un mal predictor del área geográfica, pues solo clasifica las secuencias en dos categorías: Asia y África. El porciento de buena clasificación general es cercano al 50 %, igual que al clasificar las secuencias asiáticas, mientras que las secuencias europeas las identifica con un 86 % de exactitud. El análisis está sesgado por el desbalance de los datos, al estar sobrerrepresentadas las secuencias europeas (41.9 %). Este resultado indica, de todos modos, que la simple asignación de un subtipo a una secuencia, no implica la asignación automática de un área geográfica de procedencia. El estudio solo se realiza para ofrecer un contexto al resultado anterior, y es válido a nivel de genoma, pues las clasificaciones taxonómicas empleadas se definien a este nivel. No obstante, establece una cota para la capacidad predictiva de las clasificaciones taxonómicas obtenidas con marcadores filogenéticos, pues es de esperarse que a lo sumo estas serían capaces de reproducir lo que se observa a nivel de genoma. Esto, unido al resultado obtenido con la variabilidad, indica que la deducción del área geográfica de procedencia de una nueva variante aislada, o de una nueva cepa aislada en una localidad cuya procedencia se quiera determinar, no será una tarea trivial y va a requerir de una análisis filogenético y filogeográfico más profundo, el cual puede también estar limitado por la poca diversidad observada en el virus hasta la fecha.

Para evaluar el potencial de una subregión genómica como marcador filogenético, hay que establecer un balance adecuado entre la variabilidad (y el poder de discriminación que ofrece) y la posibilidad de existencia de saturación (demasiada variabilidad deriva en pérdida de poder resolutivo). Estas características configuran la señal filogenética. Con la tasa de mutación estimada para SARS-Cov-2, y el tiempo que ha transcurrido desde el inicio del primer brote, no es de esperarse la existencia de saturación, pero sí de ruido en los alineamientos debido a la acumulación de pocas mutaciones, distribuidas a lo largo de todo el genoma.



Classification

Observed	Predicted						
	Africa	Asia	Europe	North_America	Oceania	South_America	Percent Correct
Africa	0	48	290	0	0	0	0.0%
Asia	0	537	376	0	0	0	58.8%
Europe	0	176	1120	0	0	0	86.4%
North_America	0	75	128	0	0	0	0.0%
Oceania	0	58	79	0	0	0	0.0%
South_America	0	27	178	0	0	0	0.0%
Overall Percentage	0.0%	29.8%	70.2%	0.0%	0.0%	0.0%	53.6%

Growing Method: EXHAUSTIVE CHAID
Dependent Variable: Region

Figura. 4.Resultados de procedimiento de clasificación mediante árboles por el algoritmo CHAID, realizado en el SPSS v 20. Como variables predictivas se emplearon las clasificaciones Nextstrain (la única reflejada en el árbol al ser la más influyente) y la clasificación GISAID. Se emplearon los metadatos de 3092 secuencias muestreadas en GISAID.

Antes de realizar el estudio de la señal filogenética, se realizó un preprocesamiento de los alineamientos usando el Gblocks para eliminar las regiones pobremente alineadas y Dambe para eliminar las secuencias repetidas. Las características de los datos resultantes se muestran en la *Tabla 1*. Se puede apreciar una notable reducción de la cantidad de secuencias no idénticas en la medida en la que se reduce la longitud de la región del genoma analizada. Los datos empleados en el trabajo de (Forster, Forster et al. 2020), además de incluir una secuencia de coronavirus de murciéago, RaTG13, muestreada en el 2013, cubren solo el período de diciembre de 2019 a febrero de 2020, cuando el virus apenas se estaba expandiendo por el mundo, y se había acumulado muy poca diversidad. Estas características afectan, por supuesto, cualquier generalización que se presente basada en este conjunto de datos. En lo que sigue, salvo para la evaluación de la saturación, se mantuvieron todas las secuencias para observar la influencia real del muestreo taxonómico en los resultados. Para constrastar algunos resultados, se repitieron con un conjunto de 51 secuencias provenientes de Bélgica (Material Suplementario H).

En el Material Suplementario C se muestran los resultados del análisis de saturación realizado con Dambe, así como de la evaluación de ruido por mapeo de verosimilitud realizado en Tree-Puzzle, en el conjunto de datos correspondiente al trabajo de (Forster, Forster et al. 2020). Como era de esperarse, no se observa saturación, y el test de Xia (Xia, Xie et al. 2003) lo confirma en todos los casos (solo fue necesario evaluar el esquema 1+2+3).

El ruido, por su parte, tiene un comportamiento notablemente deficiente. Como se anunciaba anteriormente, con tantas secuencias similares o idénticas, es de esperar que la resolución de las filogenias no sea elevada. Según se describe en (Schmidt and von Haeseler 2009), desde un punto de vista biológico, un análisis de mapeo de verosimilitud que muestra más del 20% –30% de los puntos en el área correspondiente a topologías de tipo estrella o red sugiere que los datos no son confiables para la inferencia filogenética (al menos para la obtención de árboles). En los resultados del anexo se puede observar que incluso para el genoma completo, el 50.9 % de los puntos cumplen con este criterio (el 48.3 % lo hace en la región totalmente no resuelta, tipo estrella), lo cual hace muy poco resolutivas la inferencia filogenética, lo cual es un factor de peso a tener en cuenta. En el Material Suplementario H (**¡Error! No se encuentra el origen de la referencia.**) se puede observar el resultado del mismo análisis para el conjunto de datos proveniente de Bélgica. Se puede observar que, incluso en este caso, la capacidad resolutiva se ve afectada, con un 28.5 % de los puntos en las regiones de no resolución, y predominio en la región central con un 27.4 %.

Aun teniendo lo anterior en cuenta, se evalúa a capacidad de cada subregión de reconstruir una filogenia similar a la que se obtiene cuando se emplea el genoma completo, o al menos una con un nivel de soporte similar. La variante del soporte se realizó a partir de la comparación de las verosimilitudes de los árboles máximo verosímiles obtenidos con cada subregión con la del árbol obtenido a partir de todo el genoma, empleando las pruebas KH, SH y AU implementadas en IQ-TREE.

En la Figura. 5 se observa la salida con los valores de significación para cada uno de estas pruebas. Se puede observar que de acuerdo a las pruebas KH y AU, solo el ORF1a permite obtener árboles cuya verosimilitud indica que son igualmente probables que el árbol obtenido con el genoma completo. Para la prueba SH ocurre lo mismo con el ORF3a y el ORF8, pero hay que tener en cuenta que esta prueba tiende a ser muy conservadora (Shimodaira 2002), por lo tanto el resultado más confiable es el aportado por la prueba AU. Los segmentos M y ORF6 no tienen sitios parsimoniosamente informativos, por lo que no se pudo realizar la comparación con ellos. En el Material Suplementario D (**¡Error! No se encuentra el origen de la referencia.**) se puede observar que, en los datos provenientes de Bélgica, además del ORF1a, también el árbol construido con la subregión de la espiga muestra un

soporte comparable al del genoma completo. En este caso solo se consideraron las subregiones genómicas ORF1a, S, ORF3a, ORF8 y N, para mostrar el comportamiento en segmentos de varias longitudes.

Tabla 1.Resultados del preprocesamiento de las secuencias en estudio

Región	**Longitud de la región (pb)**	**Cantidad de Secuencias Inicial**	**Cantidad de secuencias no idénticas luego del preprocesamiento**
Genoma	29732	160	105
ORF1a	13218	160	54
ORF1b	8088	160	37
S	3822	160	23
ORF3a	828	160	13
E	228	160	6
M	669	160	5
ORF6	186	160	2
ORF7a	366	160	5
ORF7b	132	160	2
ORF8	366	160	8
N	1260	160	16
ORF10	117	160	5

La comparación de topologías se realizó con ETE Tolkit, comparando las topologías de los árboles máximo verosímiles obtenidos en IQ-TREE. La Figura. 6 muestra la salida del procedimiento ete-compare. En la figura:

- nRF: distancia normalizada de Robinson-Foulds (RF / maxRF)
- RF: Distancia simétrica de Robinson-Foulds
- maxRF: valor máximo de Robinson-Foulds para esta comparación
- % src_br: frecuencia de ramas en el árbol de destino encontrado en el árbol de referencia (1.00 → se encuentra el 100% de las ramas)
- % ref_br: frecuencia de ramas en el árbol de referencia encontrado en el árbol de destino (1.00 → se encuentra el 100% de las ramas)

- Número de subárboles utilizados para la comparación (se aplica solo cuando se utilizan elementos duplicados para descomponer árboles objetivo)

```
Tree            logL    deltaL  bp-RELL    p-KH     p-SH        p-AU
--------------------------------------------------------------------
  ORF1a  -48995.41223       0   0.991 +  0.968 +      1 +     0.999 +
  ORF1b  -49802.61581   807.2  0.0053 -  0.0324 - 0.0717 +   0.00498 -
  S      -49893.67597  898.26  0.0018 -  0.0205 - 0.0429 -     0.004 -
  ORF3a  -50265.80098  1270.4  0.0018 -  0.0131 - 0.0176 -   0.00393 -
  E      -50597.76995  1602.4       0 -  0.0022 - 0.0033 - 5.77e-164 -
  ORF7a  -50364.91713  1369.5       0 -  0.0053 - 0.0075 -  6.57e-08 -
  ORF7b  -50549.19864  1553.8       0 -  0.0024 - 0.0034 -   3.3e-18 -
  ORF8   -49852.32591  856.91       0 -  0.0029 - 0.0966 +  0.000476 -
  N      -49897.40009  901.99  0.0002 -  0.0163 - 0.0436 -   0.00176 -
  ORF10  -61526.69716   12531       0 -       0 -      0 -  1.28e-55 -

deltaL  : logL difference from the maximal logl in the set.
bp-RELL : bootstrap proportion using RELL method.
p-KH    : p-value of one sided Kishino-Hasegawa test.
p-SH    : p-value of Shimodaira-Hasegawa test.
p-AU    : p-value of approximately unbiased (AU) test.

Plus signs denote the 95% confidence sets.
Minus signs denote significant exclusion.
All tests performed 10000 resamplings using the RELL method.
```

Figura. 5.Salida del IQ-TREE para el procedimiento de comparación de topologías, realizado con la información del genoma completo y los respectivos árboles construidos con las subregiones genómicas, partiendo de las secuencias empleadas en el trabajo de (Forster, Forster et al. 2020).

En la figura se puede observar, en concordancia con lo analizado anteriormente, que la topología menos distante (más similar) a la obtenida con el genoma completo, se obtiene con el ORF1a. No obstante, es notable que solo comparten el 56% de las ramas, por lo que existen diferencias considerables entre las topologías. Este valor no es muy diferente del obtenido con las otras subregiones. El segmento S le sigue en similitud a la topología del árbol construido con el genoma completo, y le siguen N y ORF1b. En el Material Suplementario D (**¡Error! No se encuentra el origen de la referencia.**) se observa que con los datos de Bélgica ocurre algo similar. La **¡Error! No se encuentra el origen de la referencia.** muestra las topologías d ellos árboles máximo verosímiles obtenidos con IQ-TREE para las secuencias belgas. Se observa que en la inferencia realizada a nivel de genoma el árbol está mucho más resuelto que los que se obtienen en la medida en que se asciende en el valor de distancia dado por RF, según se reporta en la tabla de la **¡Error! No se encuentra el origen de la referencia.**. En general se puede ver que para la reconstrucción de filogenias partiendo de datos de secuencias del SARS-Cov-2 en su estado actual de evolución, tanto en lo referido a la topología como al soporte, la fortaleza de la señal filogenética, y por lo tanto la probabilidad de reconstruir el historial evolutivo de las secuencias; está estrechamente ligada a la longitud del segmento seleccionado.

source	ref	eff.size	nRF	RF	maxRF	%src_br+	%ref_br+	subtrees
ORF1a	Genome	159	0.89	278.00	312.00	0.56	0.56	1
ORF1b	Genome	159	0.96	298.00	312.00	0.53	0.53	1
S	Genome	159	0.94	294.00	312.00	0.53	0.53	1
ORF3a	Genome	159	0.99	310.00	312.00	0.51	0.51	1
E	Genome	159	0.99	310.00	312.00	0.51	0.51	1
M	Genome	159	1.00	312.00	312.00	0.50	0.50	1
ORF7a	Genome	159	1.00	312.00	312.00	0.50	0.50	1
ORF7b	Genome	159	0.98	306.00	312.00	0.51	0.51	1
ORF8	Genome	159	0.98	306.00	312.00	0.51	0.51	1
N	Genome	159	0.96	298.00	312.00	0.53	0.53	1
ORF10	Genome	159	1.00	312.00	312.00	0.50	0.50	1

Figura. 6 Comparación entre las topologías de los árboles obtenidos con subregiones genómicas con respecto al obtenido con el genoma completo, realizada con el software ETE Tolkit

En la Figura. 6 se muestran los árboles máximo verosímiles obtenidos en IQ-TREE para los mismos datos, realizados empleando todo el genoma, el ORF1a y el segmento N. Sobre la capacidad de la red filogenética para reconstruir rutas de infección donde son desconocidas y representan un riesgo para la salud pública, en el trabajo de (Forster, Forster et al. 2020) se toman entre otros, como casos de estudio, los primeros aislamientos realizados en Brasil (Brazil/SPBR-02/2020|EPI_ISL_413016|2020-02-28) y en México (Mexico/CDMX-InDRE_01/2020|EPI_ISL_412972|2020-02-27, (Garcés-Ayala, Araiza-Rodríguez et al. 2020)). Para ambos casos existía un historial de viajes conocidos, y en el caso de la secuencia mexicana (Forster, Forster et al. 2020) afirman que se puede trazar la cadena de contagios a partir de información epidemiológica conocida hasta el foco de infección en Wuhan. Se observa la dependencia con respecto a la fortaleza de la señal (hay baja resolución de las relaciones filogenéticas) y a un adecuado muestreo de taxones (que en definitiva es lo que permite, aún en el marcador menos informativo, el de la subregión N, todavía poder inferir un origen geográfico de la secuencia mexicana). En las figuras **¡Error! No se encuentra el origen de la referencia.** y **¡Error! No se encuentra el origen de la referencia.** se muestra la inferencia filogeográfica realizada en la plataforma Nexstrain, empleando la reconstrucción de estados en el contexto de la inferencia por máxima verosimilitud, y empleando un muestreo voluminoso de las secuencias existentes en GISAID. Para la secuencia brasileña, que pertenece al subtipo 20A de Nextstrain (y al V de GISAID) se infiere un origen europeo, aunque no italiano, como lo indica el conocimiento epidemiológico previo. Lo mismo ocurre con la secuencia mexicana (20B de acuerdo a Nextstrain, GR de acuerdo a GISAID). Este resultado muestra, que incluso con un muestreo taxonómico mayor, si no existe específicamente un muestreo en posiciones cercanas a la ruta de transmisión, la precisión de la inferencia se reduce.

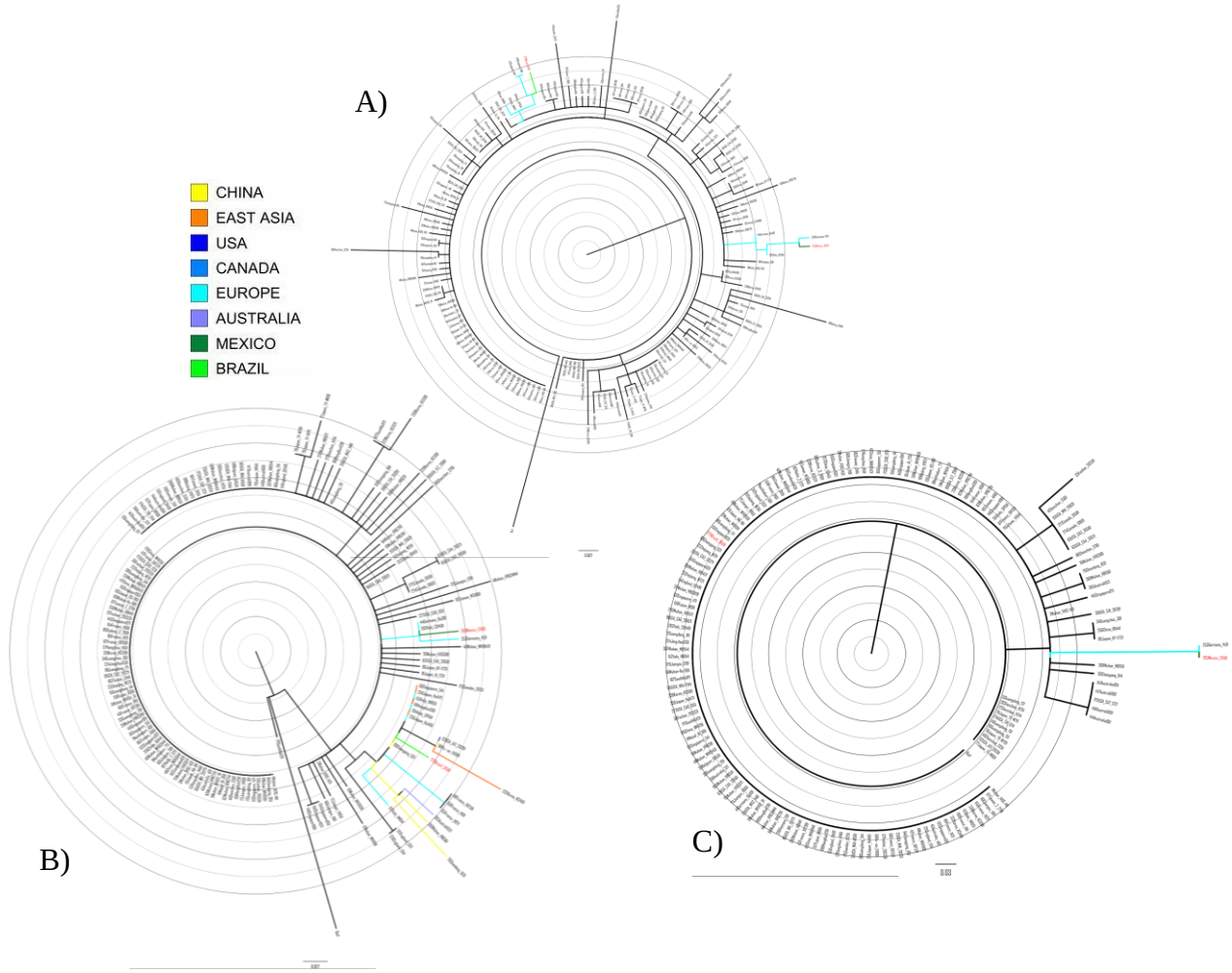


Figura. 6. Árboles máximo verosímiles obtenidos en IQ-TREE para el conjunto de secuencias empleado en (Forster, Forster et al. 2020) a partir de A) todo el genoma, B) la subregión ORF1a y C) la subregión N.

Este resultado, en primer lugar, señala a la importancia de un adecuado muestreo taxonómico, tal y como se explica en (Nabhan and Sarkar 2012), pues resulta evidente que la mejor resolución con respecto a la secuencia mexicana, estuvo estrechamente ligada a un muestreo conveniente de secuencias que formaban parte de la ruta de transmisión, o eran cercanas a las que lo hacían. Esto, combinado con la baja variabilidad acumulada en el virus hasta el momento de seleccionarse la muestra, contribuyó a una buena deducción de la ruta, por la aparición de mutaciones características de la misma.

Conclusiones

La evidencia que se presenta en este trabajo, permite resaltar algunos aspectos de la evolución presente del SARS-Cov-2, con relevancia para la selección de marcadores filogenéticos. Uno de los aspectos fundamentales es que el relativamente bajo nivel de diversidad que presenta todavía el genoma viral, deriva en una señal filogenética débil incluso a nivel de genomas, pero que pudiera mostrar un notable declive en la medida en la que se analicen fragmentos más pequeños.

Este comportamiento sugiere que, en el estado de evolución del virus en el momento en que fue realizado el estudio, los aspectos aleatorios, asociados a la evolución neutral, pudieran estar teniendo mayor influencia en los resultados de la inferencia filogenética que los aspectos asociados a presiones selectivas. La capacidad resolutiva de los métodos de reconstrucción filogenética está limitada por la baja fortaleza de la señal filogenética, y el adecuado muestreo taxonómico determinará el valor de la inferencia filogenética para cada estudio en particular. No parece existir, en concordancia con la baja diversidad observada, un número elevado de homoplasias que interfieran en la correcta inferencia de las relaciones evolutivas usando árboles.

Este estudio está limitado por los conjuntos de datos empleados, aunque algunas de las conclusiones surgen del análisis de los datos reflejados en la plataforma Nextstrain, donde se ofrece información basada en un muestreo bastante amplio de las cepas virales circulantes por todo el mundo. El análisis se basó en regiones del genoma delimitadas por la función, así que se analizaron los marcos abiertos de lectura en su totalidad y por separado, pero nada obliga a que un marcador filogenético esté restringido a un único marco abierto de lectura. Se deben realizar estudios posteriores que permitan, con un procedimiento automatizado y criterios adecuados de optimalidad, la selección específica de un segmento del genoma.

De acuerdo a lo que refleja el trabajo, el ORF1a sería el marcador cuyo uso reproduciría de forma más fiable las relaciones evolutivas que se pueden inferir a nivel de genomas, pero todo apunta a que se debe a su longitud (cerca de 14 kb), lo cual no los haría factible para su obtención por el método Sanger. Las regiones cuya longitud están en el orden del largo de lectura en el método Sanger (alrededor de 900 pb) no muestran un buen comportamiento como marcadores filogenéticos. La región de asociada a la fosfoproteína estructural de la nucleocápside (N), que cae esta categoría, y está entre las recomendadas como marcador molecular de la presencia del virus, tampoco parece comportarse como un buen marcador filogenético.

Agradecimientos

Agradecemos a la Dr.C. Monica Zoppè, del Consiglio Nazionale delle Ricerche, via Celoria 26, 20131 Milán, Italia, por las sugerencias iniciales para la realización de este trabajo. También nuestro agradecimiento para la Dr.C. Magilé Fonseca del Instituto de Medicina Tropical Pedro Kourí (IPK) por la revisión del trabajo, y sus valiosas sugerencias.

Materiales Suplementarios

Los materiales suplementarios se encuentran disponibles en https://drive.google.com/file/d/17FYUWIVOBYkIkczd-fqpHIS0qUXgu2yC/view?usp=sharing.

Referencias

Aguilar-Guerra, T. L. and G. Reed (2020). "Mobilizing Primary Health Care: Cuba's Powerful Weapon against COVID-19." MEDICC Review 22(2): 53-57.

Biggs, D., B. De Ville, et al. (1991). "A method of choosing multiway partitions for classification and decision trees." Journal of applied statistics 18(1): 49-62.

Castresana, J. (2000). "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis." Molecular Biology and Evolution 17(4): 540-552.

Ciccozzi, M., A. Lai, et al. (2019). "The phylogenetic approach for viral infectious disease evolution and epidemiology: An updating review." J Med Virol 91(10): 1707-1724.

Core, T. R. (2017). "R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http s." www. R-proje ct. org.

Darling, A. E., B. Mau, et al. (2010). "progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement." PLoS One 5(6): e11147.

Elbe, S. and G. Buckland-Merrett (2017). "Data, disease and diplomacy: GISAID's innovative contribution to global health." Global Challenges 1(1): 33-46.

Felsenstein, J. and G. A. Churchill (1996). "A Hidden Markov Model approach to variation among sites in rate of evolution." Molecular Biology and Evolution 13(1): 93-104.

Forster, P., L. Forster, et al. (2020). "Phylogenetic network analysis of SARS-CoV-2 genomes." Proceedings of the National Academy of Sciences 117(17): 9241-9243.

Friedman, M. (1940). "A comparison of alternative tests of significance for the problem of m rankings." The Annals of Mathematical Statistics 11(1): 86-92.

Garcés-Ayala, F., A. Araiza-Rodríguez, et al. (2020). "Full Genome Sequence of the first SARS-CoV-2 detected in Mexico." Archives of Virology: 1-4.

Gorry, C. (2020). "COVID-19 Case Detection: Cuba's Active Screening Approach." MEDICC Review 22(2): 58-63.

Goujon, M., H. McWilliam, et al. (2010). "A new bioinformatics analysis tools framework at EMBL-EBI." Nucleic acids research 38(Web Server issue): W695-699.

Hadfield, J., C. Megill, et al. (2018). "Nextstrain: real-time tracking of pathogen evolution." Bioinformatics 34(23): 4121-4123.

Hodcroft, E. (2020). "Clade Naming & Definitions." Retrieved April 20, 2020, from https://github.com/nextstrain/ncov/blob/master/docs/clades.md.

Holm, S. (1979). "A simple sequentially rejective multiple test procedure." Scandinavian journal of statistics: 65-70.

Huerta-Cepas, J., F. Serra, et al. (2016). "ETE 3: reconstruction, analysis, and visualization of phylogenomic data." Molecular Biology and Evolution 33(6): 1635-1638.

IBM, C. IBM SPSS Statistics for Windows, version 20. Armonk, N.Y,USA, IBM Corp.

Jia, Y., C. Yang, et al. (2020). "Characterization of eight novel full-length genomes of SARS-CoV-2 among imported COVID-19 cases from abroad in Yunnan, China." Journal of Infection.

Katoh, K. and D. M. Standley (2013). "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." Molecular Biology and Evolution 30(4): 772-780.

Kent, W. J. (2002). "BLAT—the BLAST-like alignment tool." Genome research 12(4): 656-664.

Kimura, M. (1969). "The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations." Genetics 61(4): 893.

Kishino, H., T. Miyata, et al. (1990). "Maximum likelihood inference of protein phylogeny and the origin of chloroplasts." Journal of Molecular Evolution 31(2): 151-160.

Lin, K. H., C. L. Chern, et al. (2001). "Genetic analysis of recent Taiwanese isolates of a variant of coxsackievirus A24." J Med Virol 64(3): 269-274.

Lu, R., X. Zhao, et al. (2020). "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding." The Lancet 395(10224): 565-574.

Lv, L., G. Li, et al. (2020). "Comparative genomic analysis revealed specific mutation pattern between human coronavirus SARS-CoV-2 and Bat-SARSr-CoV RaTG13." bioRxiv.

Nabhan, A. R. and I. N. Sarkar (2012). "The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy." Briefings in bioinformatics 13(1): 122-134.

Nguyen, L.-T., H. A. Schmidt, et al. (2015). "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies." Molecular Biology and Evolution 32(1): 268-274.

of the International, C. S. G. (2020). "The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2." Nature Microbiology 5(4): 536.

Pachetti, M., B. Marini, et al. (2020). "Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant." Journal of Translational Medicine 18: 1-9.

Patwardhan, A., S. Ray, et al. (2014). "Molecular markers in phylogenetic studies-a review." Journal of Phylogenetics & Evolutionary Biology 2014.

Rambaut, A. (2017). "FigTree-version 1.4. 3, a graphical viewer of phylogenetic trees." Computer program distributed by the author, website: http://tree. bio. ed. ac. uk/software/figtree.

Rambaut, A., E. C. Holmes, et al. (2020). "A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology." bioRxiv: 2020.2004.2017.046086.

Reed, G. (2020). "Stemming COVID-19 in Cuba: Strengths, Strategies, Challenges Francisco Durán MD." MEDICC Review 22(2): 47-52.

Robinson, D. F. and L. R. Foulds (1981). "Comparison of phylogenetic trees." Mathematical biosciences 53(1-2): 131-147.

Russo, C., B. Aguiar, et al. (2017). "Selecting Molecular Markers for a Specific Phylogenetic Problem." MOJ Proteomics Bioinform 6(3): 00196.

Schmidt, H. A., K. Strimmer, et al. (2002). "TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing." Bioinformatics 18(3): 502-504.

Schmidt, H. A. and A. von Haeseler (2009). "Phylogenetic inference using maximum likelihood methods." The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing 2nd Edition Cambridge University Press, Cambridge: 181-209.

Shimodaira, H. (2002). "An approximately unbiased test of phylogenetic tree selection." Systematic biology 51(3): 492-508.

Shimodaira, H. and M. Hasegawa (1999). "Multiple comparisons of log-likelihoods with applications to phylogenetic inference." Molecular Biology and Evolution 16(8): 1114-1114.

Shu, Y. and J. McCauley (2017). "GISAID: Global initiative on sharing all influenza data–from vision to reality." Eurosurveillance 22(13).

Strimmer, K. and A. von Haeseler (1996). "Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies." Molecular Biology and Evolution 13(7): 964-964.

Trifinopoulos, J., L.-T. Nguyen, et al. (2016). "W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis." Nucleic acids research 44(W1): W232-W235.

Velazquez-Salinas, L., S. Zarate, et al. (2020). "Positive selection of ORF3a and ORF8 genes drives the evolution of SARS-CoV-2 during the 2020 COVID-19 pandemic." bioRxiv.

Volz, E., H. Fu, et al. (2020). "Genomic epidemiology of a densely sampled COVID19 outbreak in China." medRxiv.

Wang, R., Y. Hozumi, et al. (2020). "Decoding SARS-CoV-2 transmission, evolution and ramification on COVID-19 diagnosis, vaccine, and medicine." arXiv preprint arXiv:2004.14114.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. Breakthroughs in statistics, Springer: 196-202.

Wu, D., G. Jospin, et al. (2013). "Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups." PLoS One 8(10): e77033.

Xia, X. and Z. Xie (2001). "DAMBE: Software Package for Data Analysis in Molecular Biology and Evolution." Journal of Heredity 92(4): 371-373.

Xia, X., Z. Xie, et al. (2003). "An index of substitution saturation and its application." Molecular Phylogenetics and Evolution 26(1): 1-7.

Yin, C. (2020). "Genotyping coronavirus SARS-CoV-2: methods and implications." Genomics.

Zhou, P., X.-L. Yang, et al. (2020). "A pneumonia outbreak associated with a new coronavirus of probable bat origin." nature 579(7798): 270-273.