11-28-2021

2019nCoVR – A comprehensive genomic resource for SARS-CoV-2 variant surveillance and COVID-19 control

Guoquing Lu

Etsuko N. Moriyama

Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal Pre-proof

2019nCoVR – A comprehensive genomic resource for SARS-CoV-2 variant surveillance and COVID-19 control

Guoqing Lu, Etsuko N. Moriyama

PII: S2666-6758(21)00075-8

DOI: https://doi.org/10.1016/j.xinn.2021.100150

Reference: XINN 100150

To appear in: *The Innovation*

Please cite this article as: Lu, G., Moriyama, E.N., 2019nCoVR – A comprehensive genomic resource for SARS-CoV-2 variant surveillance and COVID-19 control, *The Innovation* (2021), doi: https://doi.org/10.1016/j.xinn.2021.100150.

2019nCoVR – A comprehensive genomic resource for SARS-CoV-2 variant surveillance and COVID-19 control

Guoqing Lu[1,*] and Etsuko N. Moriyama[2,*]

[1]Department of Biology, University of Nebraska at Omaha, Omaha, NE 68182, USA

[2]School of Biological Sciences and Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

*Correspondence: glu3@unomaha.edu (GL); emoriyama2@unl.edu (ENM)

The coronavirus disease 2019 (COVID-19) is a once-in-a-century pandemic, and the virus, the severe acute respiratory syndrome coronavirus 2 or SARS-CoV-2, has infected more than 180 million people and claimed almost 4 million lives worldwide (as of 5 July 2021) since the first case was reported in Wuhan, China, on 31 December 2019. The China National Center for Bioinformation (CNCB) responded promptly and launched the 2019 Novel Coronavirus Resource (2019nCoVR; https://ngdc.cncb.ac.cn/ncov/) in January 2020 for rapid release and public sharing of SARS-CoV-2 genomic data and analysis tools (**Figure 1A**).[1,2] The 2019nCoVR has quickly grown to be one of the most significant SARS-CoV-2 genomic resources, allowing users to explore the global landscape of genomic

variation and conduct genomic analysis and annotation.[3] The 2019nCoVR provides valuable information that helps understand molecular evolution and epidemiological dynamics of SARS-CoV-2, which can help inform decisions about controlling the spread of the virus. In this commentary, we highlight important features of the 2019nCoVR related to SARS-CoV-2 surveillance and comment on areas that will benefit future improvement of the resource.

Genomic sequence data are of paramount importance in epidemiology and play a vital role in understanding the transmission and evolution of SARS-CoV-2 and developing COVID-19 diagnostics, vaccines, and therapeutics. The release of the first genome of SARS-CoV-2 (10 January 2020) has enabled the development of vaccines and molecular testing tools. Genomic surveillance has been providing insights into regional and global establishment and lineage dynamics of the COVID-19 epidemic.[4] In addition to accepting direct submission, the 2019nCoVR incorporates sequence information from other resources, including GISAID (https://www.gisaid.org/) and NCBI GenBank (https://www.ncbi.nlm.nih.gov/sars-cov-2/). Whereas only a few sequences were available at the end of February 2020, the 2019nCoVR has collected data for more than 2 million complete genome sequences from 167 countries and regions by the end of June 2021, indicating an unprecedented speed in sequencing SARS-CoV-2 genomes. While several other major SARS-CoV-2 genomic resources exist as listed at the NU-COVID, http://bioinfolab.unl.edu/emlab/nucovid/, it should be noted that the 2019nCoVR has developed a set of standards for genomic data integrity and quality control.[3]

The 2019nCoVR offers multiple ways to explore and visualize SARS-CoV-2 genome variations. Based on sequence alignment and variation identification against the reference genome (MN908947.3), the 2019nCoVR identified and annotated over 28,900 nucleotide mutations that correspond to 21,324 amino acid changes (5 July 2021), shown in the histograms of isolate number *versus* the number of nucleotide or amino acid substitutions. Genomic variations can be easily inquired with multiple searching options such as region, collection date, and the range of SNP numbers. Spatiotemporal dynamics of the

SARS-CoV-2 variations can be inspected through the heat maps across the time and countries, with many filter options including variant frequency, genes/regions, mutation types, and transcriptional regulation sites. For example, the variation dynamic curve of the D614G variant (S protein; nucleotide position: 23,403) demonstrated this variant emerged in February and early March 2020 and gradually became dominant, particularly in Europe and North America (**Figure 1B**), likely attributed to higher transmissibility of this variant. The comparison of variation dynamic curves clearly showed a difference in the accumulation of this mutation among countries. The most updated data can be obtained from: https://bigd.big.ac.cn/ncov/variation/annotation/variant/23403?lang=en.

The 2019nCoVR adopts the Pango lineage assignment established by Rambaut et al.[5] for all sequences in the database. In the Lineage Browser, the distributions of sampling dates and countries, as well as variants for each lineage, are summarized in interactive charts and tables. For example, the sublineage B.1.1.7 (WHO label: Alpha), the lineage emerging and extensively circulating in the UK in December 2020, is shown to have spread worldwide and peaked at the end of March 2021, then diminished dramatically by the end of June 2021 (**Figure 1C**). In contrast, the recently emerging Delta variant (Pango lineage B.1.617.2) shows a very different pattern in its temporal dynamics, i.e., the number of sampled viral isolates increasing much later (**Figure 1D**). Users, including virologists and policymakers, can also examine the evolutionary relationships and dynamics of SARS-CoV-2 by exploring phylogenetic trees and haplotype networks. In the Viral Haplotype Network, for example, the progression of the viral haplotype networks can be traced temporally as well as spatially using animation.

The 2019nCoVR has developed and made available a number of COVID-19 related resources and tools. It provides pipelines and tools for SARS-CoV-2 genome assembly, variation identification, variant and genome annotations. *De novo Assembly* allows assembling raw sequencing reads, estimating sequencing depth, and comparing the assembled contigs to the SARS-CoV-2 reference genome. The *Fastq-to-Variants* web tool can be used to align sequencing reads to the reference genome, detect SNPs

and Indels, and annotate them. With the *Variant Annotation* tool, the users can perform functional annotation of the mutations and display mutation patterns and effects. The COVID-19 pandemic has caught the massive attention of the scientific communities, and the number of scientific publications has been increasing exponentially. The literature resource in the 2019nCoVR has achieved over one hundred thousand entries, including research articles, preprints, letters, editorials, etc.

With many useful resources and tools, navigating throughout the menus and submenus to discover all contents available in the 2019nCoVR is at times tedious. Providing easy-to-follow user manuals and tutorials for many tools and resources and cross-linking information from different resources could enhance user experience with the 2019nCoVR. There are many appealing visual presentations of SARS-CoV-2 statistics. However, image rendering sometimes takes a long loading time (e.g., haplotype network dynamics for all lineages) or does not refresh promptly (e.g., linage browse), dampening the user experience. This might be an area for future improvement. It would also be helpful to make the images downloadable and of high resolution suitable for publication. While incorporating clinical data into the 2019nCoVR resource is a welcome feature, only a limited number of records were available in an earlier version. At the time of this review, the link to the Clinic and CT Image seems to be broken. For the lineage classification, the 2019nCoVR only uses the Pango lineage assignment. It could be helpful if it is cross-referenced with other viral assignments, such as those from WHO and Nextstrain (https://nextstrain.org/). Overall, the 2019nCoVR is a comprehensive and integrated SARS-CoV-2 genomic analysis platform, with many useful and practical features for analysis and annotation of SARS-CoV-2 genomes and COVID-19 epidemiological dynamics. It engages more communities through rapid data sharing in the fight against COVID-19, a global public health catastrophe.

REFERENCES

1. Gong, Z., Zhu, J.W., Li, C.P., et al. (2020). An online coronavirus analysis platform from the National Genomics Data Center. Zoological Research ***41***, 705.
2. Zhao, W. M., Song, S. H., Chen, M. L., et al. (2020). The 2019 novel coronavirus resource. Yi Chuan ***42***, 212-221.
3. Song, S., Ma, L., Zou, D., et al. (2020). The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019nCoVR. Genomics, Proteomics & Bioinformatics. https://doi.org/10.1016/j.gpb.2020.09.001.
4. du Plessis, L., McCrone, J. T., Zarebski, A. E., et al. (2021). Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science ***371***, 708-712.
5. Rambaut, A., Holmes, E. C., O'Toole, et al. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature Microbiology ***5***, 1403-1407.

ACKNOWLEDGMENTS

We thank Dr. Y. Bao and the staff at the CNCB (China National Center for Bioinformation) and NGDC (National Genomics Data Center) for answering all our inquiries. We want to thank Dr. A. Voshall for his assistance in developing the NU-COVID website (http://bioinfolab.unl.edu/emlab/nucovid/). GL acknowledges his Comparative Genomics class for insightful discussions on the 2019nCoVR platform. This work has been partially supported by grants from the University of Nebraska-Lincoln Research Council Interdisciplinary Research Grant (ENM) and the University of Nebraska Collaboration Initiative Grant (GL and ENM). The authors are grateful to the editorial office for outstanding graphical assistance and editing support.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Figure 1. Example screenshots of the 2019nCoVR. (A) The 2019nCoVR home page (https://ngdc.cncb.ac.cn/ncov/?lang=en) showing available resources and tools. (B) Variation dynamic curves of the D614G variant (genomic position 23,403) in the spike protein circulating among countries. (C) Sample distribution across different dates for the lineage B.1.1.7 (Alpha variant). (D) Sample distribution across different dates for the lineage B.1.617.2 (Delta variant). The images are slightly modified for clarity.