The Journal of Open Source Software

Nextclade: clade assignment, mutation calling and quality control for viral genomes

Ivan Aksamentov[1, 2], **Cornelius Roemer**[1, 2], **Emma B. Hodcroft**[2, 3], **and Richard A. Neher***[1, 2]

1 Biozentrum, University of Basel, Switzerland **2** Swiss Institute of Bioinformatics, Basel, Switzerland **3** Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

DOI: 10.21105/joss.03773

Software

- Review
- Repository
- Archive

Editor: Kelly Rowland

Reviewers:

- @kevinlibuit
- @DavidNickle

Submitted: 03 September 2021
Published: 29 November 2021

Summary

The variants of concern (VoCs) of SARS-CoV-2 have highlighted the need for a global molecular surveillance of pathogens via whole genome sequencing. Such sequencing, for SARS-CoV-2 and other pathogens, is performed by an ever increasing number of labs across the globe, resulting in an increased need for an easy, fast, and decentralized analysis of initial data. `Nextclade` aligns viral genomes to a reference sequence, calculates several quality control (QC) metrics, assigns sequences to a clade or variant, and identifies changes in the viral proteins relative to the reference sequence. `Nextclade` is available as a command-line tool and as a web application with completely client based processing, meaning that sequence data doesn't leave the user's browser.

Statement of need

After assembly of a consensus genome from raw read data, it is usually desirable to (i) assess the quality of the sequence, (ii) assign it to a known clade or type, and (iii) compare it to a reference sequence to detect evolutionary changes. `Nextclade` addresses this need through a command-line interface for bulk analysis of many sequences and a web-tool with the same functionality coupled to an interactive visualization. `Nextclade` is built on `Nextalign`, a codon-aware pairwise sequence aligner for similar viral genomes, which allows unambiguous calling of amino-acid changes associated with changes in the nucleotide sequence. The sequence is then placed onto a phylogenetic tree generated by the `augur` pipeline (Huddleston et al., 2021) of the `Nextstrain` tool-chain (Hadfield et al., 2018).

During the SARS-CoV-2 pandemic, `Nextclade` has already allowed countless users to quickly analyze their data, assign sequences to clades and variants of concern, and identify mutations of interest.

Implementation

`Nextclade` consists of three tools:

- Nextclade Web
- Nextclade CLI

*corresponding author

- Nextalign CLI

All tools share the common C++ library of algorithms. The CLI tools are also implemented in C++. Nextclade Web is a React web application written in Typescript, and it uses the C++ algorithms compiled to WebAssembly. All tools are meant to align multiple sequences to one common reference sequence.

Nextalign

`Nextalign` implements a banded pairwise Smith-Waterman alignment with an affine gap cost (Smith & Waterman, 1981). The bandwidth and relative shift of the two sequences are determined by seed matching. In contrast to most other existing tools (e.g.~`minimap2` (Li, 2018) or `mafft` (Katoh & Standley, 2013)), `Nextalign` can use a genome annotation specifying coding regions to make the gap-opening penalty dependent on the reading frame. This allows `Nextalign` to choose the most biologically interpretable gap-placement between otherwise equivalent alignments. In the following example, the gap could be moved forward or backward by one base with the same number of matches, but a frame-dependent gap-opening penalty locks the gap in-frame:

```
...GTT.TAT.TAC...
...GTT.---.TAC...
```

Similarly, `Nextalign` preferentially places gaps outside of genes in case of ambiguities.

In addition to nucleotide alignments, `Nextalign` will extract the aligned coding sequences, translate them, and perform pairwise amino-acid alignments. These amino-acid alignments are produced alongside the nucleotide alignment and are used by `Nextclade` to determine amino-acid changes. All alignment parameters can be configured via CLI flags.

Nextclade

`Nextclade` uses the results of `Nextalign` to determine all mutations of each query sequence relative to the reference sequence. With this set of mutations, it performs an exhaustive search for the closest match on a phylogenetic tree representing the diversity of the population. The clade of the closest match is assigned to the query sequence.

In addition, `Nextclade` determines the mutations separating the closest match from the query sequence. This set of *private mutations* is used as a QC metric: having many private mutations is often a sign of sequencing errors or miscalled bases. If such private mutations cluster in short stretches on the genome, this is an additional sign of concern. The private mutation count, a measure of SNP clusters, as well as rules quantifying sequence completeness, ambiguous bases, stop-codons, and frame-shifts are used to quantify sequence quality, individually for each metric and via an aggregate score.

Details of the algorithm and the different QC metrics are described in the documentation at docs.nextstrain.org/projects/nextclade.

Web interface

While CLI tools are most appropriate for bulk processing, analyzing up to a few hundred sequences is feasible and possibly more convenient via a graphical interface coupled to a visualization. Nextclade enables this via a completely client side web-application onto which users can drop a fasta file with sequences. The results are displayed in an interactive viewer

that highlights QC metrics and nucleotide mutations (see Figure 1), and allows users to explore the effects of complex mutations on viral proteins (see Figure 2). QC results, variant calls, and the full alignment can be downloaded from the web application for further analysis. Users can also view the placement of the query sequences in the reference tree through an interactive interface.

Figure 1: Overview of the results page with clade assignments, QC metrics, and the nucleotide mutation view. The results can be explored interactively and exported in standard tabular file formats.

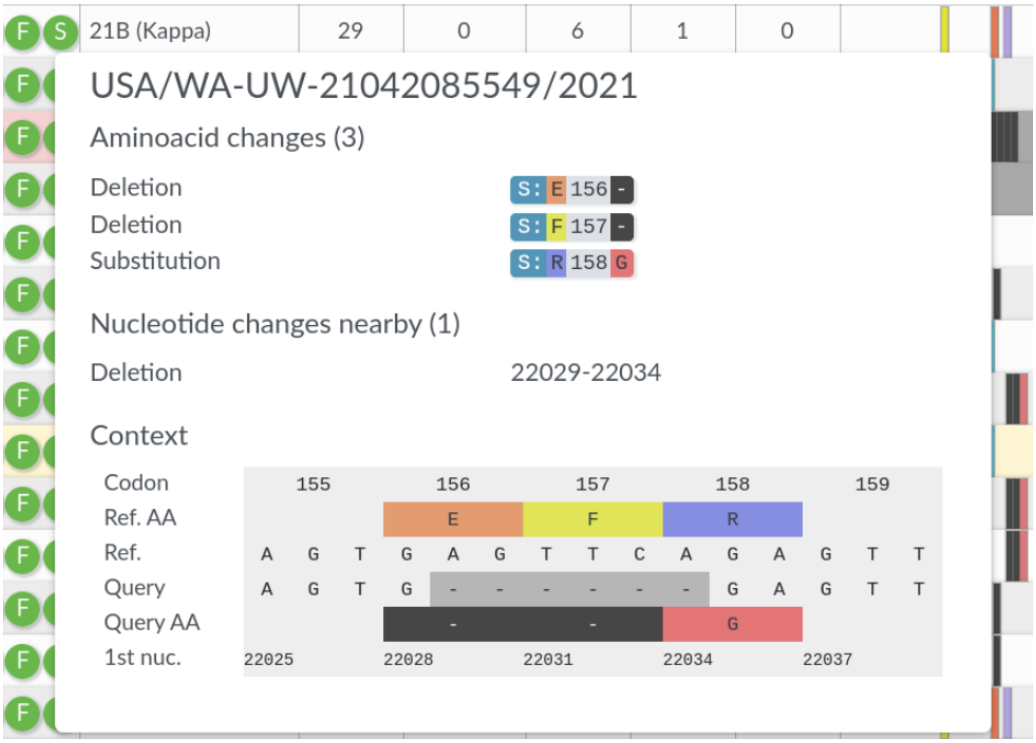

Figure 2: Mutations in each gene can be explored interactively using tool-tips that show how the changes in the nucleotide sequence correspond to changes in the viral proteins. This is particularly useful when complex mutations, such as the six base deletion in the above example, affect multiple codons.

Nextclade datasets

To run the `Nextclade` CLI, the user needs to provide a reference sequence, an annotation, a labeled tree, a QC configuration and optionally a set of primers. We currently maintain such data sets for SARS-CoV-2 and the four seasonal influenza viruses. These are automatically available in the web tool. The `nextclade` CLI tool includes `dataset list` and `dataset get` commands to explore and download available datasets. The SARS-CoV-2 tree is labeled with the Nextstrain clade annotations for SARS-CoV-2 which follow a year-letter pattern (e.g. `20A`) and coupled with the corresponding WHO variant label where available (e.g. `21A (Delta)`) (Konings et al., 2021). Influenza trees are labeled with the clades currently used by Nextstrain to describe circulating influenza virus diversity.

Discussion

`Nextclade` was developed in response to the increasing need for laboratories around the world to quickly assess the quality of their newly generated SARS-CoV-2 sequences, categorize them into different variants and clades, and investigate their mutational profiles. While `Nextclade` has some similarities to UShER (Turakhia et al., 2021), these two tools address different use cases. UShER places sequences on a comprehensive tree with hundreds of thousands of leaves and further refines the phylogenetic relationship of the user supplied sequences to analyze the fine-scale relationship between the user supplied sequences and other publicly available data. Supplied sequences need to be uploaded to UShER's servers where processing takes place. `Nextclade` provides a completely client-side analysis of sequences with a focus on QC, clade assignment, and investigation of variation. Nextalign was written for a very specific use case: fast pairwise alignment of similar sequences ($< 10\%$ divergence) with limited insertions and deletions. For more diverse data sets, tools like `mafft` or `minimap2` are likely more robust.

As sequencing of pathogens becomes more wide-spread, bioinformatic analyses of such data increasingly becomes a bottleneck. We aim to increase the number of pathogens for which `Nextclade` datasets are provided and hope that it will help users with variable experience levels easily gain as much insight into their own data as possible.

Acknowledgments

We gratefully acknowledge the generous public sharing of sequence data by many labs around the world that make tools like `Nextclade` possible and useful. We are also grateful for feedback from the `Nextstrain` team and the wider community for critical feedback and suggestions on how to improve the tools. Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing center at the University of Basel.

References

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, *34*(23), 4121–4123. https://doi.org/10.1093/bioinformatics/bty407

Huddleston, J., Hadfield, J., Sibley, T. R., Lee, J., Fay, K., Ilcisin, M., Harkins, E., Bedford, T., Neher, R. A., & Hodcroft, E. B. (2021). Augur: A bioinformatics toolkit for phylogenetic analyses of human pathogens. *Journal of Open Source Software*, *6*(57), 2906. https://doi.org/10.21105/joss.02906

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Konings, F., Perkins, M. D., Kuhn, J. H., Pallen, M. J., Alm, E. J., Archer, B. N., Barakat, A., Bedford, T., Bhiman, J. N., Caly, L., Carter, L. L., Cullinane, A., Oliveira, T. de, Druce, J., El Masry, I., Evans, R., Gao, G. F., Gorbalenya, A. E., Hamblion, E., … Van Kerkhove, M. D. (2021). SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse. *Nature Microbiology*, 1–3. https://doi.org/10.1038/s41564-021-00932-w

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Turakhia, Y., Thornlow, B., Hinrichs, A. S., De Maio, N., Gozashti, L., Lanfear, R., Haussler, D., & Corbett-Detig, R. (2021). Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics*, *53*(6), 809–816. https://doi.org/10.1038/s41588-021-00862-7