Introduction

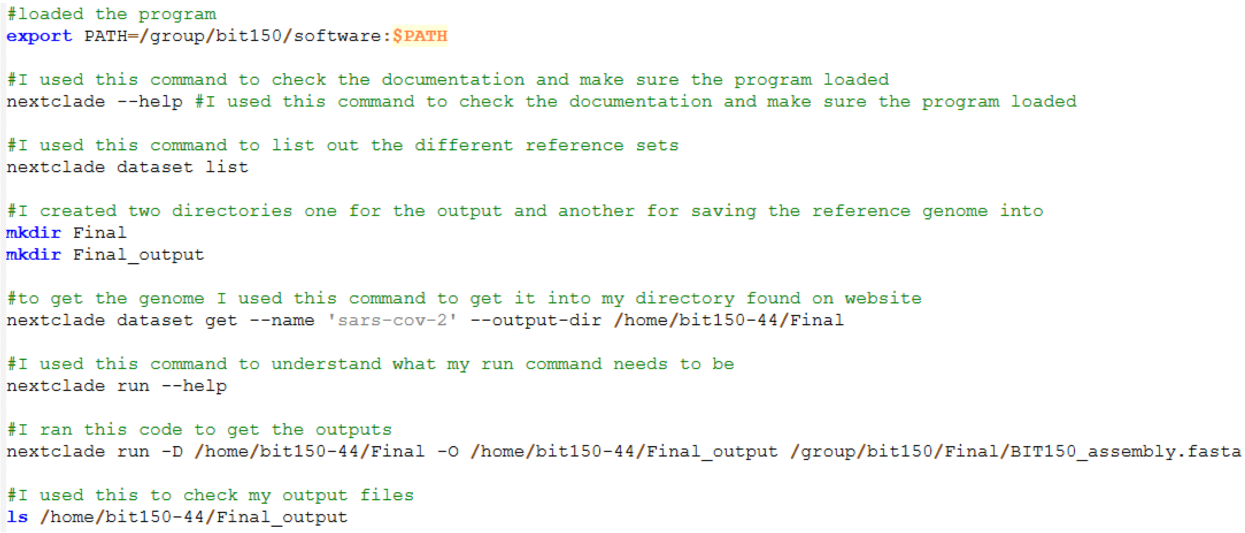
The Nextclade program is part of the Nextstrain project, which aims to utilize pathogen data for to help for scientific and public health. The project aims to help in the understanding pathogens their spread and evolution in order to better outbreak responses. The project is created and maintained by bioinformaticians part of the University of Basel and the Swiss Institute of Bioinformatics along with other community contributors. The bioinformaticians and software developers maintain the tool and update it with new data and analyses tools. The Nextclade program's apps and algorithms are open source under the terms of an open-source license this means it follows the rules highlighted by the open-source initiative such as free redistribution, all source code, and modifications of the derived works. The motivation for doing so according to the Nextclade creators lies in lowering the barriers for different health organizations and research labs to share results openly and promote collaboration.

Nextclade itself is a tool that can perform genetic alignments, clade assignment, mutation calling, phylogenetic placement, and quality checks for SARS-CoV-2, Influenza (Flu), Mpox (Monkeypox), Respiratory Syncytial Virus (RSV) and other pathogens. This tool can be a powerful way for scientists, epidemiologists, and health officials to monitor and learn about the different genomes and see where strains of pathogens lie in relation to others. It can help in understanding the linage of a pathogen such as SARS-CoV-2, see what different variants may be present and understand how mutations and changes to the genome affect behaviors and transmissions. It is important to understand the clade a particular pathogen belongs to as it visually shows the evolutionary history of the strain. For example, with SARS-CoV-2 by monitoring mutations and grouping them together it can provide a timeline for an outbreak such as the alpha, beta, delta and omicron variants. Additionally, it may help in understanding behavior patterns such as increased transmutability, severity or immune responses i.e if a certain clade carries higher severity than others. It can also help in managing public outbreaks and help inform different public health interventions that may be needed depending on known information, in the case of SARS-CoV-2 it may help in understanding which clade is dominant in what areas, vaccinations, and quarantine durations.

Nextclade consists of a web tool along with a command line tool, they both use the same algorithms and the same inputs and outputs. The advantage of the CLI is that it can help processes larger amounts of data rather than the slower more tedious way of using the web tool. Since its also on the command line large amounts of data can be processed as scripts can be created to do so. Additionally, Nextclade provides more tools such as custom reference sets and have more parameters set for the command line version. The web-tool is more user friendly and easier for smaller datasets, however, command line is advantageous when it comes to large scale data processing and automation.

Part 1 – Running Nextclade

The commands I used to run Nextclade are as follows



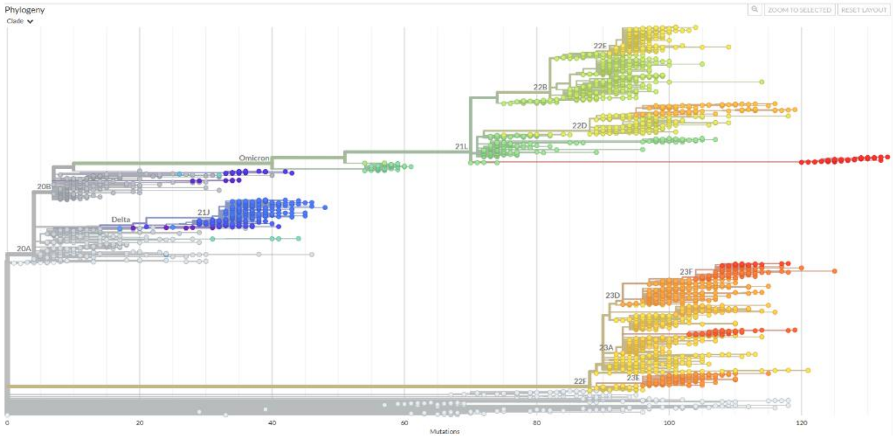
```
#loaded the program
export PATH=/group/bit150/software:$PATH

#I used this command to check the documentation and make sure the program loaded
nextclade --help #I used this command to check the documentation and make sure the program loaded

#I used this command to list out the different reference sets
nextclade dataset list

#I created two directories one for the output and another for saving the reference genome into
mkdir Final
mkdir Final_output

#to get the genome I used this command to get it into my directory found on website
nextclade dataset get --name 'sars-cov-2' --output-dir /home/bit150-44/Final

#I used this command to understand what my run command needs to be
nextclade run --help

#I ran this code to get the outputs
nextclade run -D /home/bit150-44/Final -O /home/bit150-44/Final_output /group/bit150/Final/BIT150_assembly.fasta

#I used this to check my output files
ls /home/bit150-44/Final_output
```

Part 2 – Outputs

There are many outputs from the Nextclade program. One of the outputs is the nextclade.aligned.fasta, this contains the aligned sequences produced from the sequence alignment with the reference SARS-CoV-2 genome it shows how well the input sequences align with the reference and has information about the gaps and mismatches. Another output is the nextclade.auspice.json this is a json file that has the output phylogenetic tree, by uploading this file to auspice there is a visual representation of the different evolutionary relationships along with the clade assignments. There is also an output called the nextclade.tsv this shows the results of the mutations, clade assignment, and the different quality control metrics. This file analyzes the clade and the associated mutations and gives the quality of the analysis.

Part 3 – Results

After I obtained the output using the nextclade.tsv I looked for which genes had substitutions, deletions and other mutations. I found that one of the genes that has a mutation is the ORF1a gene. This gene is the replicase polyprotein crucial for replication and transcription. ORF1a creates a polyprotein which is cleaved into viral proteases and then non-structural proteins. Mutations in this gene would probably affect viral replication and issues in protease function. Using the qc or quality control the overall score was around 8.5 which is considered a good quality score. From the tsv it can also be seen that the clade is 21J which is designated as the delta variant. Next I input the auspice.json file into the auspice website and obtained these results



From the graph it can be seen that the delta clade labelled as 21J comes from earlier clades like 20A, 20B, 20C however, it has diverged significantly from ancestral clades. Comparatively the strains have a lot more mutations than the ancestral ones as well as seen by the number of mutations. Other variants such as the omicron have a lot more mutations compared to the delta. This suggests that the delta variant evolved independently and is less mutated than variants like omicron

Part 4 – Nextclade Options

There are a few options you can adjust for example there is an option for --genes which only analyzes the genes that are specified, another option is --output dir which specifies a custom output directory. There are also options for --min-length and --max-length which sets the limit for the sequence lengths. I decided to change the genes analyzed to just target the Spike protein. When I compared my outputs the analysis report only showed the spike protein (S) and it excludes other genes. The quality control metrics were also different as the qc.frameShifts and qc.stopCodons show a 0 now since the spike protein has no frameshifts. The output fasta file is also smaller since it includes less data.

Part 5 – Script

After processing all of the files in the unknown sequences directory using the script attached, I obtained the following results

- BIT150_2021-02-17 is Epsilon 21C
- BIT150_2023-06-20 is Alpha 20I
- BIT150_2024-05-18 is Omicron 22B

So, all of these samples are not from the same clade of Sars-CoV-2 they each are from a different one.

Discussion

Nextclade overall is a powerful tool that can be used to understand genomes of some pathogens. It is easy to use due to its web version and command line, it creates a detailed output with features about the mutations and clades, and it contains different quality control metrics which helps in the understanding of the sequence in terms of errors and mismatched alignments. There are also a variety of options to use to edit your such as searching for specific genes, modifying the output names, changing the length, quality and much more. However, I felt as if there were some limitations of the program as well. The alignments performed depend significantly on the reference genome and since it simply matches each amino acid it doesn’t use any advanced alignment techniques such as one that could be self-correcting. There is also not much analysis present beyond clade and mutation data and other bioinformatic tools may be needed to predict the impact of those mutations. Additionally so far nextclade only supports certain pathogens which are SARS-CoV-2, Influenza A/B HA (H3N2,H1N1pdm,Vic,Yam), RSV, and

Mpox virus data, in order to use the tool for a different pathogen the user must provide the reference sequences, trees and annotations which is a drawback.