frontiers | Frontiers in Microbiology

TYPE Opinion
PUBLISHED 04 November 2022
DOI 10.3389/fmicb.2022.1020148

OPEN ACCESS

EDITED BY
Maria Alejandra Mussi,
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

REVIEWED BY
Leandro Souza,
Central Public Health Laboratory of Rio de Janeiro Noel Nutels (LACEN-RJ), Brazil

*CORRESPONDENCE
Asif M. Khan
asif@perdanauniversity.edu.my;
makhan@bezmialem.edu.tr

†PRESENT ADDRESS
Li Chuin Chong,
Institute for Experimental Virology, TWINCORE Centre for Experimental and Clinical Infection Research, a joint venture between Medical School Hannover (MHH) and Helmholtz Centre for Infection Research (HZI), Hannover, Germany

SPECIALTY SECTION
This article was submitted to Infectious Agents and Disease, a section of the journal Frontiers in Microbiology

RECEIVED 17 August 2022
ACCEPTED 13 October 2022
PUBLISHED 04 November 2022

CITATION
Chong LC and Khan AM (2022) Historical milestone in 42 years of viral sequencing—Impetus for a community-driven sequencing of global priority pathogens. *Front. Microbiol.* 13:1020148. doi: 10.3389/fmicb.2022.1020148

Historical milestone in 42 years of viral sequencing—Impetus for a community-driven sequencing of global priority pathogens

Li Chuin Chong[1†] and Asif M. Khan[1,2*]

[1]Beykoz Institute of Life Sciences and Biotechnology, Bezmialem Vakif University, Istanbul, Turkey, [2]School of Data Sciences, Perdana University, Kuala Lumpur, Malaysia

KEYWORDS

community, virus, SARS-CoV-2, health policy, monkeypox, sequencing, pathogen

Introduction

Sequence data are critical for the design of effective intervention (vaccines, drugs, and diagnostics) and surveillance strategies against pathogens. This need is more than ever exemplified by the ongoing global scientific effort against the COVID-19 pandemic, where the resulting sequence data of the disease agent, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is currently the largest ever for a given virus.

Trend and historical milestone in coronavirus sequencing

As many as 13,376,049 genome sequences, translating to 351,835,857 protein records (as of 4 October 2022) have been deposited to the specialist GISAID EpiCoV™ repository (Khare et al., 2021), the single largest open-access platform for SARS-CoV-2 sequence data, surpassing that of NCBI databases (NCBI Resource Coordinators, 2018), a long-standing primary sequence resource. GISAID, formally launched in 2008, has been able to encourage greater international sharing of viral sequence data through innovative policies and offerings that recognize the contributions and interests of data providers and users alike (Khare et al., 2021), starting with influenza virus. These include: (i) a trusted sharing mechanism framework that guarantees that data users will acknowledge the contributions of, and make efforts to collaborate with, data generators; (ii) a high-throughput submission portal; (iii) high quality data standards through review and curation in real-time and annotation by a global team of curators, prior to release, for all submitted data; (iv) enhancement of curated data with computed results; and (v) delivery for downstream analyses *via* customisable data feeds. Consequently, GISAID was in a position to serve as a critical blueprint for SARS-CoV-2 data deposition and real-time analyses, and is thus expected to be well-poised for future pandemic challenges.

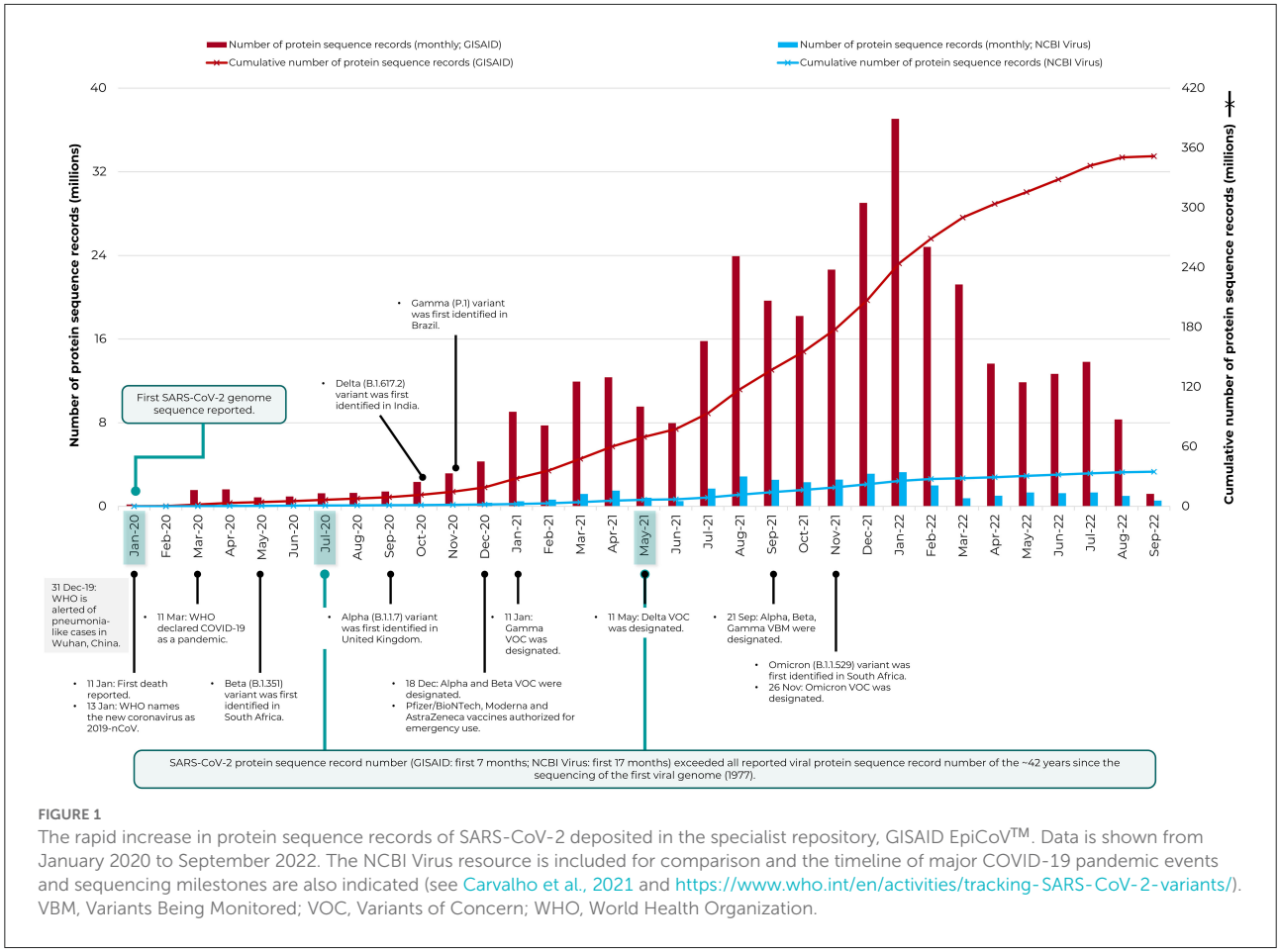

FIGURE 1
The rapid increase in protein sequence records of SARS-CoV-2 deposited in the specialist repository, GISAID EpiCoV™. Data is shown from January 2020 to September 2022. The NCBI Virus resource is included for comparison and the timeline of major COVID-19 pandemic events and sequencing milestones are also indicated (see Carvalho et al., 2021 and https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/). VBM, Variants Being Monitored; VOC, Variants of Concern; WHO, World Health Organization.

Figure 1 depicts the number of protein sequence records available for SARS-CoV-2 in the specialist databases GISAID EpiCoV™ and NCBI Virus (NCBI Resource Coordinators, 2018), with collection date starting from January 2020, before the onset of the pandemic to 33 months later, September 2022. The records in GISAID grew rapidly from ~0.02 million to ~351 million in less than three years, a number strikingly ~10-fold higher than NCBI Virus (up to ~35 million). This unprecedented, rapid collection of sequence data (GISAID) had edged past that of HIV-1 by March 2020, for which the greatest number of viral protein records had been reported, till before the pandemic (>1 million as of December 2019 in NCBI Virus; the species SARS-related coronavirus then just had only 4.4K records).

The concerted worldwide sequencing effort for SARS-CoV-2 has remarkably etched a historical milestone in virus sequencing, as the fastest and the single largest source of viral sequence data growth. The protein record data of SARS-CoV-2 (collected by first 7 months since January 2020 for GISAID or first 17 months for NCBI Virus) had exceeded the total number of protein records made available for all reported viruses since the 1977 sequencing of the first viral DNA genome, bacteriophage phi X 174 (or ΦX174) to December 2019, a period of ~42 years. There were then 5,948,418 protein records in NCBI Virus for all reported viruses (data not shown), released as of 31 December 2019. Likewise, the nucleotide record data of SARS-CoV-2 [collected by first 19 months for GISAID since January 2020 (https://gisaid.org/hcov-19-variants-dashboard) or first 22 months for NCBI Virus (data not shown)] had exceeded the number of such publicly available records of 42 years for all reported viruses (3,225,425 records in NCBI Virus as of 31 December 2019; data not shown). Notably, the GISAID data was largely of full-length genomes, whereas the prior data (of all viruses) was a mix of full-length genomes and partial sequences.

Need for a community-driven sequencing of global priority pathogens

The coronavirus sequencing historical milestone hails the value of a worldwide, community-driven pathogen sequencing effort. It heralds a global, real-time surveillance and rapid intervention response against infectious diseases. Thus, there is

a need to call for action to replicate this sequencing success as a model for other pathogens, starting with the long available list of global priority pathogens (NIAID, 2021) and perhaps with primacy given to viruses due to their smaller genome size. There is much that could be learnt from the experiences of GISAID (restricted open sharing), the International Nucleotide Sequence Database Collaboration (INSDC; promotes full open sharing), and other advocates of pathogen genomics (Black et al., 2020), where despite the differences in approach (Van Noorden, 2021), closing the global genome data gap (including metadata) for these pathogens is the collective priority (Mallapaty, 2022). Urgently needed is a global and an open platform for pathogen genomics, in the form of a consortium or an initiative and particularly involving scientific community members from affected countries, especially those of low- and middle-income countries (LMICs), with support from the vanguards of community-driven sequencing. This endeavor may possibly be sustained through assessed and voluntary funding contributions. Efforts that are already underway, such as the various national, international, and continent-wide pathogen sequencing and readiness consortiums/initiatives (Holmes et al., 2017; Makoni, 2020; NIAID, 2021; Harvard Medical School, 2022) can be adapted to expand globally (Illumina Inc., 2021). A bottom-up community-driven, open approach can be greatly complementary to the established efforts that are top-down and central, often initiated or led by large organizations, such as the CEPI (Brende et al., 2017) and Global Virome Project (Carroll et al., 2018), among others. The recipe for success is available, global collaborations are at an all-time high, and the pressing need is to rapidly model the successful coronavirus sequencing effort to all pathogens of priority first and of interest later. Granted that the model has its shortcomings, enabling global sequencing of clinical and environmental pathogen isolates is a critical first step toward the development of effective intervention and surveillance strategies.

Discussion

COVID-19 remains a vivid reminder that a single isolated outbreak is capable of rapidly bringing the world to a grinding halt for a reasonable duration of time, crippling humanity with loss of lives and battered health systems. This happened in the scientifically advanced 21st century. Thus, the new normal of the post-COVID-19 era needs to maintain the momentum and advance the progress that has been made over the catastrophic sacrifice. Remaining blind spot gaps will need to be identified and addressed. A collective community action is a necessary catalyst that can help uniformly strengthen the flailing, porous fabric of our global preparedness. The recent surge of human monkeypox cases, 72,428 confirmed in 102 geographical locations where the disease is not typically reported (as of 14 October 2022) (CDC, 2022), and the concern of fitness selection as the virus variants find more hosts for continuous evolution, is a poignant reminder.

Author contributions

AK developed the initial concepts for this paper and reviewed the writing. LC contributed to the data curation, formal analysis, visualization, and writing—original draft. Both authors approved the final version and had final responsibility for the decision to submit for publication.

Funding

AK was supported by Perdana University, Malaysia, Bezmialem Vakif University, Turkey, and the Scientific and Technological Research Council of Turkey (TÜBITAK). This publication/paper has been produced benefiting from the 2232 International Fellowship for Outstanding Researchers Program of TÜBITAK (Project No: 118C314). However, the entire responsibility of the publication/paper belongs to the owner of the publication/paper. The financial support received from TÜBITAK does not mean that the content of the publication is approved in a scientific sense by TÜBITAK.

Acknowledgments

We gratefully acknowledge the authors from the originating and submitting laboratories for the sequences deposited to GISAID's EpiCoV™ database.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Author disclaimer

The authors express that they are in no way connected to GISAID, besides being user of its various databases.

References

Black, A., MacCannell, D. R., Sibley, T. R., and Bedford, T. (2020). Ten recommendations for supporting open pathogen genomic analysis in public health. *Nat. Med.* 26, 832–841. doi: 10.1038/s41591-020-0935-z

Brende, B., Farrar, J., Gashumba, D., Moedas, C., Mundel, T., Shiozaki, Y., et al. (2017). CEPI—a new global R&D organisation for epidemic preparedness and response. *Lancet* 389, 233–235. doi: 10.1016/S0140-6736(17)30131-9

Carroll, D., Daszak, P., Wolfe, N. D., Gao, G. F., Morel, C. M., Morzaria, S., et al. (2018). The global virome project. *Sci. Adv.* 359, 872–874. doi: 10.1126/science.aap7463

Carvalho, T., Krammer, F., and Iwasaki, A. (2021). The first 12 months of COVID-19: A timeline of immunological insights. *Nat. Rev. Immunol.* 21, 245–256. doi: 10.1038/s41577-021-00522-1

CDC (2022). *2022 Monkeypox Outbreak Global Map*. Available online at: https://www.cdc.gov/poxvirus/monkeypox/response/2022/world-map.html (accessed October 4, 2022).

Harvard Medical School (2022). *Massachusetts Consortium on Pathogen Readiness*. Harvard Medical School. Available online at: https://masscpr.hms.harvard.edu (accessed February 10, 2022).

Holmes, K. K., Bertozzi, S., Bloom, B. R., Jha, P., Gelband, H., DeMaria, L. M., et al. (2017). "Major infectious diseases: key messages from disease control priorities," in *Disease Control Priorities, Third Edition (Volume 6): Major Infectious Diseases*, eds K. K. Holmes, S. Bertozzi, B. R. Bloom, and P. Jha (Washington, DC: The World Bank), 1–27. doi: 10.1596/978-1-4648-0524-0_ch1

Illumina Inc. (2021). *Illumina to Donate US $60 Million in Sequencing Capabilities to Establish a Global Pathogen Genomics Initiative*. Available online at: https://www.illumina.com/company/news-center/press-releases/press-release-details.html?newsid=04ae3ebc-4f23-4d5e-884a-4432e0de00a4 (accessed February 11, 2022).

Khare, S., Gurry, C., Freitas, L., Schultz, M. B., Bach, G., Diallo, A., et al., (2021). GISAID's role in pandemic response. *China CDC Wkly.* 3, 1049–1051. doi: 10.46234/ccdcw2021.255

Makoni, M. (2020). Africa's $100-million pathogen genomics initiative. *Lancet Microbe* 1, e318. doi: 10.1016/S2666-5247(20)30206-8

Mallapaty, S. (2022). Genome data gaps could stymie search for next COVID variant. *Nature*. doi: 10.1038/d41586-022-00894-x

NCBI Resource Coordinators (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 46, D8–13. doi: 10.1093/nar/gkx1095

NIAID (2021). *NIAID Pandemic Prepareness Plan*. Available online at: https://www.niaid.nih.gov/sites/default/files/pandemic-preparedness-plan.pdf (accessed February 15, 2022).

Van Noorden, R. (2021). Scientists call for fully open sharing of coronavirus genome data. *Nature* 590, 195–196. doi: 10.1038/d41586-021-00305-7