Sequence analysis

CovRadar: Continuously tracking and filtering SARS-CoV-2 mutations for molecular surveillance

Alice Wittig[1,2,**], Fábio Miranda[1,**], Ming Tang[1,4], Martin Hölzer[2,3], Bernhard Y. Renard[1,*] and Stephan Fuchs [2,*]

[1]Data Analytics and Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, 14482 Potsdam, Germany [2]Bioinformatics Unit (MF1), Department for Methodology and Research Infrastructure, Robert Koch Institute, 13353 Berlin, Germany [3] European Virus Bioinformatics Center, Friedrich Schiller University Jena, 07743 Jena, Germany [4]Department of Human Genetics, Hannover Medical School, 30625 Hannover, Germany

**These authors contributed equally to this work

ABSTRACT

Summary: The ongoing pandemic caused by SARS-CoV-2 emphasizes the importance of molecular surveillance to understand the evolution of the virus and to monitor and plan the epidemiological responses. Quick analysis, easy visualization and convenient filtering of the latest viral sequences are essential for this purpose. We present CovRadar, a tool for molecular surveillance of the Corona spike protein. The spike protein contains the receptor binding domain (RBD) that is used as a target for most vaccine candidates. CovRadar consists of a workflow pipeline and a web application that enable the analysis and visualization of over 250,000 sequences. First, CovRadar extracts the regions of interest using local alignment, then builds a multiple sequence alignment, infers variants, consensus sequences and phylogenetic trees and finally presents the results in an interactive PDF-like app, making reporting fast, easy and flexible.

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: Bernhard.Renard@hpi.de, FuchsS@rki.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

COVID-19 is an infectious disease caused by the novel Coronavirus SARS-CoV-2, which has been declared a global pandemic by the World Health Organization in March 2020, and currently accumulates more than 70 million cases and around 1.6 million deaths worldwide (World Health Organization, 2020).

Molecular surveillance plays a crucial role in helping us to understand molecular changes and answer epidemiological questions (Riley and Blanton, 2018). Multiple tools have been developed to facilitate real-time analysis (reviewed in (Neher and Bedford, 2018)). They summarize that the kind and content of the visualization depends on the objectives of molecular surveillance. An epidemiological approach aims to find the source and transmission chains of an outbreak. Maps, phylogenetic trees and timelines are suitable tools for this purpose. Nextstrain (Hadfield et al., 2018) is a well established example. However, a pathogen-related approach, which focuses on treatments, vaccinations and resistances, often requires specific, method-dependent visualizations, resulting in many, very specific applications.

Working groups of the public health sector, universities or other institutions often study the same pathogen using the same methods but still require slightly different visualizations, e.g. for different genes, which can result in very time-consuming adaptions.

We see the need to make sequence information more easily accessible on a daily basis for virologists and epidemiologists, particularly in a public health context. As a result, together with the German public health authorities, we developed CovRadar. CovRadar consists of an analysis pipeline and a web application. It is a highly flexible tool that can handle large data and that researchers can adjust to their needs. Both pipeline and app can be installed and run on a local system to include access-restricted sequences. All functionalities are provided at `covradar.net`.

2 METHODS

The backend of CovRadar is a pipeline running on the de.NBI cloud and written in Snakemake (Köster and Rahmann, 2012), enabling reproducible, expandable and resumable analyses. Public sequence data from EMBL-EBI (Madeira et al., 2019) and Charité (`https://civnb.info/sequences/`) are analyzed weekly and the results are stored in a MySQL database. Nevertheless, the pipeline is flexible, allowing the removal or inclusion of private third party sources, e.g., GISAID (Elbe and Buckland-Merrett, 2017). The pipeline employs pblat (Wang and Kong, 2019) to align the genomes to the spike gene of Wuhan-Hu-1 (NC_045512.2) in order to extract the spikes. Afterwards, VIRULIGN (Libin et al., 2019) performs a codon-aware multiple sequence alignment of the extracted sequences, and custom Python scripts retrieve variants and consensus sequences of different countries and calendar weeks. The workflow also infers a phylogenetic tree in Newick format using fast-tree (Price et al., 2010) and creates a static PDF report for fast access and easy archiving. In the supplements, a more detailed description of the workflow is provided.

The interactive PDF report of the SARS-CoV-2 spike protein is a web application written in Flask (`https://github.com/pallets/flask`) and Dash (`https://github.com/plotly/dash`) and connected to a MySQL database. The layout of the web application is adapted to a PDF report with interactive functionalities such as filtering, printing and navigation located in side menu bars. It is possible to save the single tables and plots as excel sheets and PNG, respectively, to use them for other purposes. All elements such as plots and tables can be changed directly.

*To whom correspondence should be addressed.

The frequency plot shows the alternative allele frequencies of the spike either in relation to Wuhan-Hu-1 or to a specific consensus sequence. The sequences can be filtered by host, origin, country and time period. The consensus table presents mutations that have occurred in the spike for each calendar week and country. CovRadar additionally enables basic quality control. Summary statistics of the pipeline steps, as well as base and sequence coverage, are provided to the user and the consensus sequence table indicates whether the sequences have been properly aligned.

3 RESULTS

CovRadar was designed to assist the molecular surveillance of the Corona spike protein used as target for most vaccine candidates. The flexible filter options facilitate both real-time and retrospective analyses. To demonstrate the usefulness of CovRadar, we applied our tool for a set of exemplary queries on the GISAID database consisting of 244,298 sequences obtained on December 9th.

The complete data set containing the sequences of all countries and dates can be used to investigate global dynamics. Two major mutations in the spike regions are D614G, first documented in January in China and Germany (Korber et al., 2020), and A222V, which occurred first in June in Spain and recently became dominant (Hodcroft et al., 2020). The consensus sequences of the individual calendar weeks show the appearance of the non-synonymous A to G mutation at position 23403, the 2nd base of codon 614, and the non-synonymous C to T mutation at position 665, the 3rd base of codon 222 (Fig. S1). Before the mutations became consensus, the gradual increase of the alternative alleles can be observed in the frequencies. It is also possible to work with a subset. This allows examining specific countries, hosts, or time intervals. Through the country filter, country-specific mutations and trends can be seen in the consensus sequence table and allele frequency plot. For instance, for Australia they show two striking features compared to other countries: A non-synonymous mutation at position 1430, which corresponds to codon 477, and a silent mutation at position 1839 corresponding to codon 613 (Fig. S2).

CovRadar can also assist in situations where zoonotic transmissions with host-specific mutations happen. One of the most significant zoonotic events to date is the host jump between mink and humans. Several mutations are currently under discussion whether they are related to mink and what significance they have for the soon-to-be-released vaccines (van Dorp et al., 2020). Using CovRadar, we looked at the positions in the available Danish mink data with high allele frequencies and compared them with the human data before and after the mink outbreak. We found an increase of alternative alleles especially at codon positions 453 and 68–70 (Fig. 1), which might be related to Y435F and del69-70.

FUNDING

This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B). BYR gratefully acknowledges support by a BMBF Computational Life Science project (grant number 031L0175B).

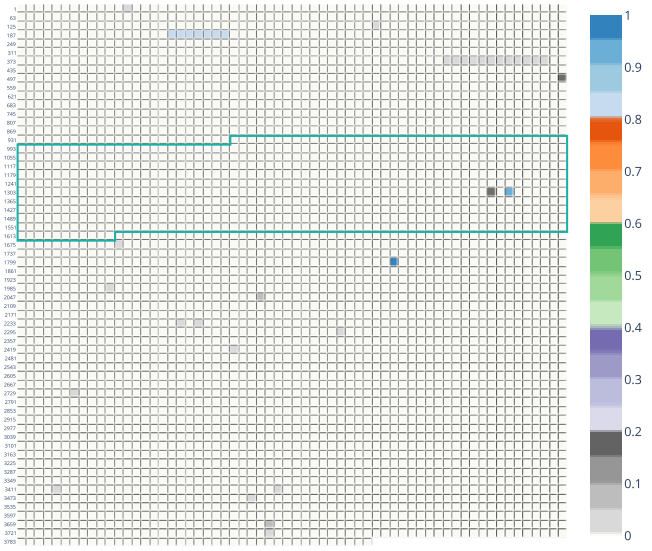

Fig. 1. Allele frequency plot of the Danish mink sequences with framed receptor binding domain (RBD). Each block represents a nucleotide in the MSA of spike gene sequences. Coordinates on the left are related to the MSA position. The plot shows frequencies > 0.8 for nucleotide positions corresponding to codons 68–70, 453 and 614.

Conflict of Interest: none declared.

REFERENCES

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

Hadfield, J. et al. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics, **34**(23), 4121–4123.

Hodcroft, E. B. et al. (2020). Emergence and spread of a sars-cov-2 variant through europe in the summer of 2020. medRxiv.

Korber, B. et al. (2020). Tracking changes in sars-cov-2 spike: evidence that d614g increases infectivity of the covid-19 virus. Cell, **182**(4), 812–827.

Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. Bioinformatics, **28**(19), 2520–2522.

Libin, P. J. et al. (2019). Virulign: fast codon-correct alignment and annotation of viral genomes. Bioinformatics, **35**(10), 1763–1765.

Madeira, F. et al. (2019). The embl-ebi search and sequence analysis tools apis in 2019. Nucleic acids research, **47**(W1), W636–W641.

Neher, R. A. and Bedford, T. (2018). Real-time analysis and visualization of pathogen sequence data. Journal of clinical microbiology, **56**(11).

Price, M. N. et al. (2010). Fasttree 2–approximately maximum-likelihood trees for large alignments. PloS one, **5**(3), e9490.

Riley, L. W. and Blanton, R. (2018). Advances in Molecular Epidemiology of Infectious Diseases: definitions, approaches, and scope of the field. Microbiology spectrum, **6**(6).

van Dorp, L. et al. (2020). Recurrent mutations in sars-cov-2 genomes isolated from mink point to rapid host-adaptation. bioRxiv.

Wang, M. and Kong, L. (2019). pblat: a multithread blat algorithm speeding up aligning sequences to genomes. BMC bioinformatics, **20**(1), 28.

World Health Organization (2020). Weekly epidemiological update - 14 December 2020. `https://www.who.int/publications/m/item/weekly-epidemiological-update---14-december-2020`. [Online; accessed 14-December-2020].