





## Use of a bioinformatic tool for the molecular epidemiology of SARS-CoV-2

### Uso de herramientas bioinformáticas para la epidemiología molecular del SARS-CoV-2

Jorge Vega-Fernández<sup>1</sup> , Sebastián Iglesias-Osores<sup>1</sup>  , Percy Tullume-Vergara<sup>2</sup> 

<sup>1</sup>Universidad Nacional “Pedro Ruiz Gallo”. Facultad de Ciencias Biológicas. Lambayeque, Perú.

<sup>2</sup>Universidad de São Paulo. Instituto de Ciências Biomédicas. Brasil

**Received:** May 15, 2020 | **Accepted:** May 21, 2020 | **Published:** May 28, 2020

**Cited as:** Vega-Fernández JA, Iglesias Osores S, Tullume-Vergara PO. Use of a bioinformatic tool for the molecular epidemiology of SARS-CoV-2. Univ Méd Pinar [Internet]. 2021 [cited: Access date]; 17(1):e530. <http://www.revgaleno.sld.cu/index.php/ump/article/view/530>

#### Mr. director

According the article “Political will and intersectorality to confront COVID-19 in Cuba”, we need to find more recommendations to fight COVID-19<sup>(1)</sup>. The first sequencing of the SARS-CoV-2 virus began to be published and stored since December 2019. The data generated from genomic studies, such as sequencing, are the basis for studies of molecular epidemiology, population genetics, and studies on diversity genetics<sup>(2)</sup>.

The use of genomes from sequencing creates new areas to study the details of outbreaks such as molecular epidemiology (MS), is a hybrid of molecular biology within epidemiological studies. Aiming to establish the order of transmission of the pathogen, associate mutations related to its pathogenicity, patterns of transmission of pathogens, reservoirs and source of pathogens<sup>(3)</sup>.

Fundamental support for MS is data that comes from high-throughput sequencing technologies (NGS), in addition to bioinformatics techniques such as genomics and phylogenetics<sup>(4)</sup>. Likewise, databases of viral genomes, bioinformatic networks for phylodynamic analyzes, together with an interactive platform for real-time visualization of epidemic outbreaks. These databases are ready to be used to obtain information on various epidemiological aspects related to the virus. The objective of this letter is to present a tool that allows us to visualize and study epidemics in real-time using DNA sequencing of pathogens.

Nexstrain is a clear example of a current bioinformatics tool for MS, this tool allows the visualization of genomes and uses repositories of genomic sequences such as NCBI ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)), GISAID ([www.gisaid.org](http://www.gisaid.org)) and ViPR ([www.viprbrc.org](http://www.viprbrc.org)). The platform performs phylodynamic analysis in addition to other tools with phylogenetic inference, temporal dating of ancestral nodes and, probable transmission events<sup>(5)</sup>.

Various countries in the world and specifically in the South American region have deposited sequences of the SARS-CoV-2 virus in the mentioned databases and are used by Nexstrain for the elaboration of phylogenetic trees, as well as a map of their possible spread in the world. A descriptive analysis of the molecular epidemiology of SARS-CoV-2 was performed based on the data observed on the Nexstrain website. The Nexstrain database was accessed to narrate events regarding the possible route of the virus since its outbreak in Wuhan, China, settling in Europe and its introduction to South America.

The results indicate the following at the South American level: Brazil is the first country to reveal the first genome of the SARS-CoV-2 virus (accession EPI\_ISL\_412964) on 02/25/2020, indicating an introduction of the virus, from Lombardy, northern Italy. Followed by Argentina (accession EPI\_ISL\_415700) on 03/01/2020, Chile (accession EPI\_ISL\_414577) on 03/03/2020, which are associated with groups from

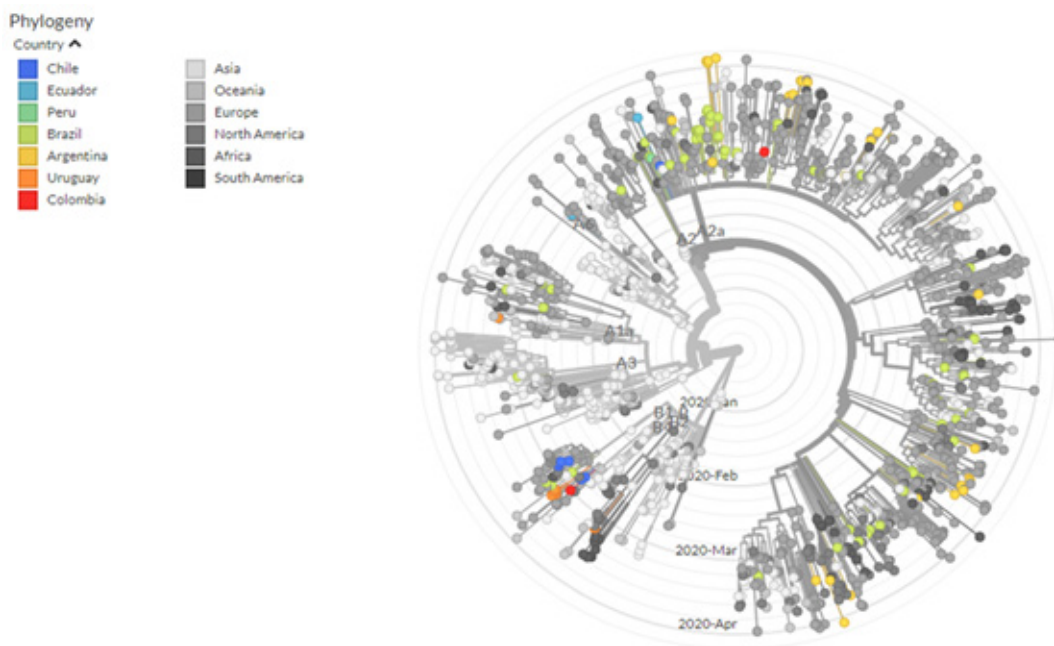
Wuhan, China and Taiwan(6); then Colombia (accession EPI\_ISL\_418262) on 03/06/2020, Peru (accession EPI\_ISL\_415787) on 03/10/2020, Uruguay (accession EPI\_ISL\_426479) on 03/19/2020, finally Ecuador (accession EPI\_ISL\_422563) revealed the virus genome on 03/30/2020 to the community.

In addition, we review the distribution of deposited genomes worldwide (Table 1). Europe exhibits 45 % of the total (788), followed by North America with 8,1 % (142), being South America (119) and Africa (89) with percentages 6,8 % and 5 % respectively, these being Minor locations represented within the global context. Likewise, we observed a significant representation of clades A2a and A1a for both South America and the various locations globally, indicating some mutations of the sequenced viruses (Figure 1).

**Table 1.** Distribution of annotated SARS-CoV-2 genomes in Nexstrain globally until May 15, 2020

Origin	Number of genomes	Clades
Europe	788	A2a, A1a
North-America	142	A2a, B1
Oceania	76	A2a, A1a
Asia	536	A2a, A1a, B
South America	119	A2a, A1a
Africa	89	A2a, A1a

We believe that it is important to focus efforts on knowing and identifying a greater number of SARS-CoV-2 samples that circulate in South American countries, especially those countries where there is only one deposited genome, as in the case of Peru. This will allow us to understand the dynamics of transmission of this and the control measures that can be taken to stop the spread of the infection and its possible readmissions. On the other hand, comparing its genomes allows us to determine the degree of existing genetic diversity, that is, how it is mutating, or if it obtained a recombination process, however, there is no mutation associated with its pathogenicity or virulence at the moment.



**Figure 1.** Phylogenetic tree of the representation of the SARS-CoV-2 virus genomes in South America. Obtained from Nexstrain.

In conclusion, Nexstrain is a tool that provides support for real-time molecular epidemiology surveillance studies, being of great help through a visualization platform for epidemiologists, virologists, and political decision-makers. Brazil has greater representation being the pioneer, however, South America is poorly represented, highlighting the European clades that were introduced. Further sequencing studies are recommended in patients from South America.

### CONFLICT OF INTERESTS

The authors declare that does not exist an interest conflicts

### AUTHORSHIP CONTRIBUTION

All the authors participated in the writing and review of the article; as well as its concluding version.

### FINANCING

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

### BIBLIOGRAPHIC REFERENCES

1. Leyva Caballero R. Voluntad política e intersectorialidad para el enfrentamiento a la COVID-19 en Cuba. *Rev. habanera. cien. méd.* [Internet]. 2020 [cited 2020 May 6];19(2):1-5. Available from: <http://www.revhabanera.sld.cu/index.php/rhab/article/view/3241/2507>
2. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, et al. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. *PLoS Med* [Internet]. 2013 [cited 2020 Apr 10];10(2). Available from: <https://doi.org/10.1371/journal.pmed.1001387>
3. Riapis LA, Filatov NN, Salova NI. Molecular epidemiology of infectious diseases. *ZhMikrobiol Epidemiol Immunobiol* [Internet]. 2003 [cited 2020 Apr 10];(5):49-54. Available from: <https://europepmc.org/article/med/14565124>
4. McNaughton AL, Roberts HE, Bonsall D, de Cesare M, Mokaya J, Lumley SF, et al. Illumina and Nanopore methods for whole genome sequencing of hepatitis B virus (HBV). *Sci Rep* [Internet]. 2019 [cited 2020 Apr 10];9(1):1-14. Available from: <https://doi.org/10.1038/s41598-019-43524-9>
5. 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 2020 Apr 10];34(23):4121-3. Available from: <https://academic.oup.com/bioinformatics/article/34/23/4121/5001388>