BIOINFORMATICS

Vol. 00 no. 00 2020
Pages 1–2

Sequence analysis

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

Alice Wittig[1,2,**], Fábio Miranda[1,**], Martin Hölzer[2], Tom Altenburg[1], Jakub M. Bartoszewicz[1,2], Marius A. Dieckmann[3], Ulrich Genske[1], Sven H. Giese[1], Melania Nowicka[1], Henning Schiebenhoefer[1], Anna-Juliane Schmachtenberg[1], Ming Tang[1,4], Bernhard Y. Renard[1,*], Stephan Fuchs [2,*]

[1]Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, 14482 Potsdam, Germany [2]Bioinformatics Unit (MF1), Robert Koch Institute, 13353 Berlin, Germany [3]Bioinformatics & Systems Biology, Justus-Liebig-University Gießen, Heinrich-Buff-Ring 58, 35390 Gießen, Hesse, Germany [4]Department of Human Genetics, Hannover Medical School, 30625 Hannover, Germany **These authors contributed equally to this work

Associate Editor: XXXXXXXXXXXXX

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 1 million 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 122 million cases and around 2.7 million deaths worldwide (World Health Organization, 2021).

Molecular surveillance plays a crucial role to understand evolutionary changes and answer epidemiological questions (Riley and Blanton, 2018). Multiple developed tools facilitate real-time analysis. The kind and content of the visualization depends on the objectives of molecular surveillance (Neher and Bedford, 2018). 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.

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 written in Snakemake (Köster and Rahmann, 2012), enabling reproducible, expandable and resumable analyses. Public sequence data from EMBL-EBI (Madeira et al., 2019) and the Charité (`https://civnb.info/sequences/`) are analyzed daily. Furthermore, the Robert Koch Institute uses CovRadar in-house to evaluate data related to Germany on a weekly basis. In addition, the removal or inclusion of licensed third party sources, e.g., GISAID (Elbe and Buckland-Merrett, 2017) is supported. First, the pipeline employs pblat (Wang and Kong, 2019) to align the genomes to the spike gene of Wuhan-Hu-1 (NC_045512.2) to extract the spike sequence. 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

*To whom correspondence should be addressed.

© Oxford University Press 2020.

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 and Dash and connected to a MySQL database. The layout is adapted to a PDF report, while all elements such as plots and tables can be directly modified and downloaded. CovRadar provides an overview of the input data and allows the data to be analyzed for amino acid and nucleotide changes in the spike region. In addition, the collaboration with the Robert Koch Institute allows a closer look at data from Germany. A map of German primary sequencing laboratories sequencing positive PCR tests displays for selected "Mutations of Concern" their geographical distribution and temporal development (Fig. S1).

3 RESULTS

CovRadar was designed to assist the molecular surveillance of the Corona spike protein which is used as a target for most vaccines. 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.
The complete data set containing the sequences of all countries and dates can be used to investigate **global dynamics**, e.g. the surveillance of "Mutations of Concern" (MOC). Mutations such as N501Y in the RBD of the spike were shown as important factors for enhanced receptor binding (Zahradník et al., 2021) and attenuated antibody response (Wang et al., 2021). The variant B.1.1.7, which was first discovered in the UK, carries in addition to N501Y further mutations such as A570D, D614G, P681H, T716I, S982A, and D1118H and the characteristic deletions del69-70 and del144. The emergence of the MOC in global scale can be monitored with CovRadar (Fig. 1).
It is also possible to work with **specific countries, hosts, or time intervals**. As B.1.1.7 is becoming the main variant in Germany (Oh et al., 2021), its mutations are also increasing in frequency and have been consensus since mid-February 2021 (Fig. S2). In addition, a map based on sequences from diagnostic primary laboratories in Germany allows investigation of local and temporal distributions of selected MOC (Fig. S1). We have observed that in recent weeks L18F, which was previously predominant, has been replaced by N501Y and del144 which are also part of the lineage-defining mutations of B.1.1.7.
CovRadar can assist in situations where **zoonotic transmissions** with host-specific mutations happen. One significant zoonotic event in 2020 was the transmission from mink to human. Using CovRadar, we found an increase of alternative alleles especially at codon positions 453 and 68–70 (Fig. S3), that correspond to Y435F and del69-70. For del69-70 it has been shown that it often co-occurs with the receptor binding motif amino acid variants (Kemp et al., 2020).

FUNDING

This work was supported by the BMBF Computational Life Science (031L0175B) to BYR and the German Network for Bioinformatics Infrastructure; de.NBI-cloud: 031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B).

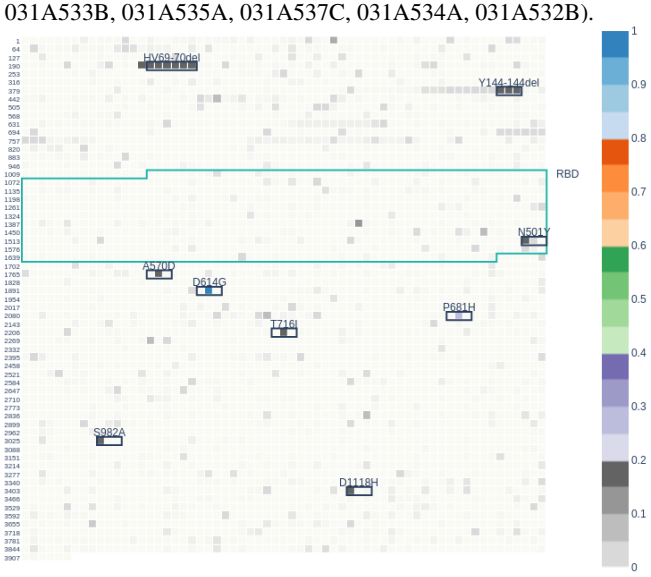

Fig. 1. Allele frequency plot of global sequences from calendar week 8/2021 with framed receptor binding domain (RBD). Each block represents a nucleotide in the MSA of spike gene sequences. Coordinates on the left correspond to the MSA position. The plot shows frequencies >0.16 for nucleotide positions within codons 69, 70, 144, 501, 570, 614, 681, 716, 982 and 1118. Data were obtained from the EMBL-EBI database and the Charité.

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.
Kemp, S. et al. (2020). Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion ΔH69/V70. bioRxiv.
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).
Oh, D.-Y. et al. (2021). SARS-CoV-2-Varianten: Evolution im Zeitraffer. Dtsch Arztebl International, **118**(9), A–460–.
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).
Wang, M. and Kong, L. (2019). pblat: a multithread blat algorithm speeding up aligning sequences to genomes. BMC bioinformatics, **20**(1), 28.
Wang, Z. et al. (2021). mrna vaccine-elicited antibodies to sars-cov-2 and circulating variants. bioRxiv.
World Health Organization (2021). Weekly epidemiological update - 23 March 2021. `https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---23-march-2021`. [Online; accessed 29-March-2021].
Zahradník, J. et al. (2021). Sars-cov-2 rbd in vitro evolution follows contagious mutation spread, yet generates an able infection inhibitor. bioRxiv.