2019nCoVR—A comprehensive genomic resource for SARS-CoV-2 variant surveillance

Guoqing Lu¹,* and Etsuko N. Moriyama²,*

¹Department of Biology, University of Nebraska at Omaha, Omaha, NE 68182, USA
²School of Biological Sciences and Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

*Correspondence: glu3@unomaha.edu (G.L.); emoriyama2@unl.edu (E.N.M.)

Received: January 29, 2021; Accepted: August 7, 2021; Published Online: August 12, 2021; https://doi.org/10.1016/j.xinn.2021.100150
© 2021 The Author(s). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Citation: Lu G. and Moriyama E.N. (2021). 2019nCoVR—A comprehensive genomic resource for SARS-CoV-2 variant surveillance. The Innovation 2(4), 100150.

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 July 5, 2021) since the first case was reported in Wuhan, China, on December 31, 2019. The China National Center for Bioinformation (CNCB) responded promptly and launched the 2019 Novel Coronavirus Resource (2019nCoVR, https://ngdc.cnbc.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 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 2019nCoVR provides valuable information that helps understand the 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 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 (January 10, 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, 2019nCoVR incorporates sequence information from other resources, including GISAID (https://www.gisaid.org/) and the NCBI GenBank (https://www.ncbi.nlm.nih.gov/sars-cov-2/). Whereas only a few sequences were available at the end of February 2020, 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 NUCoVID, http://bioinfosl.net.unl.edu/emuclab/nucovid/, it should be noted that 2019nCoVR has developed a set of standards for genomic data integrity and quality control.³

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), 2019nCoVR identified and annotated over 28,900 nucleotide mutations that correspond to 21,324 amino acid changes (July 5, 2021), shown in the histograms of isolate numbers versus the number of nucleotide or amino acid substitutions. Genomic variations can be easily investigated with multiple searching options such as region, collection date, and the range of SNP (single nucleotide polymorphism) numbers. Spatiotemporal dynamics of the SARS-CoV-2 variations can be inspected through the heatmaps across time and countries, with many filter options including mutation frequency, genes/regions, mutation types, and transcriptional regulation sites. For example, the variation dynamic curve of the D614G mutation (S protein; nucleotide position: 23,403) demonstrated that this mutation emerged in February and early March 2020 and gradually became dominant, particularly in Europe and North America (Figure 1B), likely attributed to the higher transmissibility of this mutation. 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.

2019nCoVR adopts the Pango lineage assignment established by Rambaut et al.⁶ 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 that had emerged and extensively circulated 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 delta variant (Pango lineage B.1.617.2), the variant more recently emerged, shows a very different pattern in its temporal dynamics, i.e., whose number of sampled viral isolates has started 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.

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, and 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 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 massive attention from the scientific communities, and the number of scientific publications has been increasing exponentially. The literature resource in 2019nCoVR has achieved over 100,000 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 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 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., image browse), damping 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. For lineage classification, 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, 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. Zoolog. Res. 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. https://doi.org/10.16288/j.yczyzz.20-030.
3. Song, S., Ma, L., Zou, D., et al. (2020). The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019nCoV. Genomics Proteomics Bioinformatics. https://doi.org/10.1016/j.gpb.2020.09.001.
4. du Plessis, L., McDrone, 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, Â., et al. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. 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/). G.L. 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 (to E.N.M.) and the University of Nebraska Collaboration Initiative Grant (to G.L. and E.N.M.). The authors are grateful to the editorial office for outstanding graphical assistance and editing support.

DECLARATION OF INTERESTS

The authors declare no competing interests.