Revisión

Nextstrain: una herramienta que analiza la epidemiología molecular del SARS-CoV-2

Nextstrain: a tool to analyze the molecular epidemiology of SARS-CoV-2

Sebastián Iglesias-Osores[1*] https://orcid.org/0000-0002-4984-4656
Miguel Alcántara-Mimbela[1] https://orcid.org/0000-0001-9189-9098
Zhandra Arce-Gil[2] https://orcid.org/0000-0002-8894-9186
Lizbeth M. Córdova-Rojas[3] https://orcid.org/0000-0002-9998-5019
Elmer López-López[4] https://orcid.org/0000-0002-8414-7805
Arturo Rafael-Heredia[5] https://orcid.org/0000-0001-7461-0176

[1]Universidad Nacional Pedro Ruiz Gallo, Facultad de Ciencias Biológicas. Lambayeque, Perú.
[2]Universidad Católica Santo Toribio de Mogrovejo, Facultad de Medicina Humana. Chiclayo, Perú.
[3]Universidad Nacional de Jaén. Cajamarca, Perú.
[4]Universidad Señor de Sipán, Facultad de Medicina Humana. Perú.
[5]Facultad de Medicina Humana de la Universidad Nacional de Ucayali. Ucayali, Perú.

*Autor para la correspondencia: sebasiglo@gmail.com

RESUMEN

La preocupación mundial por el nuevo coronavirus (2019-nCoV), como una amenaza global para la salud pública, fue el motor para que los análisis filogenéticos sufrieran un crecimiento exponencial. El objetivo de esta revisión fue describir el modo de funcionamiento y las bondades de la herramienta Nextstrain, así como el secuenciamiento del virus SARS-CoV-2 en el mundo. Se

usó la interfaz de la página de Nextstrain para mostrar sus funcionalidades y los modos de visualización de datos, y se descargaron estos de la web GISAID para mostrar la cantidad de secuenciamientos del SARS-CoV-2 hasta la fecha. Nextstrain es un proyecto de código abierto creado por biólogos bioinformáticos, para aprovechar el potencial científico y de salud pública de los datos de genomas de patógenos. Nextstrain consiste en un conjunto de herramientas que toman secuencias sin procesar (en formato FASTA). Nextstrain realiza una alineación de secuencia de los datos de entrada en alineación de secuencia múltiple basada en la transformación rápida de Fourier. Se basa en el uso de dos *softwares*: *Augur* y *Auspice*. Nextstrain es una herramienta eficaz para mostrar datos epidemiológicos de manera simple para un público no especializado. Puede ser usado en la salud pública, ya que muestra datos en tiempo real de las epidemias y su distribución geográfica. Se puede usar para dar seguimiento a los brotes como es el caso del COVID-19.
Palabras clave: Filogenética; SARS-CoV-2; COVID-19; Nexstrain; epidemiología.

ABSTRACT

Worldwide concern about the novel coronavirus (2019-nCoV) as a global threat to public health is the reason for the exponential growth of phylogenetic analyses. The purpose of this review was to describe the mode of operation and advantages of the tool Nextstrain, as well as the sequencing of the SARS-CoV-2 virus worldwide. The interface of the Nextstrain page was used to show its functions and data visualization modes. These were downloaded from the website GISAID to show the number of SARS-CoV-2 sequencing processes performed so far. Nextstrain is an open code project created by bioinformatics biologists to make good use of the scientific and public health potential of data about genomes of pathogens. Nextstrain consists in a set of tools operating with unprocessed sequences (in FASTA format). Nextstrain performs a sequence alignment of the input data into a multiple sequence alignment based on fast Fourier transform. Its use is based on two software applications: Augur and Auspice. Nextstrain is an efficient tool by which lay people may obtain epidemiological data in a simple manner. It may be used in the public health sector, since it shows real time data about epidemics and their geographic

distribution. It may also be used to follow-up outbreaks, as is the case with COVID-19.

Key words: Phylogenetics; SARS-CoV-2; COVID-19; Nextstrain; epidemiology.

Recibido: 25/04/2020
Aceptado: 22/12/2020

Introducción

La preocupación mundial por el nuevo coronavirus 2019 - nCoV, como una amenaza global para la salud pública, fue el motor para que los análisis filogenéticos sufrieran un crecimiento exponencial. Esto se mostró cuando el 2019-nCoV se agrupó significativamente con la secuencia de coronavirus similar a SARS de murciélago aislada en el año 2015.[1] Estas epidemias han dado lugar al desarrollo de un sistema nacional de vigilancia, como sucedió con el VIH.[2] El estudio de la propagación de epidemias es fundamental para nuestra comprensión del desarrollo de sus procesos dinámicos.[3]

Los datos de vigilancia epidemiológica proporcionan una comprensión de los riesgos de transmisión y caracterizan a las comunidades afectadas por la epidemia.[2] En los últimos años, la comunidad investigadora ha acumulado evidencia abrumadora de la aparición de patrones de conectividad complejos y heterogéneos en una amplia gama de sistemas biológicos y sociotécnicos.[3] Las redes y la epidemiología de las enfermedades infecciosas de transmisión directa están fundamentalmente vinculadas.[4]

El análisis de las interacciones y actividades de la epidemiología digital es nuevo y rápido en las plataformas de las redes sociales, las cuales pueden generar respuestas y son un campo en crecimiento.[5] Los avances en genómica e informática están transformando la capacidad de los científicos de responder a los brotes.[6] Las tecnologías de secuenciación de próxima generación son

mucho más rápidas y baratas que hace unos años. La secuenciación a gran escala de genomas se está convirtiendo en una realidad tangible.[7]

En la actualidad ha surgido una nueva forma de estudiar las mutaciones que se acumulan en el genoma de las células o los virus y pueden usarse para inferir su historia evolutiva.[8] Uno de esos casos es la herramienta llamada Nextstrain, que nos permite visualizar en tiempo real y con un análisis previo los virus secuenciados en el mundo. Esto es de importancia en el caso de organismos que evolucionan rápidamente. Los genomas pueden revelar su diseminación espaciotemporal detallada,[8] para lo cual Nexstrain muestra una de sus mejores funcionalidades: la visualización de datos de forma simple. Tales análisis filodinámicos son particularmente útiles para comprender la epidemiología de los patógenos virales que evolucionan rápidamente.[8]

A medida que la cantidad de secuencias genómicas disponibles para diferentes patógenos ha aumentado dramáticamente en los últimos años, el análisis filodinámico con métodos tradicionales se vuelve desafiante, ya que estos métodos escalan mal con el crecimiento de los conjuntos de datos.[8] Por ejemplo, el paquete *de software Bayesian Evolutionary Analysis by Sampling Trees* (BEAST) se ha convertido en una herramienta principal para la inferencia filogenética y filodinámica bayesiana, a partir de datos de secuencias genéticas.[9]

Nextstrain se basa para su análisis en BEAST, el cual unifica la reconstrucción filogenética molecular con la evolución de rasgos complejos discretos y continuos, la datación en el tiempo de divergencia y los modelos demográficos.[9] Nextstrain es una colección de herramientas de código abierto que usa repositorios abiertos para nutrirse. Esta herramienta ayuda a comprender la propagación y evolución de patógenos, especialmente en escenarios de brotes o pandemias. Su diseño nos permite usar una amplia gama de combinaciones y gama de fuentes de datos, de fácil uso, que facilita reemplazar variables para la visualización de datos. La visualización integra datos de secuenciamiento con otros tipos de datos, como información

geográfica, sexo, clados, serología o especies hospedadoras.[10] Nextstrain fue creado por *Trevor Bedford* y otros[11] y fue usada para el análisis genómico rápido para monitorear patógenos, como el virus del Ébola o el virus del Nilo Occidental, a medida que evolucionan y se propagan.

La gran cantidad de muestras secuenciadas del virus SARS-CoV-2 a nivel mundial hacen muy difícil su análisis y posterior utilización en la salud pública. El análisis de datos bioinformáticos no es de fácil lectura para el público no especializado. La utilización de herramientas, como Nextstrain, hace que la información sea universal. El objetivo de esta revisión fue describir el modo de funcionamiento y las bondades de la herramienta Nextstrain, así como el secuenciamiento del virus SARS-CoV-2 en el mundo.

Métodos

Se usó la interfaz de la página de Nextstrain para mostrar sus funcionalidades y modos de visualización de datos y se descargaron los datos de la web GISAID para mostrar la cantidad de secuenciamientos del SARS-CoV-2 desde diciembre del año 2019 hasta el 24 de julio de 2020.

Se realizó una revisión en las bases de datos MEDLINE/PubMed y SCOPUS, y en el buscador Google Scholar, donde se recuperaron los artículos publicados hasta junio del año 2020 que consignaron las palabras “Nexstrain”. Los artículos obtenidos fueron revisados a texto completo con el fin de verificar que se trataba de la herramienta y no de palabras separadas como Next strain o Next-strain por ejemplo. Se obtuvieron 28 artículos. Fueron eliminadas las publicaciones que no cumplieron con ambos criterios y los resultados duplicados. Se usó la interfaz de la página de Nextstrain para mostrar sus funcionalidades y modos de visualización de datos y se descargaron los datos de la web GISAID para mostrar la cantidad de secuenciamientos del SARS-CoV-2 hasta el 24 de julio del año 2020.

¿Qué es Nextstrain?

Comprender la propagación y la evolución de los patógenos es importante para la vigilancia y las medidas eficaces de salud pública.[10] Nextstrain es un proyecto de código abierto creado por biólogos bioinformáticos, para aprovechar el potencial científico y de salud pública de los datos de genomas de patógenos. Proporciona una vista continuamente actualizada de los datos disponibles públicamente junto con potentes herramientas analíticas y de visualización para uso de la comunidad. Su objetivo es ayudar a la comprensión epidemiológica y mejorar la respuesta al brote. Nextstrain consta de una base de datos de genomas virales, una tubería de bioinformática para el análisis de la filodinámica y una plataforma de visualización interactiva.[10] Presenta los siguientes pilares:

A) *Filogenias de patógenos:* Los patógenos tienden a acumular mutaciones aleatorias en sus genomas. Las mutaciones pueden usarse como un marcador de transmisión en el que los genomas estrechamente relacionados indican infecciones estrechamente relacionadas o donde se ha producido la mutación. Al reconstruir una filogenia podemos aprender sobre fenómenos epidemiológicos importantes, como la propagación espacial, los tiempos de introducción y la tasa de crecimiento de la epidemia.

B) *Permite hacer inferencias:* Si las secuencias del genoma del patógeno van a informar las intervenciones de salud pública, entonces los análisis deben realizarse rápidamente y los resultados deben difundirse ampliamente. Las prácticas actuales de publicación científica obstaculizan la rápida difusión de resultados epidemiológicamente relevantes.

C) *Sitio web:* Tiene como objetivo 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 científicos ciudadanos.

¿Cómo funciona Nextstrain?

Nextstrain consiste en un conjunto de herramientas que toman secuencias sin procesar (en formato FASTA, que es un formato basado en texto para representar secuencias de nucleótidos o secuencias de aminoácidos)[12] y metadatos asociados (por ejemplo, hora, sexo, región, país, publicaciones, autores y laboratorios) como entrada.

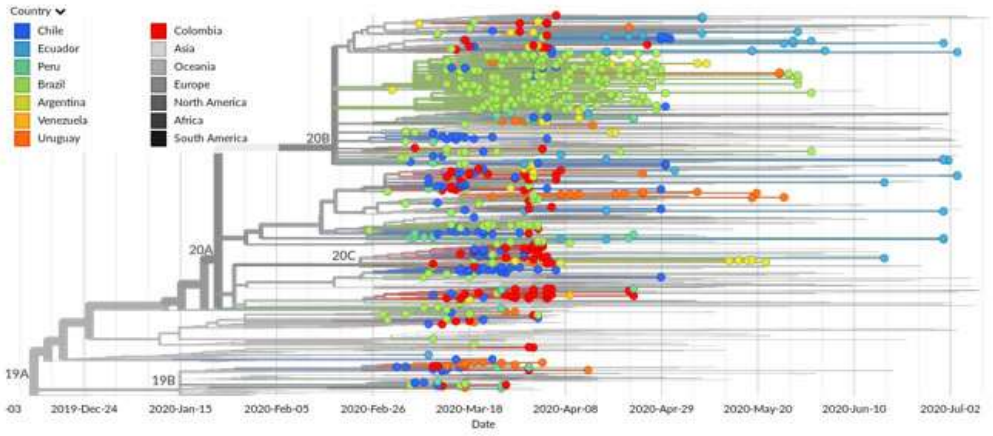
Nextstrain realiza una alineación de secuencia de los datos de entrada basada en la transformación rápida de *Fourier* (MAFFT)[13] que, a su vez, se utiliza para inferir una probabilidad máxima. Su estudio se basa en la topología del árbol de máxima verosimilitud (ML) que se infiere en RAxML usando el modelo general de tiempo de reversión (GTR) + G + I de sustitución de nucleótidos.[14]

La topología del árbol resultante se transforma en una filogenia fechada donde las ramas corresponden a unidades de tiempo real usando un enfoque de datación de mínimos cuadrados.[15] A su vez, la filogenia fechada se usa para realizar la reconstrucción del estado ancestral para inferir la ubicación probable de los nodos internos en la filogenia, usando un enfoque de probabilidad marginal.[16] Nextstrain compila muchas opciones en una única ubicación accesible, abierta a profesionales de la salud, epidemiólogos, virólogos y al público en general.[10] En términos generales, Nextstrain se basa en el uso de dos *softwares*:

> *Augur:* El análisis de la función de la proteína es un desafío y un importante "cuello de botella" en el análisis de genomas.[10] Las proteínas de la superficie bacteriana, por ejemplo, presentan información que puede ser usada en la farmacológica y el desarrollo de vacunas. Augur es una tubería de predicción automática que integra los principales algoritmos de predicción de superficie y permite el análisis comparativo, la clasificación y la visualización de microrganismos a escala genómica,[17] mediante una serie de herramientas bioinformáticas modulares (tipo Unix). Augur necesita de herramientas externas para el

alineamientos de secuencias múltiples de los genomas virales. Usa el *software* MAFFT (https://mafft.cbrc.jp/alignment/software/).[18] Además, para la inferencia filogenética de los árboles mostrados en el sitio web, usa la inferencia de máxima probabilidad, a través del *software* IQ-Tree (http://www.iqtree.org/).[19]

Auspice: Es un *software* para mostrar visualizaciones interactivas de datos filogenómicos. Puede ejecutarse en su computadora o integrarse en sitios web (https://github.com/nextstrain/auspice). En este conjunto de datos, la filogenia está coloreada por "país" (país de muestreo) de Suramérica y se muestra como un árbol en una línea de tiempo. Se obtuvieron los datos de Nexstrain sobre el número de muestras secuenciadas de SARS-CoV-2 de Argentina (36), Asia (343), Brasil (324), Chile (134), Colombia (118), Ecuador (46), Perú (34), Uruguay (61) y Venezuela (2) (Fig. 1).

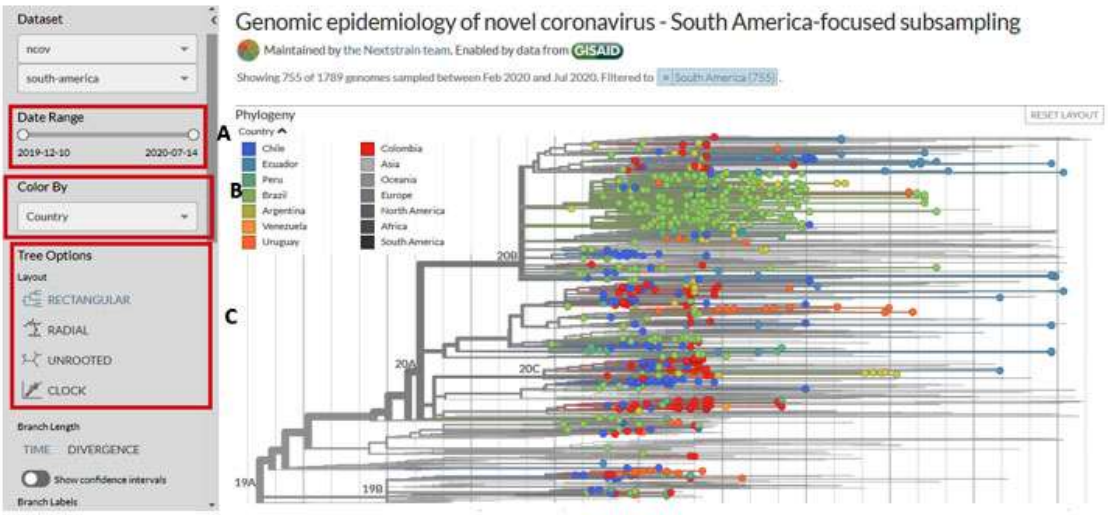


Fuente: https://nextstrain.org/ncov/south-america

Fig. 1 - Árbol filogenético de las muestras secuenciadas del virus SARS-CoV-2 de Suramérica clasificado por clado y coloreado por países.

El panel de control de la izquierda también puede cambiar (Fig. 2), usando los controles deslizantes en cualquier extremo del "intervalo de fechas" (A). Luego, usando el menú desplegable "color por" (B), intentamos colorear la filogenia por región, autor y fecha. También podemos cambiar el diseño de la filogenia:

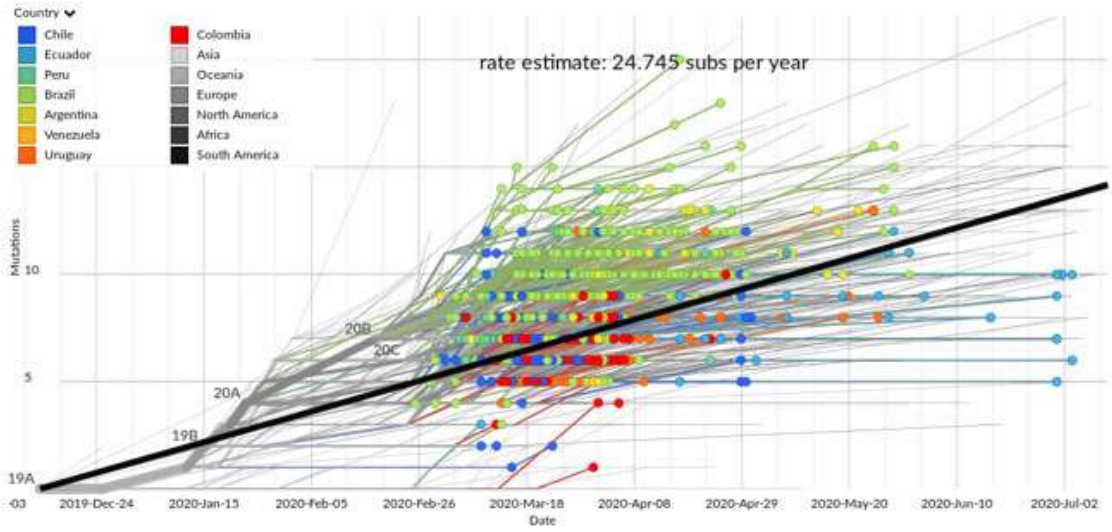
intentamos hacer clic en "Radial" y "Sin raíz" en "Opciones de árbol - Diseño" y vemos cómo cambia el árbol (C).



Fuente: https://nextstrain.org/ncov/south-america.

Fig. 2 - Panel de control de Nextstrain del virus SARS-CoV-2 de Suramérica clasificado por clado y coloreado por países.

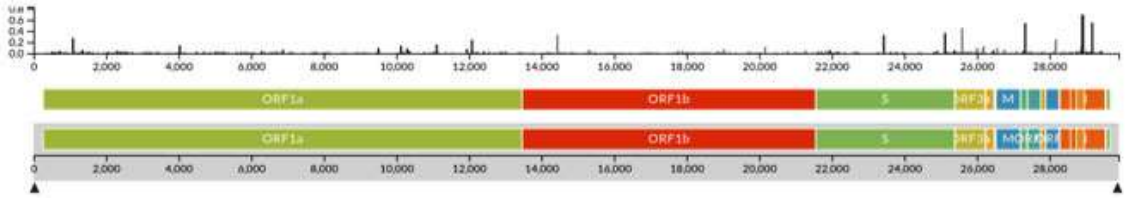
La figura 3 nos muestra divergencia (número de mutaciones desde el ancestro común) en el eje “y” y la fecha de muestreo en el eje “x”. En el caso ideal, todos los puntos caen aproximadamente en una línea y la pendiente de esta línea es la tasa evolutiva, es decir, el número de sustituciones que se acumulan por año.



Fuente: https://nextstrain.org/ncov/south-america

Fig. 3 - Árbol filogenético en reloj de las muestras secuenciadas del virus SARS-CoV-2 de Suramérica clasificado por clado y coloreado por países, donde se muestra la divergencia.

La figura 4 muestra que el panel de diversidad es una buena forma de ver las mutaciones en la filogenia. Le permite detectar regiones en el genoma que evolucionan más rápidamente que otras; por ejemplo, porque una región en particular es el objetivo del sistema inmunitario del huésped y cambia rápidamente.

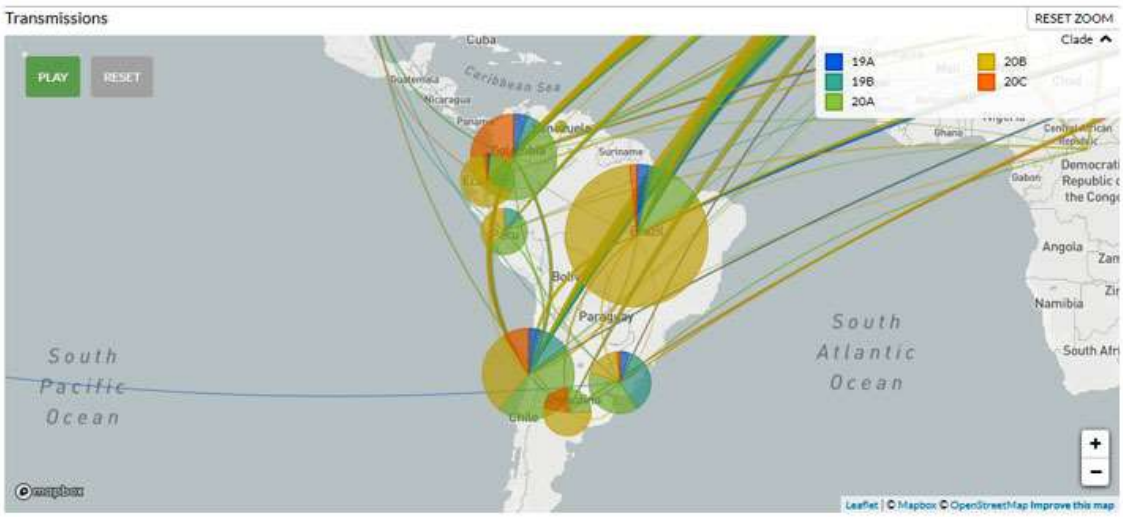


Fuente: https://nextstrain.org/ncov/south-america

Fig. 4 - Diversidad de muestras secuenciadas del virus SARS-CoV-2 de Suramérica.

El panel de mapa no es tan interactivo como el panel de filogenia, pero cambia para reflejar acciones, como la de hacer un acercamiento o cambiar las fechas en la filogenia. Usando el botón "reproducir" en la parte superior izquierda del panel del mapa, puede moverse a través del tiempo en la filogenia y ver cómo

el virus puede haberse propagado geográficamente. La precisión de estas reconstrucciones depende de muchas cosas, por lo que es importante tener en cuenta si mostrar dichos enlaces es apropiado para sus datos. Puede ajustar la velocidad de la animación utilizando las "Opciones de mapa" en el panel de control. También puede cambiar el nivel de detalle de la ubicación que se muestra en el mapa. Podemos cambiar la "Resolución geográfica" a "región" (Fig. 5).



Fuente: https://nextstrain.org/ncov/south-america

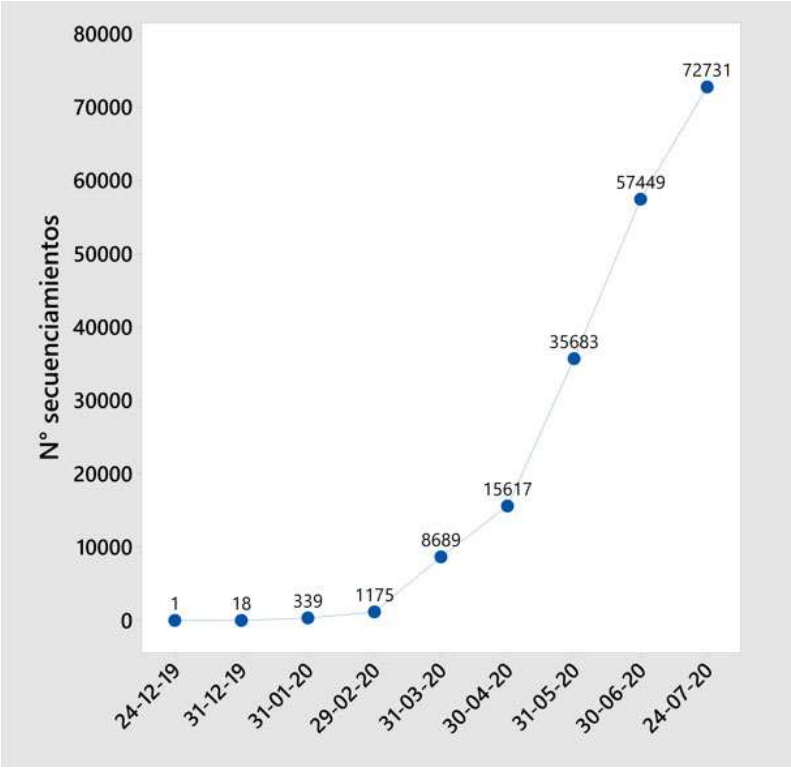
Fig. 5 - Mapa de muestras secuenciadas del virus SARS-CoV-2 de Suramérica clasificadas en clados.

Todos los datos para el análisis y el intercambio de datos son almacenados por los científicos con el código utilizado en los repositorios de GitHub (https://nextstrain.org/), y los resultados también se almacenan en estos repositorios. La herramienta de *software* en la que se basa GitHub se llama Git. Fue creado en el año 2005 por el codificador Linus Torvalds para gestionar el desarrollo del sistema operativo de código abierto Linux, un gran proyecto que involucró a miles de programadores independientes. [20]

Los datos de Nextstrain se basan en una red mundial de científicos que comparten datos abiertos a través de GISAID (*Global Initiative on Sharing All Influenza Data*). Los investigadores continúan secuenciando muestras virales que ayudan a completar las ramas del árbol evolutivo del nuevo coronavirus, y

ese trabajo continuará brindando a los visitantes de Nextstrain una mirada colorida sobre el pasado, el presente y, quizás, el futuro de la pandemia de COVID-19.

GISAID proporciona la plataforma de intercambio de datos particularmente utilizada por GISRS (*Global Influenza Surveillance and Response System*), a través de la cual los datos de secuencia considerados por la Organización Mundial de la Salud (OMS) en la selección de virus recomendados para su inclusión en vacunas estacionales y prepandémicas se comparten abiertamente, y de los cuales dependen los investigadores científicos, funcionarios de salud pública y animal y la industria farmacéutica. Tal apertura de los datos más actualizados ayuda a comprender y a mejorar la credibilidad de las recomendaciones de la OMS para la composición de estas vacunas estacionales y las pandemias potenciales.[21] El número de secuenciamientos va en aumento, y se puede observar en la figura 6.



Fuente: GISAID.

Fig. 6 - Número de secuenciamientos a nivel mundial del SARS-CoV-2 hasta el 24 de julio de 2020.

Motivado por la diversidad genética del virus SARS-CoV-2, GISAID introdujo un sistema de nomenclatura para clados principales, que se basa en mutaciones de marcadores dentro de seis agrupaciones filogenéticas de alto nivel de la división temprana de S y L, para una mayor evolución de L en V y G y más tarde de G en GH y GR divididos en clados. Un clado, también conocido como grupo monofilético o grupo natural, es un grupo de organismos compuesto por un ancestro común y todos sus descendientes lineales.[22]

Los clados de GISAID se incrementan con linajes más detallados, asignados por la herramienta Phylogenetic Assignment of Named Global Outbreak LINeages (PANGOLIN), lo que ayuda a comprender los patrones y los determinantes de la propagación global de la cepa pandémica que causa COVID-19. Las definiciones de clados en GISAID están informadas por la distribución estadística de las distancias del genoma en grupos filogenéticos,[23] seguido de la fusión de linajes más pequeños en clados principales basados en variantes de marcadores compartidos. En lugar de las letras genéricas A, B, C, elegimos letras reales de mutaciones marcadoras (alfabeto para sustituciones no sinónimas y número para sinónimos) para hacer que el sistema sea más tangible y específico para este virus. Las extensiones de clado y nombre se activan cuando un clado se puede subdividir.

Otra característica única es permitir la caída de letras y números en frente para evitar cadenas de letras/números monótonas, como se usan en otros virus. Usando combinaciones específicas de nueve marcadores genéticos, el 95 % de los datos de hCOV-19 en GISAID se pueden clasificar en seis clados de tamaño equilibrado. Por ejemplo, comenzando con S y L,[24] S continuó a niveles moderados y L se dividió en versiones G y V, inicialmente iguales con G, que alcanzó el 50 % en marzo del año 2020 y se dividió aún más en GR y GH. La lista de las nueve variantes de marcador es la siguiente (Fig. 7 y 8):

S: C8782T, T28144C incluye NS8-L84S

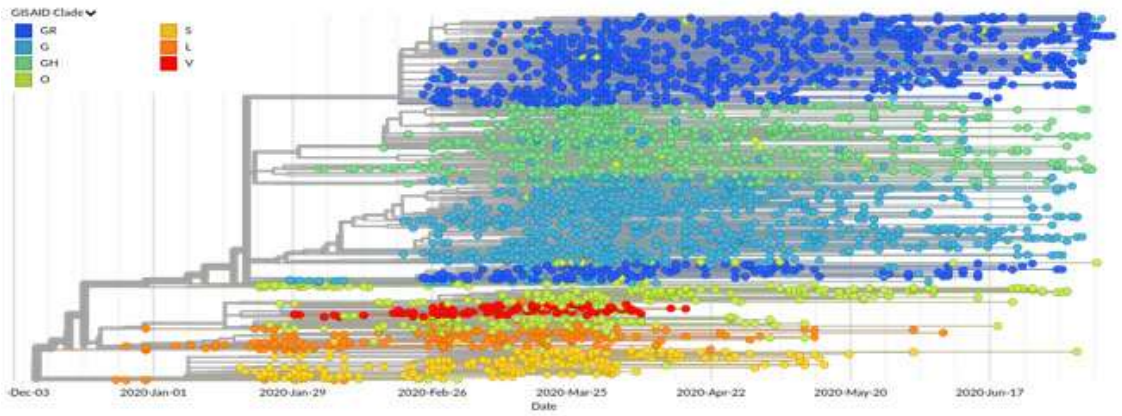
L: C241, C3037, A23403, C8782, G11083, G25563, G26144, T28144, G28882 (secuencia de referencia WIV04)

V: G11083T, G26144T NSP6-L37F + NS3-G251V

G: C241T, C3037T, A23403G incluye S-D614G

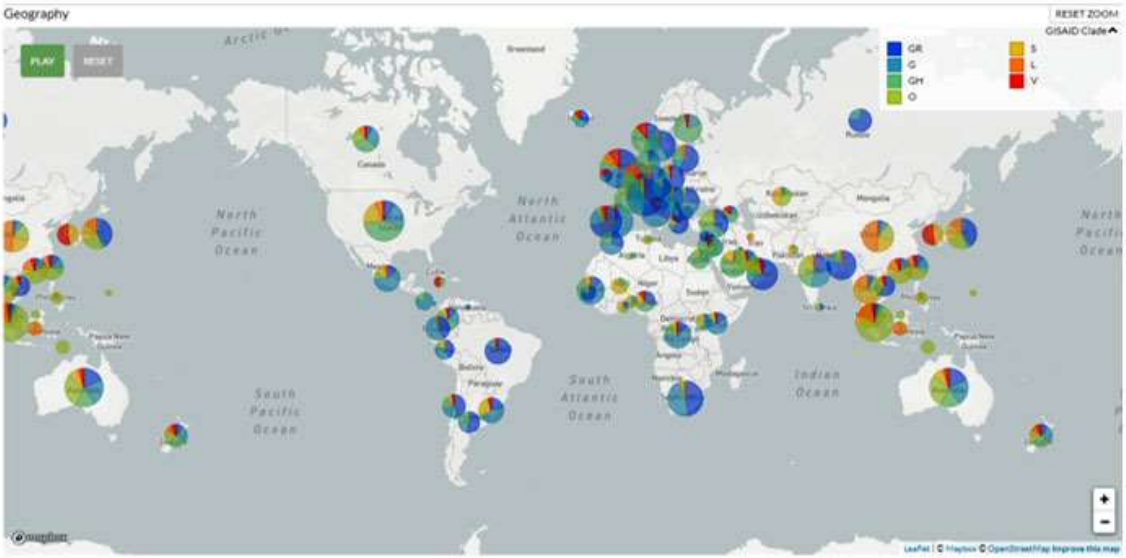
GH: C241T, C3037T, A23403G, G25563T incluye S-D614G + NS3-Q57H

GR: C241T, C3037T, A23403G, G28882A incluye S-D614G + N-G204R



Fuente: https://nextstrain.org/ncov/south-america

Fig. 7 - Árbol filogenético que muestra clados determinados y coloreados por GISAID.



Fuente: https://nextstrain.org/ncov/south-america

Fig. 8 - Mapa que muestra los clados determinados y coloreados por GISAID.

Las definiciones de clado en GISAID se aumentan con linajes más detallados, asignados por la herramienta filogenética de asignación de nombres globales de brote global (PANGOLIN) de *Rambaut* y otros,[25] un esfuerzo adicional que ayuda a comprender los patrones y determinantes de la propagación global de la cepa pandémica que causa la COVID-19. Para cada cepa, la información del

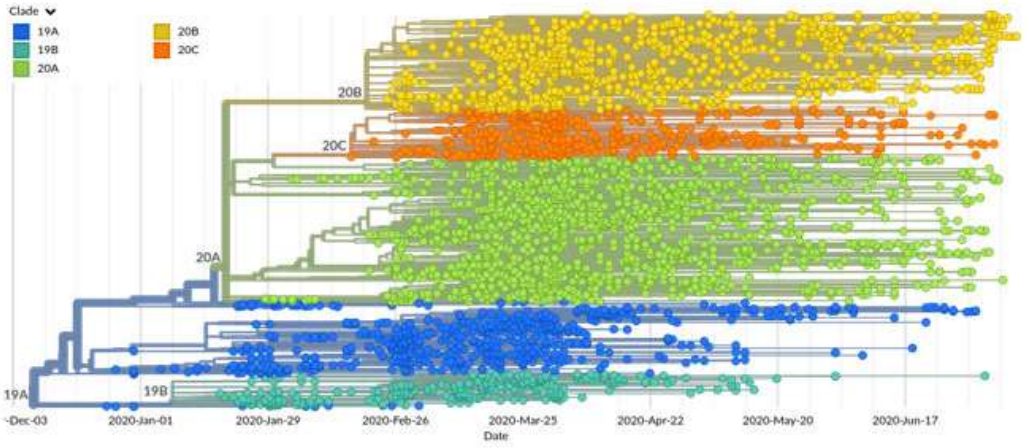
clado se proporciona en la sección "Detalles del virus" de los metadatos. Los linajes asignados por esta herramienta de *software* se caracterizan por una combinación de soporte genético y epidemiológico. Esta nomenclatura jerárquica y dinámica describe un linaje como un grupo de secuencias vistas en una región geográficamente distinta con evidencia de transmisión continua en esa región. Se tienen en cuenta múltiples fuentes de información, incluida la información filogenética, así como una variedad de metadatos asociados con esa secuencia. La escala más fina de este sistema de nomenclatura puede ayudar a separar las investigaciones de brotes y, a medida que aumentan las tasas de viajes internacionales, facilitará el seguimiento de las importaciones virales en todo el mundo.

Otro esfuerzo, por *Hodcroft* y otros,[26] utiliza una nomenclatura de letras de año para facilitar la discusión de los patrones de diversidad a gran escala de hCoV-19 y los clados de etiquetas que persisten durante, al menos, varios meses y tienen una distribución geográfica significativa. Cada nombre de clado consiste en el año en que surgió el clado y una letra mayúscula que comienza con A para cada año. Los clados se definen por mutaciones de firma. Los nuevos clados principales se nombran una vez que la frecuencia de un clado excede el 20 % en una muestra global representativa y ese clado difiere en, al menos, dos posiciones de su clado padre, actualmente utilizando los clados 19A, 19B, 20A, 20B y 20C.

Nextstrain también introdujo designaciones de clado informales para SARS-CoV-2 el 4 de marzo del año 2020, en gran medida para ayudar a las discusiones internas y crear enlaces URL que permitieran el "acercamiento automático" a un área del árbol que era de interés. Estos nombres de clados eran combinaciones de letras y números ad-hoc (por ejemplo, A2a) y nunca pretendieron ser un sistema de nombres permanente (nunca visibles por defecto). Sin embargo, estos clados han sido utilizados en algunos casos para discutir diferentes aspectos de la filogenia en Nextstrain, lo que subraya la necesidad de una propuesta formal a más largo plazo para designar clados SARS-

CoV-2. Se designaron las etiquetas de los clados genéticamente bien definidos que han alcanzado una frecuencia significativa y una extensión geográfica.

Los dos primeros clados son 19A y 19B que corresponden a la división marcada por las mutaciones C8782T y T28144C. Estos clados prevalecieron en Asia durante los primeros meses del brote. El siguiente clado que se nombró fue el 20ª, correspondiente a un gran brote europeo que existió a principios del año 2020. Se distingue de su padre 19A por las mutaciones C3037T, C14408T y A23403G. Después de esto, hemos visto aparecer otros dos clados: 20B (otro clado europeo separado claramente por tres mutaciones consecutivas: G28881A, G28882A y G28883C) y 20C (un clado mayormente norteamericano, distinguido por las mutaciones C1059T y G25563T). Las definiciones de clado se codifican como un archivo tabular que define una firma genotípica para cada clado. Proporcionamos un script que genera una tabla con asignaciones de clado para un conjunto de secuencias (Fig. 9).



Fuente: https://nextstrain.org/ncov/south-america

Fig. 9 - Árbol filogenético mostrando clados determinados y coloreados por Nextstrain.

Consideraciones finales

El enfoque filodinámico se ha convertido en un elemento fundamental para el desarrollo de herramientas filogenéticas comparativas capaces de incorporar

datos de vigilancia epidemiológica con secuencias moleculares en un único marco estadístico.[27] Se usó Nextstrain principalmente para la visualización de datos filogenéticos en la actual pandemia de COVID-19.[28] También se usó para analizar al H3N2 y sus frecuencias de mutación máximas,[29] y ha permitido el seguimiento de cepas en todo el mundo aislado entre los años 2016 y 2018.[30]

Esta herramienta innovadora ha mejorado en gran medida las investigaciones científicas de los orígenes temporales y geográficos, la historia evolutiva y los factores de riesgo ecológicos asociados con el crecimiento y la propagación de virus, como el virus de la inmunodeficiencia humana (VIH), el zika y el dengue; las bacterias y la resistencia a la meticilina, como *Staphylococcus aureus*,[27] y el análisis de la evolución del serotipo 2 del virus del dengue en las Américas y el Caribe desde Nextstrain.[31] Puede usarse para rastrear la propagación y la evolución del virus del Nilo Occidental, así como también descubrir nuevas medidas de control específicas para ayudar a aliviar su carga de salud pública.[11] También se usa para analizar otros microrganismos, como la diversidad genómica de los aislados de ToBRFV (El virus rugoso del tomate).[32]

La recopilación, visualización y análisis de datos de brotes se están volviendo cada vez más complejos, por motivo de la diversidad en los tipos de datos.[33] El papel del análisis filogenético en la aclaración de la fuente de infección de un paciente de COVID-19, basado en las representaciones de Nextstrain, concluye que es mucho más probable que este se haya infectado durante su viaje al extranjero,[34] y para estudiar la propagación de costa a costa del SARS-CoV-2 durante la epidemia temprana en los Estados Unidos,[35] a través de visualizaciones de datos interactivas se permite la exploración de conjuntos de datos continuamente actualizados, proporcionando una herramienta de vigilancia novedosa para las comunidades científicas y de salud pública.[10]

Existen herramientas similares como GLUE (*Genes Linked by Underlying Evolution*) que organizan datos de secuencia a lo largo de líneas evolutivas, capturando no solo datos de nucleótidos, sino también elementos asociados tales como alineaciones, definiciones de genotipos, anotaciones y motivos del

genoma.[36] Nextflu es una aplicación web que muestra un árbol filogenético, el cual puede ofrecer información, como el genotipo viral en sitios específicos, la ubicación de muestreo y las estadísticas derivadas.[37] La nomenclatura de clados y los linajes ayudan en los estudios de epidemiología genómica de los virus hCoV-19 activos.[38]

Nextstrain es una herramienta eficaz para mostrar datos epidemiológicos de manera simple para un público no especializado. Puede ser usado en la salud pública, ya que muestra datos en tiempo real de las epidemias y su distribución geográfica, y se puede emplear para dar seguimiento a los brotes, como es el caso de la COVID-19.

Referencias bibliográficas

1. Benvenuto D, Giovanetti M, Ciccozzi A, Spoto S, Angeletti S, Ciccozzi M. The 2019-new coronavirus epidemic: Evidence for virus evolution. J Med Virol [Internet]. 2020 [acceso: 25/07/2020];92(4):455-9. Disponible en: https://onlinelibrary.wiley.com/doi/full/10.1002/jmv.25688

2. Fleming PL, Wortley PM, Karon JM, DeCock KM, Janssen RS. Tracking the HIV epidemic: Current issues, future challenges [Internet]. Am J Publ Health. 2000 [acceso: 25/07/2020];90(7):1037-41. Disponible en: https://www.pmc/articles/PMC1446284/?report=abstract

3. Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A. Epidemic processes in complex networks. Rev Mod Phys [Internet]. 2015 [acceso: 25/07/2020];87(3):925. Disponible en: https://journals.aps.org/rmp/abstract/10.1103/RevModPhys.87.925

4. Keeling MJ, Eames KT. Networks and epidemic models. J Roy Soc Interface [Internet]. 2005 [acceso: 25/07/2020];2(4):295-307. Disponible en: https://royalsocietypublishing.org/doi/10.1098/rsif.2005.0051

5. Brockmann D. Digital epidemiology. Bundesgesundh Gesundheits Gesundh. 2020;63(2):166-75.

6. Ladner JT, Grubaugh ND, Pybus OG, Andersen KG. Precision epidemiology for infectious disease control. Nat Med. 2019;25(2):206-11.

7. von Bubnoff A. Next-Generation Sequencing: The Race Is On. Cell Press. 2008;132:721-3.
8. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4(1):1.
9. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol [Internet]. 2018 [acceso: 25/07/2020];4(1). Disponible en: https://pubmed.ncbi.nlm.nih.gov/29942656/
10. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics [Internet]. 2018 [acceso: 10/04/2020];34(23):4121-3. Disponible en: https://academic.oup.com/bioinformatics/article/34/23/4121/5001388
11. Hadfield J, Brito AF, Swetnam DM, Vogels CBF, Tokarz RE, Andersen KG, et al. Twenty years of West Nile virus spread and evolution in the Americas visualized by Nextstrain [Internet]. PLoS Pathogens: Public Library of Science; 2019 [acceso: 25/07/2020]. p. e1008042. Disponible en: https://doi.org/10.1371/journal.ppat.1008042
12. Pearson WR. Finding Protein and Nucleotide Similarities with FASTA. Curr Protoc Bioinform [Internet]. 2016 [acceso: 26/07/2020];53(1):391-925. Disponible en: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi0309s53
13. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software: Improvements in Performance and Usability. Mol Biol Evol [Internet]. 2013;30(4):772-80. DOI: https://doi.org/10.1093/molbev/mst010
14. Lanave C, Preparata G, Sacone C, Serio G. A new method for calculating evolutionary substitution rates. J Mol Evol [Internet]. 1984 [acceso: 24/07/2020];20(1):86-93. Disponible en: https://link.springer.com/article/10.1007/BF02101990
15. To TH, Jung M, Lycett S, Gascuel O. Fast Dating Using Least-Squares Criteria and Algorithms. Syst Biol [Internet]. 2015;65(1):82-97. DOI: https://doi.org/10.1093/sysbio/syv068
16. Junqueira DM, Wilkinson E, Vallari A, Deng X, Achari A, Yu G, et al. New genomes from the Congo Basin Expand History of CRF01_AE Origin and

Dissemination. AIDS Res Hum Retrovir [Internet]. 2020 [acceso: 24/07/2020];36(7):574-82. Disponible en: https://www.liebertpub.com/doi/10.1089/aid.2020.0031

17. Billion A, Ghai R, Chakraborty T, Hain T. Augur - a computational pipeline for whole genome microbial surface protein prediction and classification. Bioinformatics [Internet]. 2006;22(22):2819-20. DOI: https://doi.org/10.1093/bioinformatics/btl466

18. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res [Internet]. 2002 [acceso: 25/07/2020];30(14):3059-66. Disponible en: https://pubmed.ncbi.nlm.nih.gov/12136088

19. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268-74.

20. Perkel J. Democratic databases: Science on GitHub. Nature [Internet]. 2016 [acceso: 25/07/2020];538(7623):127-8. Disponible en: http://www.nature.com/news/democratic-databases-science-on-github-1.20719

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

22. Seberg O, Petersen G. Assembling the Tree of Life [Internet]. Oxford University Press; 2006 [acceso: 26/07/2020]. p. 33-46. Disponible en: https://books.google.com.pe/books/about/Assembling_the_Tree_of_Life.html?id=6lXTP0YU6_kC&redir_esc=y

23. Han AX, Parker E, Scholer F, Maurer-Stroh S, Russell CA. Phylogenetic Clustering by Linear Integer Programming (PhyCLIP). Mol Biol Evol [Internet]. 2019;36(7):1580-95. DOI: https://doi.org/10.1093/molbev/msz053

24. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev [Internet]. 2020;7(6):1012-23. DOI: https://doi.org/10.1093/nsr/nwaa036

25. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic

epidemiology. Nat Microbiol [Internet]. 2020 [acceso: 25/07/2020];1-5. Disponible en: http://www.nature.com/articles/s41564-020-0770-5

26. Hodcroft EB, Hadfield J, Neher RA, Bedford T. Year-letter genetic clade naming for SARS-CoV-2 on nextstrain.org [Internet]. Nextstrain. 2020 [acceso: 25/07/2020]. Disponible en: https://nextstrain.org/blog/2020-06-02-SARSCoV2-clade-naming

27. Rife BD, Mavian C, Chen X, Ciccozzi M, Salemi M, Min J, et al. Phylodynamic applications in 21st century global infectious disease research. Glob Heal Res Policy. 2017;2(1):1-10.

28. Monteil V, Kwon H, Prado P, Hagelkrüys A, Wimmer RA, Stahl M, et al. Inhibition of SARS-CoV-2 Infections in Engineered Human Tissues Using Clinical-Grade Soluble Human ACE2. Cell. 2020;181(4):905-13.

29. Lee JM, Huddleston J, Doud MB, Hooper KA, Wu NC, Bedford T, et al. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. Proc Natl Acad Sci USA [Internet]. 2018 [acceso: 26/07/2020];115(35):E8276-85. Disponible en: https://www.pnas.org/content/115/35/E8276

30. Yamayoshi S, Kawaoka Y. Current and future influenza vaccines. Nat Med [Internet]. 2019 [acceso: 12/04/2020];25(2):212-20. Disponible en: http://www.nature.com/articles/s41591-018-0340-z

31. Dolan PT, Whitfield ZJ, Andino R. Mechanisms and Concepts in RNA Virus Population Dynamics and Evolution. Annu Rev Virol [Internet]. 2018 [acceso: 26/07/2020];5(1):69-92. Disponible en: https://www.annualreviews.org/doi/abs/10.1146/annurev-virology-101416-041718

32. van de Vossenberg BTLH, Visser M, Bruinsma M, Koenraadt HMS, Westenberg M, Botermans M. Real-time tracking of Tomato brown rugose fruit virus (ToBRFV) outbreaks in the Netherlands using Nextstrain. bioRxiv [Internet]. 2020 [acceso: 26/07/2020];06(02):129395. Disponible en: http://biorxiv.org/content/early/2020/06/02/2020.06.02.129395.abstract

33. Vega-Fernández J, Iglesias-Osores S, Tullume-Vergara P. Use of a bioinformatic tool for the molecular epidemiology of SARS-CoV-2. Univ Méd Pinar [Internet]. 2020 [acceso: 14/04/2020];16(3):3-5. Disponible en: http://revgaleno.sld.cu/index.php/ump/article/view/530

34. Wang JT, Lin YY, Chang SY, Yeh SH, Hu BH, Chen PJ, et al. The role of phylogenetic analysis in clarifying the infection source of a COVID-19 patient. J Infect. 2020;81(1):147-78.

35. Fauver JR, Petrone ME, Hodcroft EB, Shioda K, Ehrlich HY, Watts AG, et al. Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. Cell. 2020;181(5):990-6.

36. Singer JB, Thomson EC, McLauchlan J, Hughes J, Gifford RJ. GLUE: A flexible software system for virus sequence data. BMC Bioinformatics [Internet]. 2018 [acceso: 26/07/2020];19(1):1-18. Disponible en: https://link.springer.com/articles/10.1186/s12859-018-2459-9

37. Neher RA, Bedford T. Nextflu: real-time tracking of seasonal influenza virus evolution in humans. Bioinformatics [Internet]. 2015;31(21):3546-8. DOI: https://doi.org/10.1093/bioinformatics/btv381

38. Iglesias-Osores S, Iglesias-Osores S, Tullume-Vergara PO, Acosta-Quiroz J, Saavedra-Camacho JL, Rafael-Heredia A. Epidemiología genómica del virus SARS-CoV-2 con una plataforma bioinformática. Univ Méd Pinar [Internet]. 2020 [acceso: 25/07/2020];16(3):e555. Disponible en: http://revgaleno.sld.cu/index.php/ump/article/view/555

Conflicto de intereses

Se declara que no tienen conflicto de intereses.

Financiamiento

Autofinanciado.