



Sequence analysis

CovRadar: continuously tracking and filtering SARS-CoV-2 mutations for genomic surveillance

Alice Wittig ^{1,2,†}, Fábio Miranda ^{1,†}, Martin Hölzer ², Tom Altenburg ¹, Jakub M. Bartoszewicz ^{1,2}, Sebastian Beyvers³, Marius A. Dieckmann³, Ulrich Genske¹, Sven H. Giese ¹, Melania Nowicka ¹, Hugues Richard², Henning Schiebenhoefer ¹, Anna-Juliane Schmachtenberg¹, Paul Sieben¹, Ming Tang ^{1,4}, Julius Tembrockhaus¹, Bernhard Y. Renard ¹ and Stephan Fuchs^{2,*}

¹Digital Engineering Faculty, Hasso Plattner Institute, University of Potsdam, Potsdam 14482, Germany, ²Methods Development, Research Infrastructure and Information Technology (MFI), Bioinformatics and Systems Biology, Robert Koch Institute, Berlin, Germany, ³Department of Biology and Chemistry, Justus-Liebig-University Gießen, Gießen 35390, Germany and ⁴Department of Human Genetics, Hannover Medical School, Hannover 30625, Germany

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Janet Kelso

Received on April 2, 2021; revised on May 13, 2022; editorial decision on June 13, 2022; accepted on June 13, 2022

Abstract

Summary: The ongoing pandemic caused by SARS-CoV-2 emphasizes the importance of genomic surveillance to understand the evolution of the virus, to monitor the viral population, and plan epidemiological responses. Detailed analysis, easy visualization and intuitive filtering of the latest viral sequences are powerful for this purpose. We present CovRadar, a tool for genomic surveillance of the SARS-CoV-2 Spike protein. CovRadar consists of an analytical pipeline and a web application that enable the analysis and visualization of hundreds of thousand sequences. First, CovRadar extracts the regions of interest using local alignment, then builds a multiple sequence alignment, infers variants and consensus and finally presents the results in an interactive app, making accessing and reporting simple, flexible and fast.

Availability and implementation: CovRadar is freely accessible at <https://covradar.net>, its open-source code is available at <https://gitlab.com/dacs-hpi/covradar>.

Contact: FuchsS@rki.de

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

COVID-19 is an infectious disease caused by SARS-CoV-2 and has been declared as a pandemic in March 2020. More than 500 million cases and over six million deaths have been reported worldwide ([World Health Organization, 2022](#)).

Genomic surveillance plays a crucial role to understand evolutionary changes and answer epidemiological questions ([Riley and Blanton, 2018](#)). With decreasing costs and broader access to sequencing technologies, whole-genome sequencing has proven to be an indispensable tool for genomic surveillance of SARS-CoV-2 ([van Dorp et al., 2021](#)).

During the pandemic, various tools like [nextstrain.org](#) ([Hadfield et al., 2018](#)), [covspectrum.org](#) ([Chen et al., 2022](#)) and [outbreak.info](#) were developed or extended to analyze and visualize genomic and epidemiological data to help track evolution and spread

of the virus. We specifically discovered a lack in analyses and visualizations that can summarize large sequence collections over certain time frames and geographic areas while still providing mutation-based information and can be executed locally on custom sequence data.

Here, we present CovRadar, as a tool to focus on individual mutations of the *spike* gene, which encodes an important target for vaccines, and to provide another view on the large amount of SARS-CoV-2 sequence information for virologists and epidemiologists, especially in a public health context.

CovRadar functions primarily at the mutational level, allowing for early observations even before a new SARS-CoV-2 variant is assigned an official lineage name. The analytical workflow can process millions of genome sequences and can also combine different data sources. With the web application, researchers can interactively explore in spatio-temporal resolution the already pre-processed genome data at [covradar.net](#) or custom data on a local system.

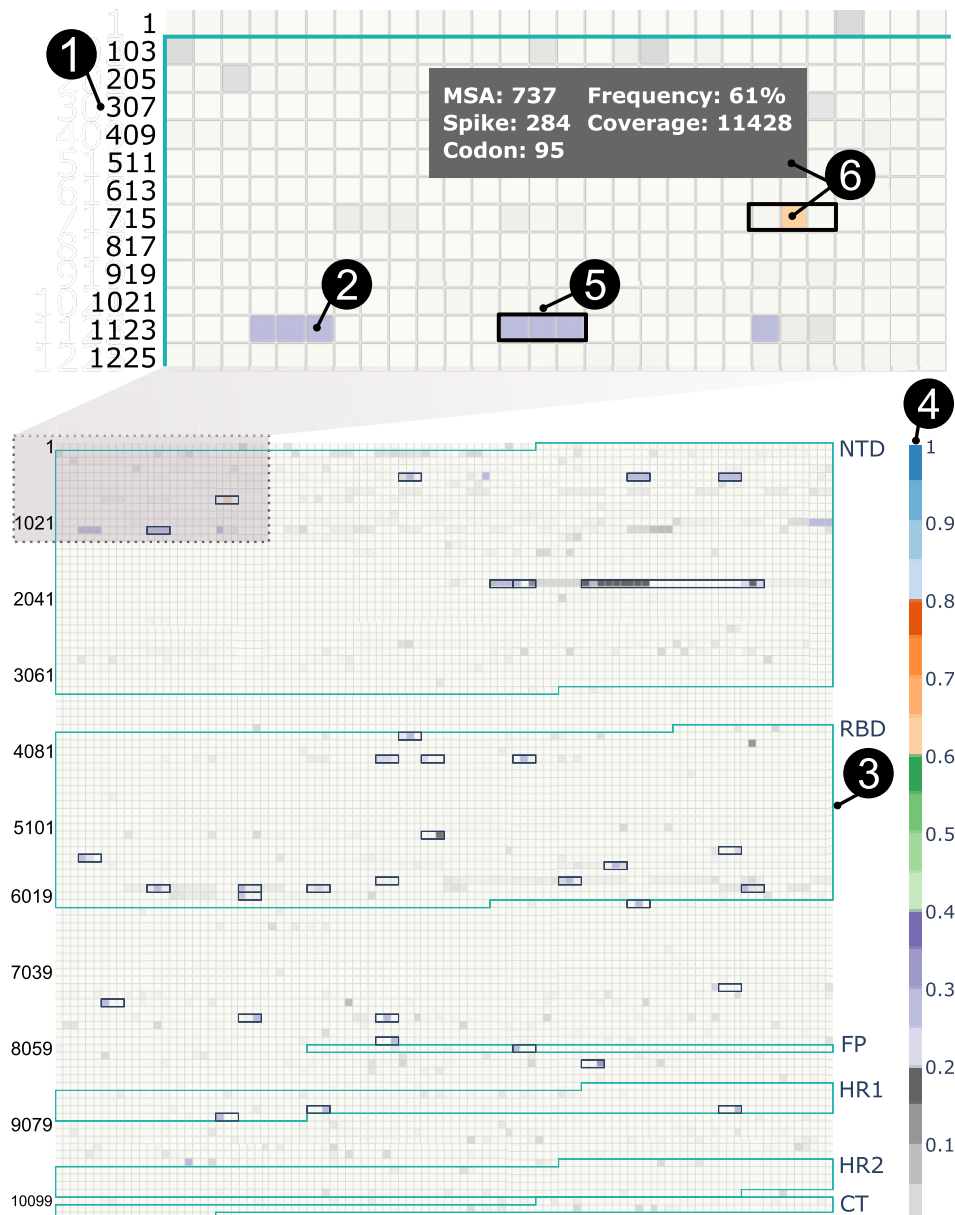


Fig. 1. AAF plot of German sequences from CWs 51 and 52 of 2021 compared to the German consensus sequence of CW50. Coordinates on the left correspond to the MSA positions (1). Each block represents a nucleotide position in the MSA (2). The different Spike protein domains (NTD, RBD, FP, HR1, HR2, CT) are labelled and highlighted with colored frames (3). The AAF is color-coded (4). The plot shows increased frequencies for nucleotide positions common for Omicron [mutations with >75% frequency based on outbreak.info are highlighted by frames (5)]. The highest AAF here (~61%) belongs to codon 95 (6). Shown data are based on 11 428 German sequences obtained from the German Electronic Sequence Data Hub (10.5281/zenodo.6519480)

2 Implementation

CovRadar consists of an analytical pipeline and a web application (Supplementary Figs S1 and S2). The pipeline is written in Snakemake (Köster and Rahmann, 2012) enabling reproducible and expandable analyses. The pipeline takes as input sequence FASTA files and TSV files with their metadata. One or more different data sources can be used simultaneously, e.g. the COVID-19 Data Portal provided by EMBL-EBI (Harrison *et al.*, 2021) and in-house sequence data. First, the pipeline aligns the genomes to the *spike* gene of the index sequence NC_045512.2 to extract the *spike* sequence. Afterwards, it performs a codon-aware multiple sequence alignment (MSA) of the extracted sequences, and retrieves variants and consensus sequences of different countries and calendar weeks (CWs).

After completion of the analysis pipeline, helper scripts can be used to directly access the results and create customized plots. Furthermore, the results can be imported to a database for easier post-analysis or for using the interactive report with CovRadar's app.

The interactive report covers the results for the SARS-CoV-2 Spike protein calculations in a web application written in Flask (Grinberg, 2018) and Dash (Hossain, 2019). The core features are (i) alternative allele frequency plot (Fig. 1) based on consensus sequences for each country and CW with sequence filter options like sampling time, lineage and host; (ii) nucleotide and amino acid mutation distributions per site and CW (Supplementary Fig. S4) and (iii) a global map (Supplementary Fig. S7) showing spatio-temporal resolution of common Spike mutations ($\geq 75\%$ frequency) of the

Variants of Concern (VOC) taken from outbreak.info. The design principles of CovRadar's web application are focused on intuitive overviews that allow users to customize the visualized data based on their needs and to further download the filtered datasets for specialized analyses. At covradar.net we provide pre-processed, daily updated results for the latest SARS-CoV-2 sequences of the COVID-19 Data Portal.

In the supplements, more detailed descriptions of the workflow and web server are provided. In Section 'Use Cases', we demonstrate CovRadar's functionalities for a set of exemplary queries: SARS-CoV-2 in Danish mink farms in 2020, spatio-temporal occurrences of the Spike mutation L452R in Germany, and the rise of characteristic Omicron mutations in Germany (Fig. 1).

3 Conclusion

CovRadar was designed to assist the genomic surveillance of the SARS-CoV-2 Spike protein which is used as a target for vaccines. The flexible filter options facilitate both near real-time and retrospective analyses with a focus on the mutation level and customizable spatial and time analyses. We believe that CovRadar will aid researchers around the world to get better and faster access to SARS-CoV-2 mutation profiles and metadata to support the ongoing fight against the COVID-19 pandemic on the genomic surveillance level.

Acknowledgements

We are very grateful to the GISAID Initiative, the COVID-19 Data Portal hosted by EMBL-EBI, the German Electronic Sequence Data Hub and all data contributors, i.e. the authors from the originating laboratories responsible for obtaining the specimens and the submitting laboratories where genetic sequence data were generated and shared and on which this research is based. We thank Elizabeth Yuu for proofreading the manuscript and David Fischer for short-term implementation support.

Funding

This work was supported by the European Centers for Disease Control [ECDC GRANT/2021/008 ECD.12222] to S.F., the Bundesministerium für

Bildung und Forschung [031L0175C, 01KI1905D] to B.Y.R. as well as Bundesministerium für Wirtschaft und Klimaschutz Daten- und KI-gestütztes Frühwarnsystem zur Stabilisierung der deutschen Wirtschaft [01MK21009E] to B.Y.R. and S.F. and the German Network for Bioinformatics Infrastructure; de.NBI-cloud [031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B].

Conflict of Interest: none declared.

Data availability

The data underlying this article are available in Zenodo (<https://doi.org/10.5281/zenodo.6519480>) and in GISAID (<https://www.gisaid.org>) and can be accessed with GISAID EPI_SET ID 20220509yn.

References

- Chen, C. *et al.* (2022) CoV-spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics*, **38**, 1735–1737.
- Grinberg, M. (2018) *Flask Web Development: Developing Web Applications With Python*. O'Reilly Media, Inc, Sebastopol, CA, USA.
- Hadfield, J. *et al.* (2018) nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, **34**, 4121–4123.
- Harrison, P.W. *et al.* (2021) The COVID-19 Data Portal: accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing. *Nucleic Acids Res.*, **49**, W619–W623.
- Hossain, S. (2019) Visualization of bioinformatics data with dash bio. In: Calloway, C. *et al.* (eds) *Proceedings of the 18th Python in Science Conference*. SciPy, Austin, Texas, pp. 126–133.
- Köster, J. and Rahmann, S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.
- Riley, L.W. and Blanton, R. (2018) Advances in molecular epidemiology of infectious diseases: definitions, approaches, and scope of the field. *Microbiol. Spectr.*, **6**, 1–12.
- van Dorp, L. *et al.* (2021) COVID-19, the first pandemic in the post-genomic era. *Curr. Opin. Virol.*, **50**, 40–48.
- World Health Organization (2022) *Covid-19 Weekly Epidemiological Update (edition 90)*. Technical documents.