Bioinformatics, 38(6), 2022, 1735–1737
https://doi.org/10.1093/bioinformatics/btab856
Advance Access Publication Date: 25 December 2021
Applications Note

Sequence analysis

CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants

Chaoran Chen [1,2,*], Sarah Nadeau[1,2], Michael Yared[3], Philippe Voinov[3], Ning Xie[4], Cornelius Roemer[2,5] and Tanja Stadler[1,2,*]

[1]Department of Biosystems Science and Engineering, ETH Zürich, CH-4058 Basel, Switzerland, [2]Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland, [3]Department of Computer Science, ETH Zürich, CH-8092 Zürich, Switzerland, [4]Department of Informatics, University of Zurich, CH-8050 Zürich, Switzerland and [5]Biozentrum, University of Basel, CH-4056 Basel, Switzerland

*To whom correspondence should be addressed.
Associate Editor: Can Alkan

Received on June 11, 2021; revised on November 26, 2021; editorial decision on December 17, 2021; accepted on December 21, 2021

Abstract

Summary: The CoV-Spectrum website supports the identification of new SARS-CoV-2 variants of concern and the tracking of known variants. Its flexible amino acid and nucleotide mutation search allows querying of variants before they are designated by a lineage nomenclature system. The platform brings together SARS-CoV-2 data from different sources and applies analyses. Results include the proportion of different variants over time, their demographic and geographic distributions, common mutations, hospitalization and mortality probabilities, estimates for transmission fitness advantage and insights obtained from wastewater samples.

Availability and implementation: CoV-Spectrum is available at https://cov-spectrum.org. The code is released under the GPL-3.0 license at https://github.com/cevo-public/cov-spectrum-website.

Contact: chaoran.chen@bsse.ethz.ch or tanja.stadler@bsse.ethz.ch

1 Introduction

Most mutations in the SARS-CoV-2 genome do not cause phenotypic changes in the virus. However, some mutations may change the virus such that it is (i) more transmissible, (ii) causes a more severe disease outcome or (iii) has the ability to evade immunity after infection or vaccination. A new variant with one of these properties is classified as a variant of concern (VOC; World Health Organization, 2021). It is crucial to rapidly identify and characterize new variants of concern such that public health measures can be adapted to emerging threats.

Demonstrating that one of the VOC properties (i)–(iii) is met by a new variant in real time is an ongoing challenge for public health. In particular, observing a quickly spreading variant does not necessarily imply a transmission advantage. In fact, there are many factors that can influence a variant's observed spread: geographic biases and different interventions in different regions, demographic biases and different behaviors in different populations, varying contact tracing efforts and varying test regimes, among others. For example, the variant named 20E (EU1) (B.1.177) spread across Europe in summer 2020. However, rather than having a transmission advantage, it appears that superspreading events and travel activities drove this spread (Hodcroft *et al.*, 2021). Therefore, it is important to investigate a broad set of data from different regions to evaluate the risk posed by new variants. This was done for the variant Alpha (B.1.1.7) which showed a consistent relative growth in many countries, indicating an intrinsic transmission advantage of 43–90% (Davies *et al.*, 2021a).

The CoV-Spectrum website aims to help track known VOCs and facilitate early identification of new ones. It brings together the global public dataset of genomic sequences and additional epidemiological data (Section 2.1) to provide a multifaceted view of a variant. The website's variant search feature allows users to track combinations of amino acid and nucleotide mutations, in addition to already designated lineages.

2 Materials and methods

2.1 Data sources and data presentation

The primary data presented by CoV-Spectrum are genomic sequences. We currently provide two instances of CoV-Spectrum: one that uses data provided by GISAID (Elbe and Buckland-Merrett, 2017), and another one that uses data from NCBI GenBank provided through Nextstrain (Hadfield *et al.*, 2018). These are whole genome sequences of SARS-CoV-2 from countries across the globe as well as basic metadata such as the sampling date, location (often at the level of national divisions) and, for some sequences, the age and sex of the infected individual. We clean the location data with Nextstrain's geo location rules (https://github.com/nextstrain/ncov-ingest/blob/master/source-data/gisaid_geoLocationRules.tsv) and we run Nextclade

(Aksamentov *et al.*, 2021) to obtain aligned nucleotide and amino acid sequences. These data are updated daily.

CoV-Spectrum uses this data to create plots summarizing the raw data and to perform statistical analyses. The plots include the prevalence, estimated number of cases, demographic and geographic distributions and the common mutations of a variant. Some of the plots can further be stratified by geographic divisions. These are presented in a grid, enabling the user to visually check whether the same dynamic is present in different divisions.

For Switzerland, we receive additional metadata from the Swiss Federal Office of Public Health. The metadata is linked to the whole genome sequences and includes, e.g. additional demographic, hospitalization and mortality information. CoV-Spectrum uses this unique dataset to compute the hospitalization and mortality probabilities for different age groups and shows a plot that compares the hospitalization and mortality probabilities of confirmed cases infected with a selected variant with other variants. This enables direct assessment of VOC property (ii) (severe outcome). With this feature, we can see that the hospitalization probability of cases infected with the Alpha variant is indeed higher in older age groups, as suggested by other studies (Challen *et al.*, 2021; Davies *et al.*, 2021b).

2.2 Statistical analysis

In addition to presenting the raw data, CoV-Spectrum applies statistical analyses to them. For instance, first, CoV-Spectrum shows the mutations that occur in sequences of a variant and, by ranking them by their Jaccard similarity, it helps identify the mutations that are specific to a particular variant.

Second, CoV-Spectrum integrates a model to estimate variant transmission fitness advantages, as described in Chen *et al.* (2021). Chen *et al.* (2021) presents static results for the Alpha variant in Switzerland, while CoV-Spectrum allows users to explore results for any variants and countries. This enables assessment of VOC property (i) (increased transmissibility). For Switzerland, we additionally receive estimates of the proportion of different variants in wastewater samples from collaborators. The underlying procedure is described in Jahn *et al.* (2021), and is currently applied to a selection of variants for which characteristic mutations are manually chosen. This allows to assess if mutations that are identified as spreading in the population based on clinical data is confirmed in wastewater data.

2.3 Linking other services

Many COVID-19 dashboards and web tools have been developed since the start of the pandemic, each with their own specific use case. CoV-Spectrum can serve as a hub between several of these services. Namely, the website directly integrates external services so that users can obtain more information about selected variants at the click of a button. For example, CoV-Spectrum can send a list of sequence identifiers to UShER (Turakhia *et al.*, 2021), which will then place the sequences on a predefined tree. It can also redirect the user to Taxonium (Sanderson, 2021), which highlights the selected variant in a precomputed global tree with millions of nodes. Finally, it links to CoVariants (Hodcroft, 2021), which provides users with curated information about a variant.

2.4 Sharing of results

To promote dissemination of real-time results, CoV-Spectrum's plots and tables are made available to external websites via iframes. These plots remain interactive and will be automatically updated as new data arrives. We used this technique, e.g. to integrate plots into a dedicated website explaining the spread of the Alpha variant in Switzerland (https://cevo-public.github.io/Quantification-of-the-spread-of-a-SARS-CoV-2-variant/).

2.5 Implementation

The frontend of CoV-Spectrum is a single-page React application written in TypeScript. It retrieves data from two REST APIs. First, CoV-Spectrum's own server application provides the non-sequence data. Then, our Lightweight API for Sequences (LAPIS; Chen and Stadler, 2021) provides the sequence data. LAPIS is a general API to query sequences that is maintained as a separate project. The servers are written in Kotlin and Java using the Spring Boot framework. Finally, the data are stored in a PostgreSQL database.

3 Conclusion

The CoV-Spectrum website facilitates rapid detection and characterization of circulating SARS-CoV-2 variants around the globe. The website offers users a convenient way to assess the available SARS-CoV-2 sequencing data together with its metadata. It provides rich information by providing timely figures and tables produced based on globally shared data. As mentioned, evaluating variants requires careful consideration of potential biases in the sequencing data. CoV-Spectrum aims to help with this task by providing the appropriate geographic and demographic context, wherever possible. Users should, however, be aware of possible sampling biases in the raw data, which may carry through to results presented on CoV-Spectrum. Thus, any results should be interpreted and communicated accordingly.

CoV-Spectrum is only possible due to the ongoing efforts of the international community to perform sequencing and make the data rapidly and openly available on GISAID and GenBank. However, some crucial tasks like assessing VOC properties (ii) (severe outcome) and (iii) (immune/vaccine breakthrough) require additional metadata, such as the severity of infections or vaccine status (Gomez *et al.*, 2021). We call for global sharing of such metadata. The sharing of properly anonymized and aggregated data will facilitate the rapid identification of VOCs. This will be crucial for timely global public health responses.

Acknowledgements

CoV-Spectrum is enabled by data from GISAID and open data. We acknowledge the GISAID team and all originating and submitting labs. Further, we acknowledge the Federal Office of Public Health in Switzerland for providing metadata for Swiss sequences.

Data availability

CoV-Spectrum uses public datasets provided by GISAID (https://www.gisaid.org/), GenBank (https://www.ncbi.nlm.nih.gov/genbank/) and Our World in Data (https://ourworldindata.org/). It further uses non-public data provided by the Federal Office of Public Health in Switzerland.

Funding

This work was supported by the Swiss National Science foundation [Special Call on Coronaviruses; 31CA30 196267 and 31CA30 196348] to T.S.

Conflict of Interest: none declared.

References

Aksamentov,I. *et al.* (2021) Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Softw.*, **6**, 3773.

Challen,R. *et al.* (2021) Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study. *BMJ*, **372**, n579.

Chen,C. and Stadler,T. (2021) *Lightweight API for Sequences (LAPIS)*. https://github.com/cevo-public/LAPIS (25 November 2021, date last accessed).

Chen,C. *et al.* (2021) Quantification of the spread of SARS-CoV-2 variant B.1.1.7 in Switzerland. *Epidemics*, **37**, 100480.

Davies,N.G. *et al.* (2021a) Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, **372**, eabg3055.

Davies,N.G. *et al.* (2021b) Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature*, **593**, 270–274.

Elbe,S. and Buckland-Merrett,G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Chall.*, **1**, 33–46.

Gomez,G.B. *et al.* (2021) Uncertain effects of the pandemic on respiratory viruses. *Science*, **372**, 1043–1044.

Hadfield,J. *et al.* (2018) Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, **34**, 4121–4123.

Hodcroft,E.B. (2021) *CoVariants: SARS-CoV-2 Mutations and Variants of Interest*. https://covariants.org/ (25 November 2021, date last accessed).

Hodcroft,E.B. *et al.* (2021) Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature*, **595**, 707–712.

Jahn,K. *et al.* (2021) Detection and surveillance of SARS-CoV-2 genomic variants in wastewater. medRxiv.

Sanderson,T. (2021) *Taxonium*. https://github.com/theosanderson/taxonium (25 November 2021, date last accessed).

Turakhia,Y. *et al.* (2021) Ultrafast Sample placement on Existing tRees (USHER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.*, **53**, 809–816.

World Health Organization (2021) Special edition: proposed working definitions of SARS-CoV-2 Variants of Interest and Variants of Concern. *Technical report*.