Impact of data access models on the detection of SARS-CoV-2 variants of concern

Lucy Hui[1,2], Art Poon[1,3]

[1] Department of Pathology and Laboratory Medicine, [2] Department of Physiology and Pharmacology, [3] Department of Microbiology and Immunology, Schulich School of Medicine & Dentistry, Western University, Canada

INTRODUCTION

One of the unique developments of the global SARS-CoV-2 pandemic has been the rapid generation and dissemination of virus genome sequence data[1]. To date, nearly 15 million SARS-CoV-2 genomes have been published through the Global Initiative for Sharing All Influenza Data (GISAID) database. These data have played a central role in tracking the spread of this virus, and the identification of "variants of concern" (VoCs) associated with rapid outbreaks[2].

However, there has been controversy around GISAID's model for data access, which is controlled by a user registration system and Terms of Use that restricts the re-distribution of data or derived results[3]. As a result, some researchers have advocated for more open models of data sharing. The Nextstrain development team curated an open data feed of SARS-CoV-2 genome records that have been deposited in the NCBI GenBank database[2]. Although the Nextstrain open data feed comprises a substantial number of genomes, there are considerable differences in the temporal and geographical distribution of samples covered by the databases, which impact downstream analyses.

OBJECTIVES

- Examine the impact of different data access models on the time until emerging VoCs could be detected from genomic data..
- Examine the global distribution of SARS-CoV-2 samples covered by different data access models.

METHODS

GISAID N = 10.6 million *v.s.* **Nextstrain** N = 4.7 million

Data frame processing

VoC	Collection Date	Submission Date	Country
Alpha	⋯		
⋮			
Beta			
Gamma			
Delta			
Omicron			

Publication delay

Publication delay = Submission date – Collection date

Detection of VoCs

- Simulate limited data access during the SARS-CoV-2 pandemic by constraining data frames based on monthly submission date cut-offs starting from October 2020 to May 2022.
- Logistic regression conducted to determine the earliest dates at which a significant differences in relative growth rates could be detected for each emerging VoC.

Geographical distribution

- Proportion of genome submissions contributed by the top nine participating countries were determined

PUBLICATION DELAY

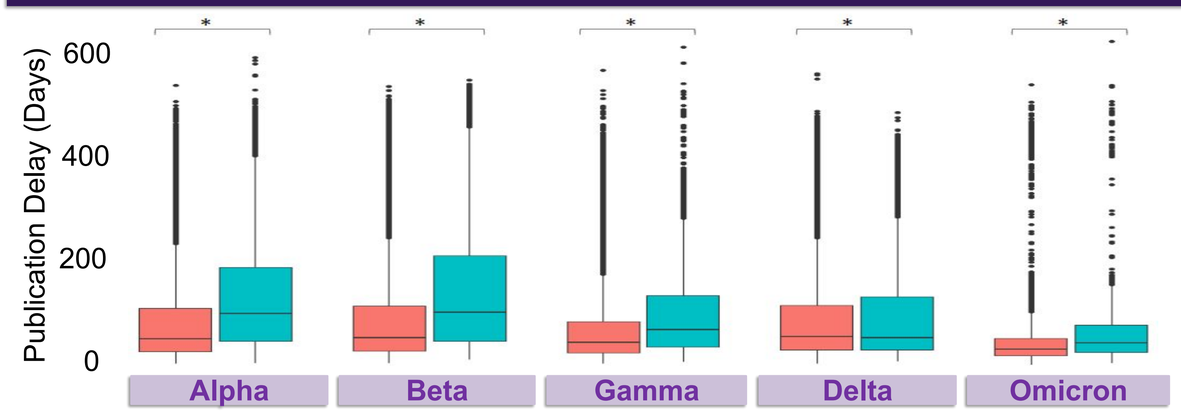

Figure 1. *Publication delay of SARS-CoV-2 VoCs in **GISAID** and **Nextstrain**. The average publication delay is significantly shorter in GISAID compared to Nextstrain across all VoCs. The asterisks denote statistical significance at $p < 0.001$.*

DETECTION OF VoCs

	GISAID	Nextstrain
Alpha	December 2020	February 2021
Beta	December 2020	March 2021
Gamma	February 2021	March 2021
Delta	April 2021	October 2021
Omicron	December 2021	March 2022

Table 1*. Logistic regression of SARS-CoV-2 variants of concerns (alpha, beta, gamma, delta, omicron) estimate and significant submission cut-off dates in GISAID and Nextstrain. All significant submission cut-off dates were earlier in GISAID compared to Nextstrain (Mean = 3.2 months, Standard deviation = 1.9 months, Range = 1 to 6 months).*

GEOGRAPHICAL DISTRIBUTION

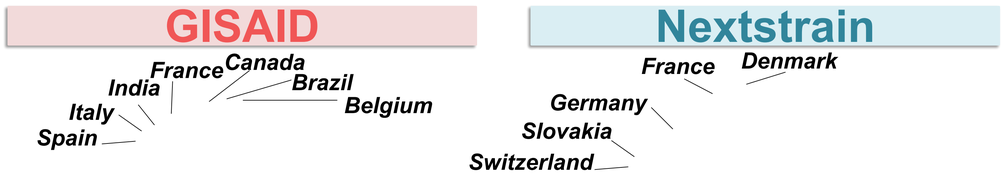

Figure 2. *Proportion of SARS-CoV-2 genome submissions by country in GISAID and Nextstrain. GISAID contains a greater number of submissions per country in contrast to Nextstrain. The geographical distribution of genome submissions is more balanced in GISAID compared to Nextstrain.*

CONCLUSIONS

- In comparison to GISAID, the detection of VoCs was significantly delayed using Nextstrain. The geographical distribution of SARS-CoV-2 samples was also limited as opposed to GISAID.
- These difference may be attributed to a multitude of factors such as hesitance to submit sequence data into open data feeds and a delay in the establishment of Nextstrain as a data access model for SARS-CoV-2 after the maturation of GISAID.
- In consideration of public health efforts to implement timely emergency responses, GISAID is the recommended data access model for timely detection of SARS-CoV-2 VoCs.

REFERENCES

1. Bernasconi, A., Canakoglu, A., Masseroli, M., Pinoli, P., & Ceri, S. (2021). A review on viral data sources and search systems for perspective mitigation of COVID-19. *Briefings in bioinformatics*, 22(2), 664–675.
2. Bedford, T., Hadfield, J., Hodcoft, E., Huddlestone. J., Neher, R., & Sibley, T. (2021, July 8) Extension of SARS-CoV-2 data processing to incorporate Open Data through GenBank. *Nextstrain*. H
3. Wadman, M. (2021, March 10). Critics decry access, transparency issues with key trove of coronavirus sequences. *ScienceInsider.*

ACKNOWLEDGEMENTS

I would like to thank Dr. Art Poon for his ongoing support and guidance throughout this research study.