

Tracking SARS-CoV-2 variants and resources

Bas B. Oude Munnink & Marion Koopmans

 Check for updates

Outbreak.info empowers real-time variant monitoring and tracing of associated publications and resources during the ‘infodemic’ of SARS-CoV-2.

During the SARS-CoV-2 pandemic researchers urgently generated and shared information about the SARS-CoV-2 virus, genomes and other data such as preprints, publications, datasets, protocols, images, computational tools and clinical trial data. However, keeping track of the overwhelming number of new sequences and the growing number of data sources has become a tremendous task. For instance, a query of ‘SARS-CoV-2’ in the PubMed search engine results in over 187,000 hits, more than 14 million sequences have been shared on GISAID¹, as of 19 December 2022, and over 5.5 million raw sequence datasets have been shared on the European Nucleotide Archive (ENA) through the COVID-19 Data Portal (as of 1 December 2022)². In the current issue of *Nature Methods*, a Resource³ and Brief Communication⁴ present outbreak.info, which can be used to track and trace SARS-CoV-2 sequence variants based on current classification systems or on specific mutations in the viral genome³. In addition, outbreak.info assembles and unifies various data resources to enable researchers to quickly search through the latest research, using an interface that allows subdividing the search by category⁴. An overview by virus or variant with a specific mutation can be generated, as can an overview or a comparison of the characteristic mutations (in case of a variant). Outbreak.info provides a summary of the global prevalence, the daily prevalence in given parts of the world, the prevalence during a particular time period, and publications and resources for more information regarding the variant or particular mutation. The combination of different resources for queries is one of the real strengths of outbreak.info (see Fig. 1 for a schematic overview).

One presented case study⁴ shows how publications on a particular variant rapidly increase after the initial detection and that the scientific research response with regards to publications, including preprints, was more rapid following the emergence of Omicron than the emergence of Alpha. This case study also shows that research on variants lags behind the spread of these variants, as it takes time to perform fundamental studies on, for instance, the mechanism of action of virus entry. This results in publications of observations on viruses that are no longer detected in global surveillance. The other example presented³ provides an overview of the prevalence of different variants of concern over time and by geographical region during the current pandemic. Using this approach, the emergence and evolution of different lineages can be monitored on a global scale; a selection can also be made to monitor (for instance) the lineage distribution of different Omicron variants that have been circulating in a particular country in the past 2 months, which is convenient for researchers interested in the situation in their own country.

Several applications have been developed during the pandemic to track SARS-CoV-2 variants globally, enabled by the rapid, almost real-time sharing of genomic information by different institutions

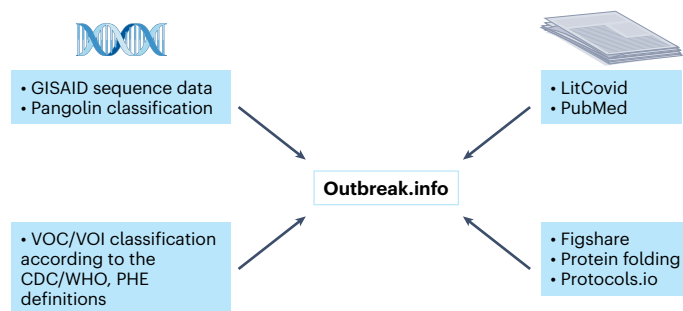


Fig. 1 | Overview of the combination of different resources that outbreak.info compiles on their website. CDC, Centers for Disease Control and Prevention; PHE, Public Health England; VOC, variant of concern; VOI, variant of interest.

worldwide. In addition to outbreak.info, websites such as the COVID-19 Data Portal², Nextstrain⁵ and CoV-Spectrum⁶ all keep track of the genomic diversity and distribution of SARS-CoV-2 lineages. All of them have their strengths, but the unique feature of outbreak.info is that it is largely automated and that it combines and centralizes different data streams. Nonetheless, it needs to be emphasized that the quality of information depends on the underlying data that are provided; although the overviews provided are intuitive and informative, they are also currently biased. In general, genomic surveillance efforts are very much biased toward the Western world⁷ and coverage is rapidly decreasing. Different genomic and other surveillance strategies were applied in different countries during the pandemic, which calls for caution when comparing data provided in overviews generated by outbreak.info throughout the entire pandemic. For instance, sometimes a prescreening was performed using a variant or mutation-specific real-time PCR or samples were selected on the basis of recent travel history, thus sampling a subgroup that is not representative of the general community. In principle, these metadata are all fields that can be entered during data submission to GISAID and that are essential for proper interpretation of the data. Unfortunately, although genomic sequences are shared, most of the time these metadata are missing owing to privacy concerns, stigmatization or other reasons. This is an issue that severely reduces the utility of the global genomic data resources, and that should be resolved to meaningfully use or reuse sequences that have been submitted to GISAID. Moving forward, the World Health Organization (WHO) is encouraging member states to continue SARS-CoV-2 surveillance, which in the longer term will most probably fall under the governance of public-health institutes and include a mechanism for the annotation of findings. However, it remains to be seen whether the spirit of real-time sharing will continue to be part of this surveillance.


As the authors also rightfully point out in both studies^{3,4}, the current scientific reward system is well-established for scientific manuscripts but less so for datasets and the researchers who produce them. Although there is wide agreement about the importance of sharing

data (and of the combined analysis of all available data), there is some debate as to the best ways of doing so. GISAID encourages people to share their data⁸ but the user agreement does not allow researchers to perform a combined analysis and to subsequently publish on all of the data present in the database; some researchers have therefore urged for fully open data sharing to open up the repository for analytical studies beyond the core team of GISAID⁹. Another recurring issue is what the benefit is for data providers, other than contributing to the public good. Downstream applications such as outbreak.info should be encouraged, but we should also find ways of acknowledging data providers to ensure that researchers continue to submit their data to semi-open or open data repositories.

As an example, the ENA has now started to provide data DOIs for SARS-CoV-2 studies to make them citable. In addition, datasets can be claimed using ORCID iDs. When combined with quality criteria (regarding the completeness of a metadata file), this could incentivize sharing and result in higher-quality datasets. These data DOIs could be extended to different data sources (such as image objects and clinical trials) and these could then be combined to obtain a more comprehensive overview on outbreak.info, even though the data sources might not all be submitted to the same database or use the same metadata template.

Bas B. Oude Munnink   & **Marion Koopmans**  

Viroscience Department, Erasmus Medical Center, Rotterdam, the Netherlands.

 e-mail: b.oudemunnink@erasmusmc.nl; m.koopmans@erasmusmc.nl

Published online: 15 March 2023

References

1. Shu, Y. & McCauley, J. *Euro Surveill.* **22**, pii=30494 (2017).
2. Harrison, P. W. et al. *Nucleic Acids Res.* **49**, W619–W623 (2021).
3. Gangavarapu, K. et al. *Nat. Methods* <https://doi.org/10.1038/s41592-023-01769-3> (2023).
4. Tsueng, G. et al. *Nat. Methods* <https://doi.org/10.1038/s41592-023-01770-w> (2023).
5. Hadfield, J. et al. *Bioinformatics* **34**, 4121–4123 (2018).
6. Chen, C. et al. *Bioinformatics* **38**, 1735–1737 (2022).
7. Brito, A. F. et al. *Nat. Commun.* **13**, 7003 (2022).
8. Maxmen, A. *Nature* **593**, 176–177 (2021).
9. Van Noorden, R. *Nature* **590**, 195–196 (2021).

Acknowledgements

The authors are financed by the NWO Stevin Prize by the Netherlands Organisation for Scientific Research (NWO) and the European Union's Horizon 2020 research and innovation programme Versatile Emerging infectious disease Observatory (VEO) under grant number 874735.

Competing interests

The authors declare no competing interests.