ISSN-E: 1990-7990 RNPS:2008
Univ Méd Pinareña. September-December 2020; 16(3):e555

SPECIAL ARTICLE

Genomic epidemiology of SARS-CoV-2 virus with a bioinformatics platform

Epidemiología genómica del virus SARS-CoV-2 con una plataforma bioinformática

Sebastian Iglesias-Osores[1], Percy Omar Tullume-Vergara[2], Johana Acosta-Quiroz[1], Johnny Leandro Saavedra-Camacho[1], Arturo Rafael-Heredia[3]

[1]Universidad Nacional Pedro Ruiz Gallo. Facultad de Ciencias Biológicas. Lambayeque, Perú.
[2]Universidad de São Paulo. Instituto de Ciencias Biomédicas. São Paulo, Brasil.
[3]Universidad Nacional de Ucayali. Facultad de Medicina Humana. Ucayali, Perú.

Received: June 11, 2020 | **Accepted:** July 8, 2020 | **Published:** July 16, 2020

Cite as: Iglesias-Osores S, Tullume-Vergara PO, Acosta-Quirós J, Saavedra-Camacho JL, Rafael-Heredia A. Genomic epidemiology of SARS-CoV-2 virus with a bioinformatics platform. Univ Méd Pinareña [Internet]. 2020 [Cited: Access date];16(3):e555. Disponible en: http://revgaleno.sld.cu/index.php/ump/article/view/555

ABSTRACT

Introduction: viral epidemics have presented a risk to human health since they can turn into pandemics and affect a large part of the population, especially for poor developing countries. In 2020, the worldwide pandemic of COVID-19 is underway. Research is currently being carried out showing data that combines genetic and social information that can change our understanding of the dynamics of the epidemic.
Objective: to describe data science-based technology tool called Nextstrain that allows epidemics to be visualized with data as up to date as possible using academic databases.
Development: there are currently viral sequences from 57 countries on 6 continents. The common ancestor of the virus circulating in the world emerged in Wuhan, China, in late November or early December 2019, and from where it is supposed to have mutated towards humans, from bats and pangolins. Regarding monitoring, research work is already being carried out using this tool, such as in Taiwan, France, and Finland, which were able to determine where the SARS-CoV-2 strains that were causing outbreaks in their respective country originated. Besides, Nextstrain allows to freely share the phylogenetic analyzes of various authors from different countries and allows us to see the great work in the epidemiology of the virus.
Conclusions: Nextstrain is a tool based on big data that gives us a better view of the worldwide epidemiology of pathogens of interest. Its use is based on bioinformatic tools and it shows us this information through a pleasant and understandable ecosystem.

Keywords: Coronavirus Infections; Severe Acute Respiratory Syndrome; Betacoronavirus; Genome, Viral; Semantic Web.

RESUMEN

Introducción: las epidemias virales han demostrado ser un riesgo para la salud humana, ya que pueden convertirse en pandemias y afectar a gran parte de la población, especialmente a los países pobres en vías de desarrollo. En este año 2020 la pandemia mundial por la COVID-19 está en marcha. Actualmente se están llevando a cabo investigaciones que muestran datos que combinan tanto la información genética como la social, las cuales pueden cambiar nuestra comprensión acerca de la dinámica de la epidemia.
Objetivo: describir la herramienta tecnológica basada en la ciencia de datos, denominada Nextstrain, que permite visualizar las epidemias con los datos más actualizados en tiempo real, utilizando las bases de datos académicas.
Desarrollo: actualmente hay secuencias virales de 57 países en 6 continentes. El antepasado común del virus que circula en el mundo surgió en Wuhan (China), a finales de 2019, y desde donde se supone

CC-BY-NC 4.0

que ha mutado hacia los seres humanos a partir de murciélagos y pangolines. En lo que respecta al seguimiento, se han realizado trabajos de investigación con este instrumento en países como Taiwán, Francia y Finlandia, que han podido determinar el lugar de origen de las cepas del SARS-CoV-2 que causan los brotes en sus respectivos países. Nextstrain permite compartir libremente los análisis filogenéticos de varios autores de diferentes países, así como analizar el amplio trabajo realizado sobre la epidemiología del virus.
Conclusiones: Nextstrain es una herramienta creada a partir de grandes bases de datos que brinda una mejor visión acerca de patógenos de interés para la epidemiología mundial. Su uso se apoya en herramientas bioinformáticas y muestra esta información a través de un entorno agradable y comprensible.

Palabras clave: Infecciones por Coronavirus; Síndrome Respiratorio Agudo Grave; Betacoronavirus; Genoma Viral; Web Semántica

INTRODUCTION

The emerging viral epidemics represent a risk for human health, which may turn into pandemics affecting a large part of the population, especially in developing countries[1]. In 2019, a new coronavirus arose in Wuhan, province of Hubei, China, this virus is named SARS CoV-2, causing a severe acute respiratory syndrome, and spreading rapidly to other parts of China and currently in approximately 200 countries worldwide[2].

This coronavirus belongs to the Coronaviridae family ordered by Nidovirales, with a positive-sense single-stranded RNA and a genome of approximately 30 kilobases with a variable number of open reading frames, composed of 10 proteins[3]. The main form of transmission of the viruses is through direct contact from person to person, as well as through respiratory drops produced when an infected person touches the starling[4]. Seven types of coronaviruses affecting humans have been identified: HCoV-NL63, HCoV-229E, HCoV-OC43, HCoV-HKU1, SARS-CoV, MERS-CoV and SARS-CoV-2, can cause a severe respiratory syndrome in humans as well as four others causing mild illnesses of the upper respiratory tract[5].

Through phylogenetic analysis, it has been determined that SARS-CoV-2 is related to coronaviruses derived from bats that share an 87,99 % sequence identity with the SL-CoVZC45 and 87,23 % with the SLS-CoVZXC2, as well as SARS-CoV y MERS-CoV the 79 % and 50 % sequence identities respectively[6]. Nextstrain is developing innovative databases showing data that combine genetic, social information for a better understanding the dynamics of the epidemic, yet there is no cohesion in the isolated bell[7]. These data from different studies are not properly disseminated among the public in the academic and many times are not used in public policies of governments.

Various areas are in progress such as genomics and computing, transforming researchers' ability to respond to epidemic outbreaks in recent years[8]. Also, since next-generation sequencing technologies are more efficient and less costly than a few years ago, large-scale sequencing of entire genomes is becoming a tangible reality[9]. Therefore, this should go hand in hand with rapid and efficient dissemination of results in real-time. The speed in generating the genome sequences of various pathogens has vastly improved the ability to better understand past epidemics and generate public health interventions in a population[8].

A modern technology platform that we have of support based on data science is called Nextstrain (http://nextstrain.org) that allows visualizing the phylodynamic and evolutionary profile of epidemics with updated data in real-time using various public databases of virus genomic sequences through genomic epidemiology[10]. Nextstrain consists of data analysis and visualization components from public repositories: NCBI (National Center for Biotechnology Information) (www.ncbi.nlm.nih.gov), GISAID (Global Initiative Sharing All Influenza Data) (www.gisaid.org), ViPR (Virus Pathogen Database and Analysis Resource) (www.viprbrc.org), GitHub, and other sources in real-time. This tool performs phylodynamic analyzes, which can be used for epidemiological surveillance with molecular sequences, temporal and geographic origins, in addition to evolutionary history and ecological risk factors[11]. Nextstrain has built-in phylogenetic analysis tools such as TreeTime, which provides a tree where all clades are scaled so that the positions of the terminal nodes correspond to their sampling times, and the internal nodes are located at the most probable moment of divergence[12].

CC-BY-NC 4.0

ISSN-E: 1990-7990 RNPS:2008
Univ Méd Pinareña. September-December 2020; 16(3):e555

This article wants to describe data science-based technology tool called Nextstrain that allows epidemics to be visualized with data as up to date as possible using academic databases.

DEVELOPMENT

SARS-CoV-2 phylogenetic clade based temporal and spatial visualization tool

The phylogeny currently shown in Nextstrain, has the city of Wuhan, China as the origin of the new coronavirus SARS-CoV-2, possibly in November-December 2019. There is uncertainty regarding the date estimates and the reconstruction of the geographic spread since patterns are only a hypothesis. Site numbering and genome structure use Wuhan-Hu-1/2019 as a reference. Phylogeny originates from the earliest samples from Wuhan. The virus has spread to various countries of the world, currently reaching 188 regions and countries worldwide. In this brief review, we use Nextstrain to learn some relevant details about the spread of the SARS-CoV-2 virus in time and space in different countries of the world.

The Nextstrain platform has three panels to be able to extract data from the beginning of the replacement of the first genome to the last ones deposited until the current date of this revision. Among the functions that generate relevant information, it is to make groupings based on certain characteristics of the phylogenetic algorithm, in this case, clades, mutations, as well as other functions based on geography such as countries, regions, and less relevance by the author per laboratory, etc. For the following analysis, we use clustering based on clades generated through the Nextstrain phylogenetic algorithm. So far, five clades have been reported (19A, 19B, 20A, 20B, 20C) distributed in the different countries that have reported their viral genomes in public databases (Figure 1).

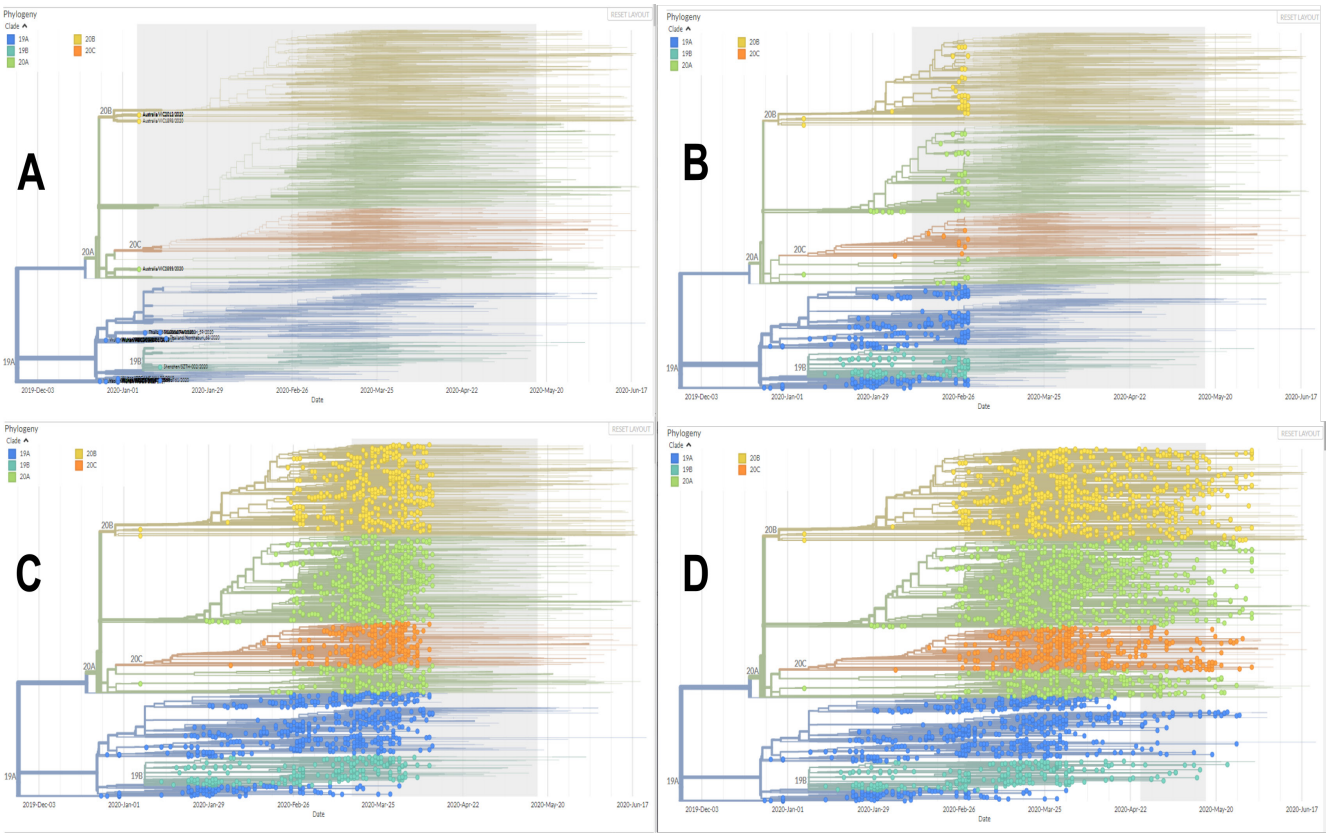

Figure 1. Phylogenetic tree of the expansion of the five clades reported on the Nexstrain platform. The images show how the virus originates in Wuhan, China (A), through Clade 19A, then clades 19B, 20A and 20C (B) appear, figure C shows how clade 20A begins to show dominance in the tree, finally D is almost the current scenario where viruses belonging to clades 20A, 20B and 20C are reported in most countries. Adapted from Nexstrain

CC-BY-NC 4.0

ISSN-E: 1990-7990 RNPS:2008
Univ Méd Pinareña. September-December 2020; 16(3):e555

The first genome reported in Nextstrain was the original 26/12/2019 (EPI_ISL 406798) of the city of Wuhan, China (Figure 1A), reported by Chen et al[13]. The chronological order in which the clades appeared is as follows, the first clade that is exhibited is the pioneer clade 19A, followed by the second appearance on 01/12/2020 reporting clade 19B. The third clade to appear is 20A on 02/03/2020 in England. Finally, the last clade to emerge was 20C dated 02 16 2020.

The distribution of the phylogenetically generated clades worldwide through Nextstrain (Figure 2) shows effectively that the most widespread clade worldwide is the 20A and 20B clades, this is according to the genomes that are on the platform. However, it gives us clear lights that, is the clade that has a greater world distribution at present.

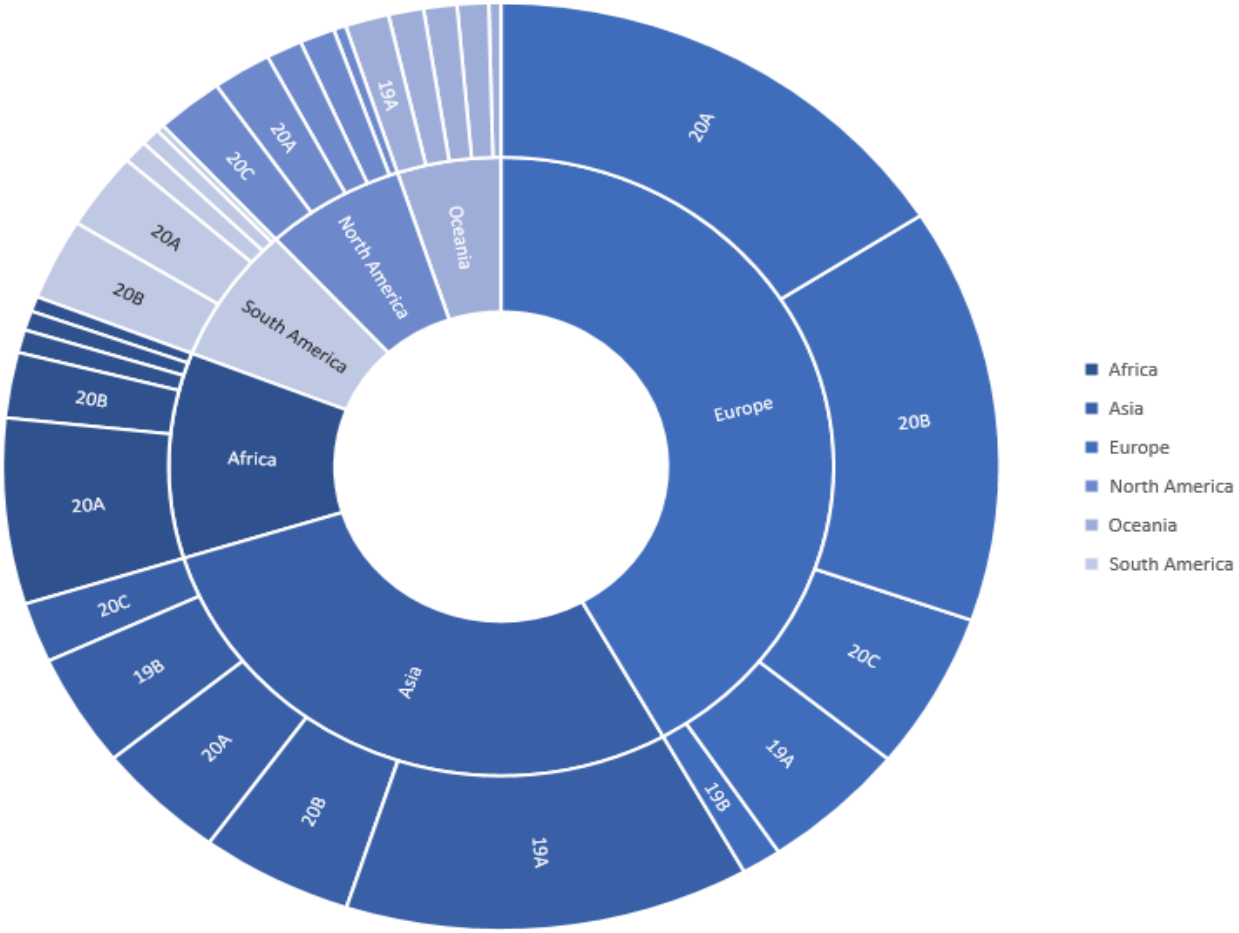

Figure 2. Distribution of phylogenetically generated clades of the SARS-CoV-2 virus across the continents.

Analysis through space and time, the virus originating in China begins its phase of expansion to Europe, neighboring Asian countries, and Australia (Figure 3A), then its expansion stage is to the United States where clades 19A and 19B are reported on the day 26/12/2019 and 13/01/2020 respectively. In China (Shenzhen), clade 19B is reported on 13/01/2020. In Europe, in England report clade 20A on 03/02/2020. England reported clade 20B on 16/02/2020. The last clade to appear is 20C on 16/02/2020, in Singapore as well.

Figures 1 and 3 show the events in time and space, generating a screenshot of four moments, of the different days of the pandemic worldwide on different dates. Although clade 19A has been the pioneer in Wuhan, clades 20A and 20b are practically are practically circulating in a large number of countries on the five continents (Figure 2), and according to the figure it is the clades with the highest spread, then it was introduced to Italy from where it has been disseminated to different countries in Europe, Asia and the United States (Figures 3B and 3C). Even though the other clades have also been reported in other

CC-BY-NC 4.0

ISSN-E: 1990-7990 RNPS:2008
Univ Méd Pinareña. September-December 2020; 16(3):e555

countries, their dissemination has not been as effective as that of clade 20A, also followed by clade 20B. The least distributed clade is 19B, it is interesting to know why it was not widely disseminated, reported in the United States, interestingly without mutations.

We also show the following from January to early February 2020, the second month of the epidemic, the appearance of introductions is observed beginning with individual cases in North America, Europe, and Oceania (Figure 3B). Since the spread of the SARS-CoV-2 virus began to the rest of the world, specifically in North America and Europe, the community transmission is evident in Europe, North America, and Oceania (Australia). Likewise, the creation of a European node where it spread throughout Europe in early February, in the third month the North American node is created. The spread to Latin America was from countries in Europe and North America and in addition to the African continent in March, Figure 3C. This explains the various intercontinental introductions during this period, indicating a mix of various viral lineages, shown in five final clades thus far, Figure 3D. Interestingly, figure 3 shows the total expansion from east to west, then south and north, this is due to the migratory movement of different tourists at the beginning of the pandemic.

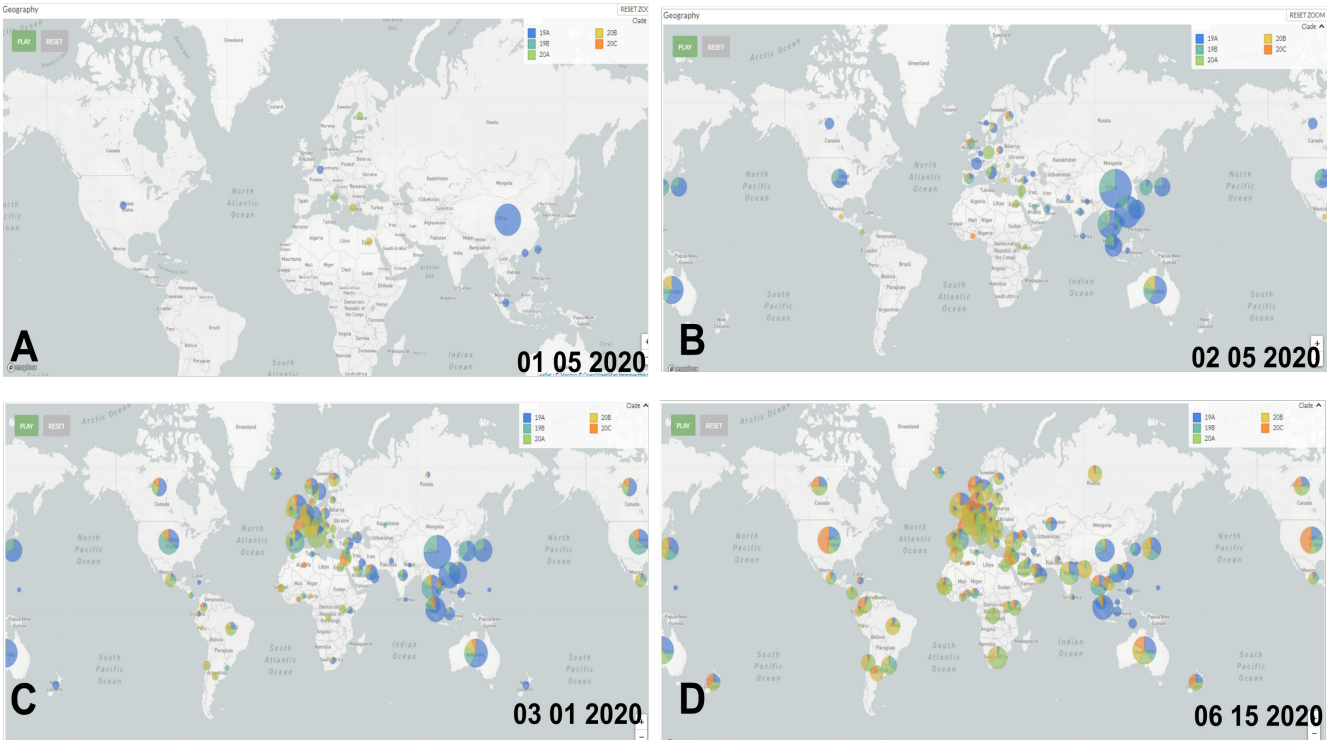

Figure 3. Expansion of the SARS-CoV-2 virus across the continents. This display by dates is done in Nextstrain by the date range function. Each image shows how the virus spread from its origin in China: (A) As of January 15, it begins to spread, entering Europe, Australia and countries close to China (B) Relevant is the introduction to the United States and new introductions to Europe (C) The various introductions of the virus in South America are beginning to be seen (D) The distribution of the virus is completely on all continents shown in a complex network. Adapted from Nextstrain.

Besides, we show a figure 4 with the genomes written down for each week since the start of the pandemic. The frequency distribution presents high bars on March and the beginning of April, where there are the dates when the genomes have been sent to the most that contribute to the release of information to face the novel coronavirus. United States and Spain, followed by China and the United Kingdom, which have sent the most sequences up to the moment of this analysis.

CC-BY-NC 4.0

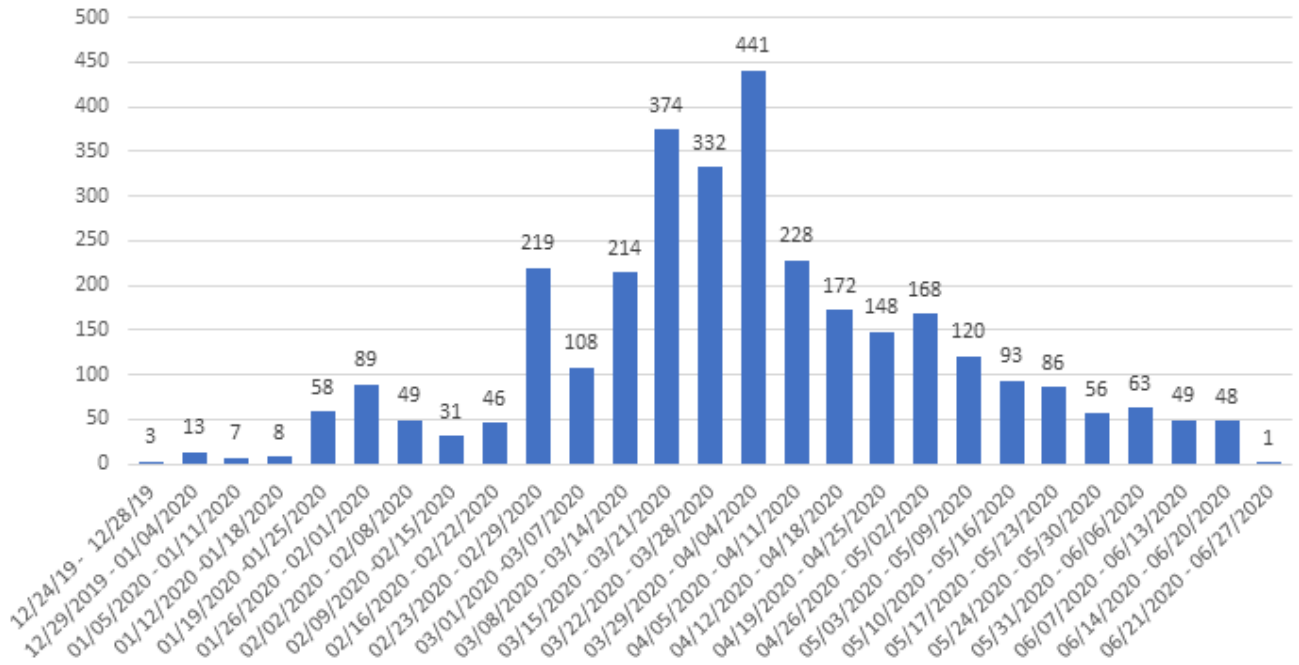

Figure 4. Distribution of the SARS-CoV-2 genomic sequences shown in the Nextstrain platform repository. The data is shown every week from the beginning of the genome deposits.

Surveillance and distribution of the SARS-COV-2 based on Nextstrain and literature review

The Nextstrain tool shows three main panels, being the phylogenetic tree panel, the geographical panel with the world map, and a panel showing the entropy of the virus in its entire sequence (30 kb). Each panel presents various manipulation options to be able to extract data according to the search strategy. Likewise, there is real-time monitoring of how mutations are occurring in different viruses from different regions and places, in addition to the generation of new clades within the phylogenetic tree. Through the formation of clades or groups, the origins of the viruses that were brought and introduced by travelers from different parts of the world can be associated.

Nextstrain-based association studies and supporting scientific literature are shown below. A study conducted in Taiwan, where a patient with the SARS-CoV-2 virus whose source of infection was unknown is presented. Later the source of infection was revealed with the help of phylogenetic analysis of the isolated virus, and together with data to help with epidemiological information, they indicated that the 66-year-old patient traveled to Dubai from January 29 to February 10, 2020, then to Egypt from February 11 to February 21, 2020. In Egypt, he participated in an eight-day sightseeing trip on a Nile River cruise. He returned to Taiwan via an international airline on February 21, 2020, from February 18, 2020, began to suffer from symptoms. Additionally, it was reported that 16 other people in the same tour group had similar symptoms at the time, showing positive results for the novel coronavirus.

Genome sequencing was performed to isolate the SARS-CoV-2 virus (NTU03) from a throat swab collected on March 2, 2020. The NTU03 sequence was grouped with clade 20A sequences showing 5 nucleotide differences, these included two synonymous mutations (between Orf 1b and 5410 and between orf 3a and 819), two non-synonymous mutations (between Orf 1b and 799 and between orf 3a and 57), and a mutation within the 5'UTR region. Likewise, an average of 12 nucleotide differences was observed between NTU03 and sequences within other clades. The phylogenetic analysis concludes that the NTU03 sequence belongs to clade 20A. Through the results obtained from the SARS-CoV-2 (NTU03 sequence), virus genome and its phylogenetic analysis, Taiwan patients, who tested positive for SARS-CoV-2 virus, spread the virus across Europe, supported by the information provided by the laboratories that submitted the virus sequences to the Global Influenza All Data Sharing Initiative (GISAID)[14].

CC-BY-NC 4.0

In another study, the first European cases of SARS-CoV-2 were analyzed, of which six genomes were associated with a group of transmissions in the French Alps in late January 2020. Of the six patients who tested positive for SARS -CoV-2, the three samples with the highest viral loads were selected for analysis. A nasopharyngeal swab was collected from a patient on February 7 (sample 1) and the other two samples were collected from the same asymptomatic patient on February 8 and 9 (sample 2 and sample 3). Comparison with the reference SARS-CoV-2 virus (NC_045512), it was possible to identify a deletion of three nucleotides in the open reading frame 1a (ORF1a) at genomic positions 1607 and 1609. It is important to highlight that this deletion was also identified in all the readings of sample 1 and 3. Using the CoV-GLUE resource, this mutation was found to lead to the deletion of amino acid 268 in nonstructural protein 2 (nsp2). This deletion in nsp2 (Asp 268) was also characterized in 37 of 571 complete genome sequences. Phylogenetic analysis based on the entire genome found 15 strains containing this specific deletion, they were relatively close to the strains isolated in China between December 2019 and the beginning of February 2020. While 23 strains with this deletion collected in the Netherlands have slightly diverged. However, longitudinal samples from the asymptomatic patient (Sample 2 versus Sample 3) were compared using a minimum depth of coverage of 100x to make a preliminary assessment of genetic variability within the host. Three SNVs are exhibited between the two samples: C366A (between nsp1 and S34Y), A20475G (synonymous mutation in nsp15), and T24084A (between protein S and L841H), suggesting the evolution of the virus within the host[15].

A third study of SARS-CoV-2 virus isolated from reported patient zero from Finland showed a nucleotide substitution C21707T compared to the reference strain from Wuhan China in December 2019 (NC_045512), which has led to a substitution of histidine to tyrosine (H49Y) in the N-terminal domain of the N domain of the spike glycoproteinor S protein .The genome sequence of the Finnish patient was almost identical to the Wuhan reference strain, indicating an early introduction from China[16].

Something to consider is the versatility to be able to work within Nextstrain and to be able to extract information from various mutations that occur in the different genomic sequences, they are synonymous and non-synonymous nucleotide and amino acid mutations, divergence rates, date and clade. Observing the various studies, the importance of this platform can be seen as an essential current tool to fight the SARS-Cov-2 pandemic, mainly in the study of geographic distribution, association studies in clades, rates of virus evolution through in time and space.

A model for epidemiological surveillance using genomic virus data

Nextstrain is a platform that mixes genetic data with epidemiological data to generate a dynamic visualization of the virus circulating in different regions at the beginning of epidemic or pandemic outbreaks. It uses public primary databases as support to generate evolutionary data through phylogeny and phylogenomic. Mutations that arose in the different genomic regions of the SARS-CoV-2 virus, such as mutations of the Spike protein, ORFa, ORFb and ORF14 in the different countries, and associating them with the pathogenicity and infectivity of the virus, from December 2019 to July 2020.

To know the distributions of SARS-CoV-2 sequences, throughout the world, we use the GISAID repository, which is the supporting database for Nexstrain.

As of July 7, 60,742 genomes had been sampled worldwide, with Europe being the continent that has reported the most genomes with 38,280 followed by North America (12,574) South America (1,057), Oceania (2,601), Asia (5,381) and Africa (836). In South America, Brazil is the one that reported the most with 630 genomes, followed by Chile with 167 and Colombia along with Uruguay with 126 and 74, respectively[10]. Therefore, Nexstrain and GISAID are useful for the scientific community, since it does not only allows us to see the distribution and evolution of SARS-CoV-2 in real-time but also helps to carry a report generated through scientific work done by country, by an institution and even in hosts the virus has been isolated. Data are shown in Figure 5.

CC-BY-NC 4.0

ISSN-E: 1990-7990 RNPS:2008
Univ Méd Pinareña. September-December 2020; 16(3):e555

Role of genomic epidemiology in Peru

At this time with no global presence, innovation in science plays an important role in tackling this global public health problem caused by the outbreak of the novel SARS-CoV-2 coronavirus. Therefore, countries with high budgets in science and technology are leading and dealing with high-impact research and producing large amounts of genomic data to understand the epidemiology and evolution of the virus globally in the short term. However, it is not a distant problem to countries in South America and Africa, where smaller amounts of genomic sequences of the SARS-CoV-2 virus are reported (Figure 5). Likewise, countries such as Venezuela and Peru are not oblivious to this reality, where only 2 and 1 genome, respectively, have been reported, deposited in GISAID, and displayed in Nextstrain.

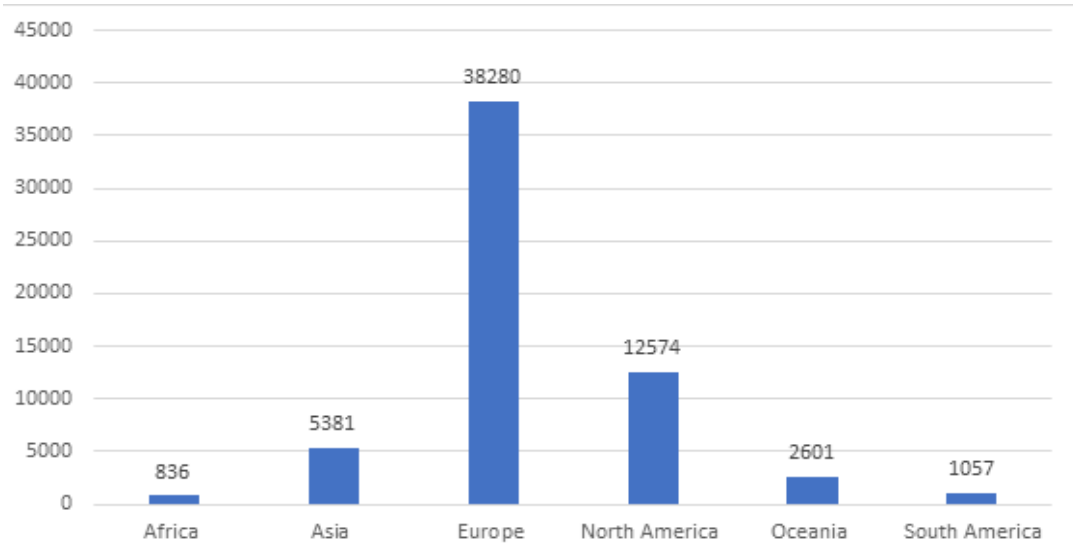

Figure 5. Number of genomes sequenced by continent to date

In the case of Peru, the reported data is insufficient to be able to obtain conjectures of the possible genotypes that are currently circulating in the different regions of Peru, in addition to studying possible unique mutations of strains that are circulating in the regions and associating them with their rate of lethality. Other questions to answer are how many introductions of the virus occurred in Peru, it is the same strain of the SARS-CoV-2 virus that is in all the regions with the highest incidence, such as departments such as Callao, Loreto, Piura, and Lambayeque. Although a SARS-CoV-2 strain was recently almost completely sequenced from an oropharyngeal swab from a Peruvian COVID-19 patient who had returned from Italy, the entire genome has 29,856 base pairs, the phylogenetic analysis indicates it is grouped in clade G (variant G), however, it is found in clade 20A according to the Nextstrain grouping, which coincides with other cases in South America, such as Brazil, Colombia, Chile, Argentina and Uruguay. Furthermore, some mutations in coding regions have been found that generated amino acid variations such as 1433P, P4720L, and D6909G in the polyprotein encoded by the orf1ab gene, also D614G in the S glycoprotein[17]. This introduction of the virus comes from Europe (Italy and Luxembourg) and has been expanding for now according to data in Lima, we need to deepen our efforts for the sequencing of genomes and to know what the current distribution of the virus is like in Peru and how it is distributed after almost six months.

Studies framed in this subject are of crucial impact to generate public health policies in each country, adapting their plans and fighting the pandemic in our region and the different countries of South America. Therefore, the Nextstrain platform for epidemiological surveillance is crucial during and after the pandemic to locate possible outbreaks in a period of intelligent isolation. After the quarantines have been removed in different places, the virus will continue to circulate with its various strains in various countries. According to reports, this would happen as long as we do not have the vaccine until possibly 2022, therefore, we insist on improving the genomic epidemiological surveillance systems to know and

CC-BY-NC 4.0

understand the transmission of the virus within the following years in both the South American region and especially in Peru.

Knowing the interior of the nextstrain platform

Nextstrain is a computer project with free access to different researchers, both bioinformatics and data science[10]. All the documentation is in the open-source repository called Github (https://github.com/), being the place of construction for software developers as well as data exchange for the scientific data community. Here in this database three repositories are the fundamental basis of the Nextstrain platform (https://github.com/nextstrain), being the following Sponsors, http://www.nextstrain.org and Augur.

The sponsor is the software to display an interactive visualizations of genomic cutting-edge data. The software can be downloaded and used locally on your computer or used through the website http://www.nextstrain.org. It is a software written in the java language and is in version 2 to be used through the following installation tutorial (https://nextstrain.github.io/auspice/introduction/how-to-run). Likewise, http://www.nextstrain.org is the website to help the public health and scientific community to obtain data on emerging pathogens such as the Dengue virus, influenza, SARS-CoV-2. The site is constantly updated on public display of data with analytical tools for community use.

The third software is Augur, which is the core part of the Nextstrain ecosystem for building phylogenetic trees pathogens. The source code is also in the Github repository. Augur provides a collection of commands to perform common bioinformatics tasks that are designed to be compostable in the pipeline for longer runs. Augur needs external tools for multiple sequence alignments of viral genomes using the MAFFT software (https://mafft.cbrc.jp/alignment/software/), in addition to the phylogenetic inference of the trees shown on the website, uses the maximum likelihood inference method, through the IQ-Tree software (http://www.iqtree.org/).

Augur outputs are a series of text files, which can be viewed through Auspicia. It presents its tutorial for consulting the different internal commands (https://nextstrain-augur.readthedocs.io/en/stable/usage/cli/cli.html).To face the pandemic, the platform created a new hcov-19 project available (http://nextstrain.org/ncov) to house the data from the SARS-CoV-2 virus genomics project. The data was shared and extracted via GISAD, which is submitted by the different authors and laboratories, on which this research is based. Here in this tool, we can analyze our information as input. We need two files: the genetic data in FASTA format and our metadata in TSV format. The table format consists of 23 columns of the genome data per sample. The team that forms Nextstrain is made up of six developers who keep the platform updated and above all that it is free without cost (open access). Furthermore, the platform uses three Python programming languages, shell, JavaScript, for its various programs contained in its pipeline.

The source and support that provides the genomic data

GISAID, is a German repository, which was created as an innovative political effort to promote the international exchange of genetic data and viruses associated with influenza[18]. However, this database in epidemic outbreaks and efficient way to share data generated by various laboratories on the 5 continents and to be useful to the scientific community and global health policies in this event of COVID-19. GISAID has played an important role since previous pandemics such as in 2008 in Indonesia caused by the H5N1 strain, then in 2009 by the H1N1 outbreak in the United States. Likewise, China reported in 2013 the H7N9 flu outbreak, sequenced the same day, and submitted it to the database.

GISAID is currently a great contribution to global health, highlighting the following: it is a collated repository of high-quality influenza sequences, facilitating the rapid exchange of virus data in the event of a possible pandemic, establishing trust in low-income countries is key to pandemic preparedness. Currently, the GISAID base has created a repository to host the genomic sequences of the SARS-CoV-2 virus that is currently causing a pandemic outbreak. The objective is to share in real-time the genomic sequences for research by the scientific community. Furthermore, which is the support for a Nextstrain

CC-BY-NC 4.0

interactive web server, GISAID currently contains 60 741 genomic virus sequences, of which 38 878 are of high coverage. The phylogenetic tree presented is based on markers that are variants such as S, O, L, V, G, GH, GR, and other clades. Also, it has metadata of the epidemiology of the sample, origin, date, sex, sequencing technology, among others.

Sequencing technologies used for the study of genomic epidemiology

The success of rapid detection, monitoring, and tracking of the novel SARS-CoV-2 virus in real-time worldwide, all this success has been thanks to the support, also of next-generation sequencing technologies (NGS). NGS, or also called parallel mass sequencing, has been developed in recent years and allows the sequencing of millions of DNA fragments. Currently, the use of NGS has almost completely replaced conventional Sanger sequencing and has many applications**(19)**. For example, these sequencing technologies are being used in the identification of new emerging viral pathogens such as recently in a Zika outbreak in Brazil[20].

Among the various platforms for genetic sequencing is Oxford Nanopore - MinION this device is a miniature sequencer that contains 512 nanopores, responsible for detecting single-stranded DNA, until now the use of this tool has focused on DNA sequencing, however it has been exploring its application in the analysis of RNA and proteins, it has been applied in sequencing and bacterial identification works, until the recent finding of its efficacy in detecting pathogens in plasma[21]. The Illumina HiSeq platform uses sequencing using synthesis technology in which fluorescent, market reversible terminator nucleotides bind to growing DNA strands and images are obtained through their excitation by fluorochrome at the point of attachment, this tool gives a true base-by-base sequence that practically eliminates errors[22].

Most of the SARS-CoV-2 virus sequencing projects have used the two technologies mentioned above, proof of which many genomes have high coverage or full coverage. It is important to have high-quality sequencing for bioinformatics and phylogenetic analysis such as substitution rate, virus variants, synonymous and non-synonymous mutations. Some patterns associated with virulence strains could be elucidated, so sequencing with a low error rate is important.

CONCLUSIONS

Nextstrain is a tool based on big data that gives us a better view of the worldwide epidemiology of pathogens of interest. It is shown that it is the support for different people both of researchers with computer skills, as well as those trained in aspects of public health and health policy. Furthermore, it is a source of continuous monitoring of the movement and outbreaks of the virus in various countries, with an emphasis on those with little scientific and technological capacity. Finally, Its use is based on bioinformatics tools and shows us this information through a pleasant and understandable ecosystem.

CONFLICT OF INTERESTS

All authors declare that they have no competing interests.

AUTHORSHIP CONTRIBUTION

SI was responsible for the conception of the study. SI, PT, JA, and JS collected the data. SI, PT, JA, and AR contributed to the data analysis. SI, PT, JA, JS, and AR wrote the initial draft with all authors providing critical comments and edits for later revisions. All authors approved the final draft of the manuscript.

FINANCING

The authors did not receive funding for the writing of this article

CC-BY-NC 4.0

BIBLIOGRAPHIC REFERENCES

1. Fong IW, Fong IW. Emerging Animal Coronaviruses: First SARS and Now MERS. In: Emerging Zoonoses [Internet]. Springer International Publishing; 2017 [cited 14/06/2020]. p. 63–80. Available from: https://dx.doi.org/10.1007/978-3-319-50890-0_4

2. Chan JFW, Kok KH, Zhu Z, Chu H, To KKW, Yuan S, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microbes Infect [Internet]. 2020 [cited 14/06/2020];9(1):221–36. Available from: https://doi.org/10.1080/22221751.2020.1719902

3. Kamel Boulos MN, Geraghty EM. Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: How 21st century GIS technologies are supporting the global fight against outbreaks and epidemics. Int J Health Geogr [Internet]. 2020 [cited 14/06/2020];19(1):8. Available from: https://doi.org/10.1186/s12942-020-00202-8

4. Burke RM, Midgley CM, Dratch A, Fenstersheib M, Haupt T, Holshue M, et al. Active Monitoring of Persons Exposed to Patients with Confirmed COVID-19 - United States, January-February 2020. MMWR Morb Mortal Wkly Rep [Internet]. 2020 [cited 14/06/2020];69(9):245–6. Available from: https://www.cdc.gov/mmwr/volumes/69/wr/mm6909e1.htm

5. Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, et al. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses [Internet].Trends Microbiol [Internet]. 2016 [cited 14/06/2020]. 24(6):490-502. Available from: https://doi.org/10.1016/j.tim.2016.03.003

6. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet [Internet]. 2020 [cited 14/06/2020];395(10224):565–74. Available from: https://doi.org/10.1016/S0140-6736(20)30251-8

7. Pybus OG, Fraser C, Rambaut A. Evolutionary epidemiology: preparing for an age of genomic plenty. Philos Trans R Soc B Biol Sci [Internet]. 2013 [cited 28/04/2020];368(1614):20120193. Available from: https://royalsocietypublishing.org/doi/10.1098/rstb.2012.0193

8. Ladner JT, Grubaugh ND, Pybus OG, Andersen KG. Precision epidemiology for infectious disease control. Nat Med [Internet]. 2019 [cited 14/06/2020];25(2):206–11. Available from: https://dx.doi.org/10.1038/s41591-019-0345-2

9. Von Bubnoff A. Next-Generation Sequencing: The Race Is On. Cell Press [Internet]; 2008 [cited 14/06/2020]. 132:721–3. Available from: https://doi.org/10.1016/j.cell.2008.02.028

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 [cited 10/04/2020];34(23):4121–3. Available from: https://academic.oup.com/bioinformatics/article/34/23/4121/5001388

11. 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 [Internet]. 2017 [cited 14/06/2020];2(1):1–10. Available from: https://dx.doi.org/10.1186/s41256-017-0034-y

12. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol [Internet]. 2018 [cited 14/06/2020]; 4(1):vex042. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29340210

CC-BY-NC 4.0

13. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet [Internet]. 2020 [cited 14/06/2020];395(10223):507-13. Available from: https://doi.org/10.1016/s0140-6736(20)30211-7

14. 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. Journal of Infection [Internet]. 2020 [cited 14/06/2020]; 81(1): 147-178. Available from: https://dx.doi.org/10.1016/j.jinf.2020.03.031

15. Bal A, Destras G, Gaymard A, Bouscambert-Duchamp M, Valette M, Escuret V, et al. Molecular characterization of SARS-CoV-2 in the first COVID-19 cluster in France reveals an amino-acid deletion in nsp2 (Asp268Del). Clinical microbiology and infection [Internet]. NLM (Medline); 2020 [cited 14/06/2020]; 26(7): 960-962. Available from: https://dx.doi.org/10.1016/j.cmi.2020.03.020

16. Haveri A, Smura T, Kuivanen S, Österlund P, Hepojoki J, Ikonen N, et al. Serological and molecular findings during SARS-CoV-2 infection: the first case study in Finland, January to February 2020. Eurosurveillance [Internet]. 2020 [cited 28/04/2020];25(11):2000266. Available from: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.11.2000266

17. Padilla-Rojas C, Lope-Pari P, Vega-Chozo K, Balbuena-Torres J, Caceres-Rey O, Bailon-Calderon H, et al. Near-Complete Genome Sequence of a 2019 Novel Coronavirus (SARS-CoV-2) Strain Causing a COVID-19 Case in Peru. Microbiol Resour Announc [Internet]. 2020 [cited 14/06/2020];9(19):[aprox.10 p]. Available from: https://dx.doi.org/10.1128/MRA.00303-20

18. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Challenges [Internet]. 2017 [cited 14/06/2020];1(1):33-46. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31565258

19. Kamps R, Brandão RD, van den Bosch BJ, Paulussen ADC, Xanthoulea S, Blok MJ, et al. Next-generation sequencing in oncology: Genetic diagnosis, risk prediction and cancer classification. International Journal of Molecular Sciences [Internet]. 2017 [cited 14/06/2020]; 18(2): 308. Available from: https://dx.doi.org/10.3390/ijms18020308

20. Kato K. Impact of the next generation DNA sequencers. Int J Clin Exp Med [Internet]. 2009 [cited 28/04/2020];2(2):193-202. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19684890

21. Chua EW, Ng PY. MinION: A novel tool for predicting drug hypersensitivity? Front Pharmacol [Internet]. 2016 [cited 14/06/2020];7(JUN):1-7. Available from: https://doi.org/10.3389/fphar.2016.00156

22. Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. Nat Rev Genet [Internet]. 2016 [cited 14/06/2020];17(6):333-51. Available from: https://doi.org/10.1038/nrg.2016.49

CC-BY-NC 4.0