Using Nextstrain and Nextclade to Analyse Dengue Sequences

Ammar Aziz – VIDRL

Doherty Institute

THE UNIVERSITY OF MELBOURNE

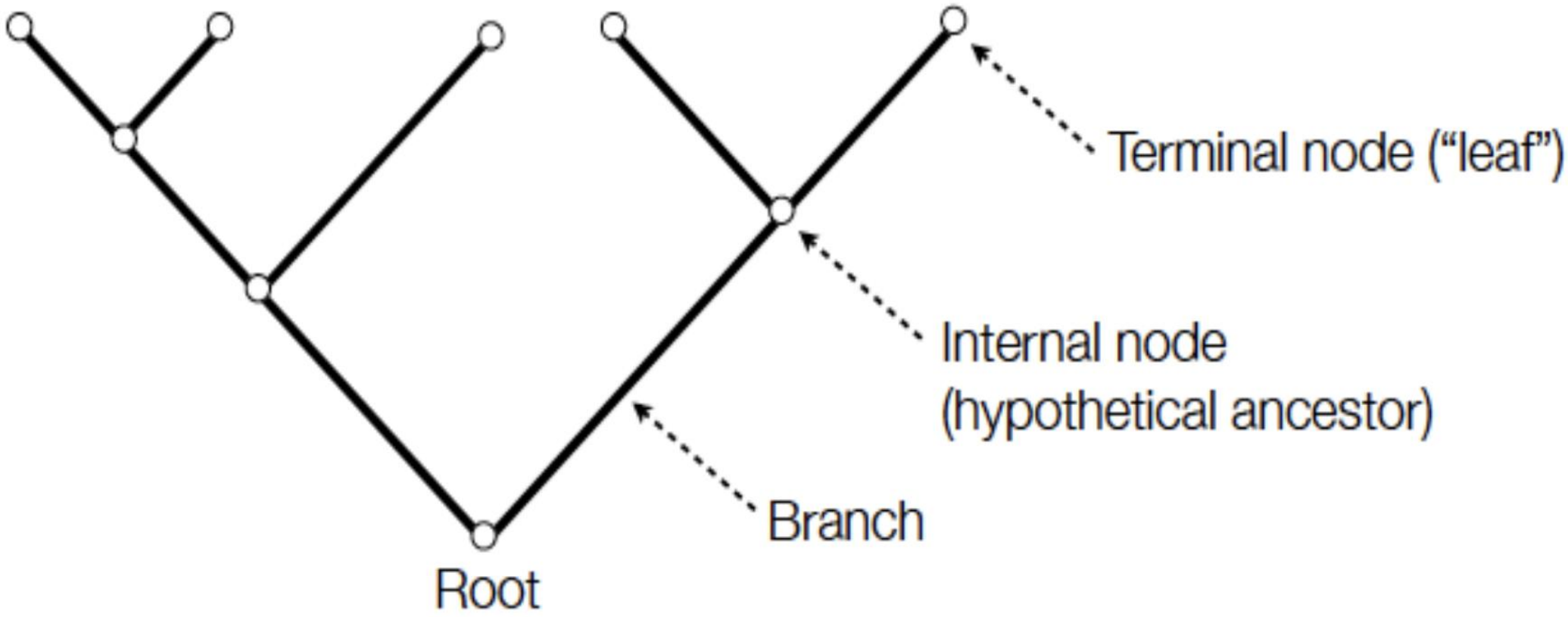
The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital

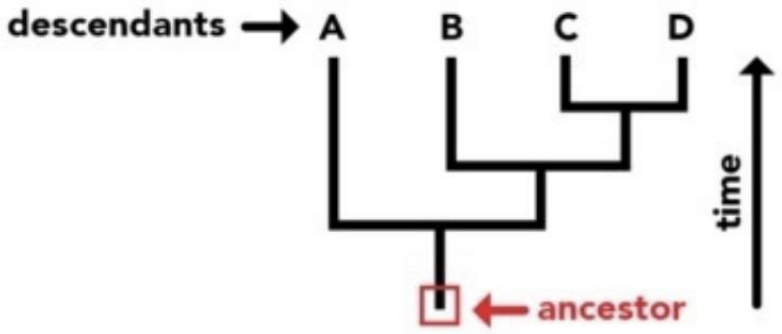
Objectives

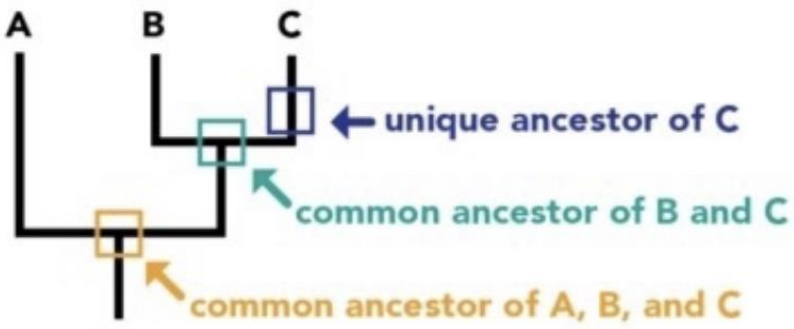
- Refresher on Phylogenies
- Dengue global surveillance using Nextstrain
- Clade/Lineage analysis using Nextclade

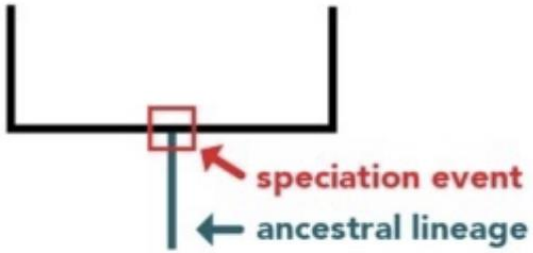
Trees - Terminology

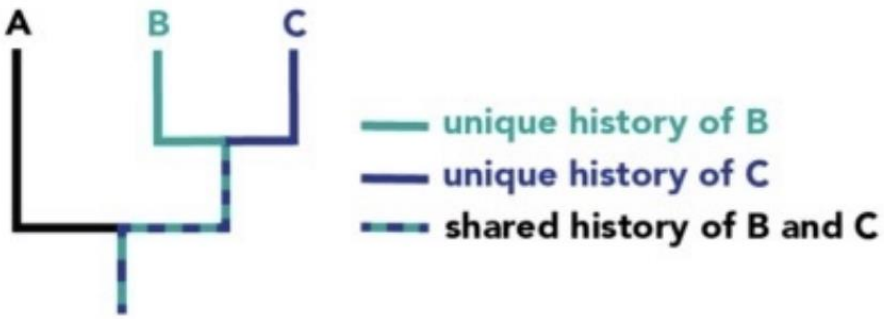


Reading Trees

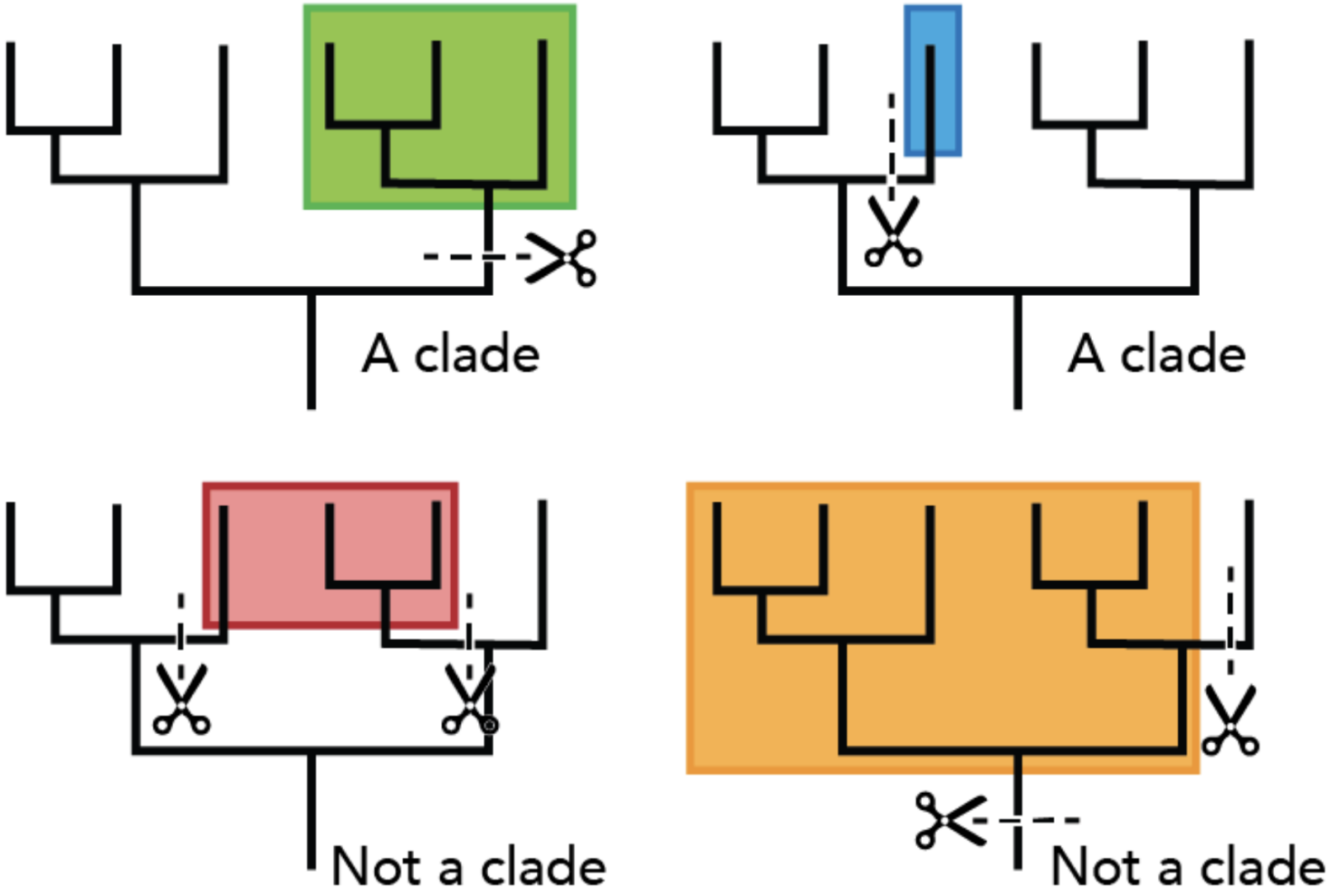




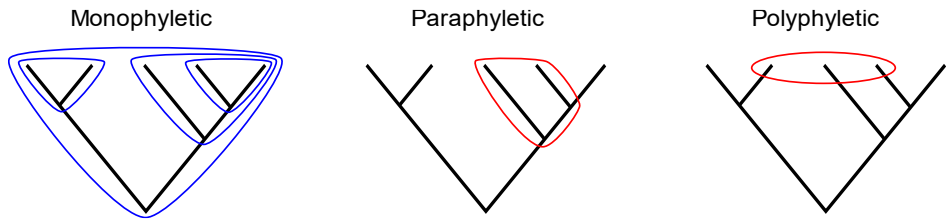




What is a Clade?



- <u>Monophyletic group</u> - All and only the descendants of a common ancestor (including that ancestor).
- <u>Paraphyletic group</u> - Does not include all of the descendants of the common ancestor.
- <u>Polyphyletic group</u> - Does not include the common ancestor.



What is Nextstrain?

Nextstrain is a project to harness the scientific and public health potential of pathogen genome data. Our goal is to aid epidemiological understanding of pathogen spread and evolution and improve outbreak response.

Pathogen surveillance

Our website, nextstrain.org, provides real-time snapshots of evolving pathogen populations such as SARS-CoV-2, influenza, and Ebola. We use interactive visualizations to enable exploration of curated datasets and analyses which are continually updated when new genomes are available. This offers a powerful pathogen surveillance tool to virologists, epidemiologists, public health officials, and community scientists. In many cases old snapshots of these analyses are able to be easily accessed, see viewing previous analyses for more.

Open-source software

The software we write to power all parts of Nextstrain—bioinformatics, visualizations, analysis pipelines, data management, and more—is entirely open-source and available to the public. We aim to empower the wider genomic epidemiology and public health communities to tweak our analyses, create new ones, and communicate we do.

Nextstrain – Reading trees

- https://nextstrain.org/narratives/trees-background/

Nexstrain Demostration

- Dengue E tree: https://nextstrain.org/dengue/all/E
- Filtering:
  - By country
  - By serotype
- Change colors:
  - Date
  - Country
  - Region
- Interactively explore tree
- Tanglegram (WGS – E tree)

Nextclade: analysis of viral genetic sequences

Nextclade is an open-source project for viral genome alignment, mutation calling, clade assignment, quality checks and phylogenetic placement.

Nextclade consists of a set of related tools:

- Nextclade Web - a web application available online at clades.nextstrain.org
- Nextclade CLI - a command-line tool

Both tools are powered by the same algorithms, they consume the same inputs and produce the same outputs, but they differ in the user interface, the features included, and the degree of customization. It is recommended to start with Nextclade Web and later proceed to CLI tools if you have more advanced use-cases (for example, repeated batch processing, bioinformatics pipelines).

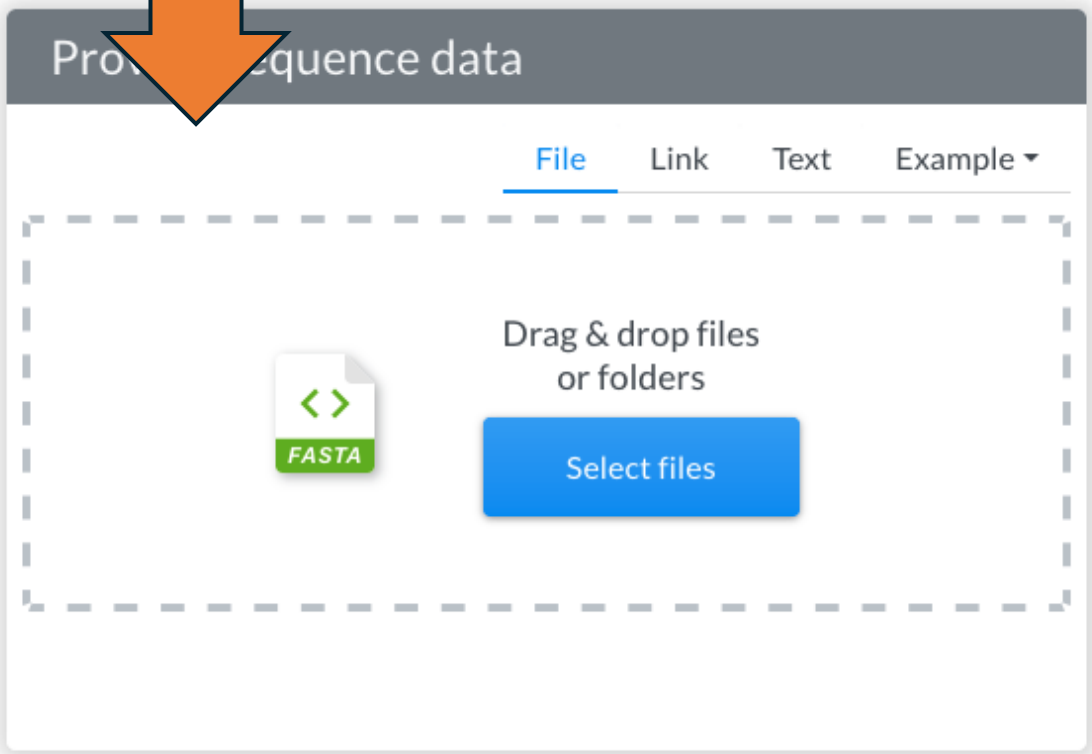
Download sequencing data

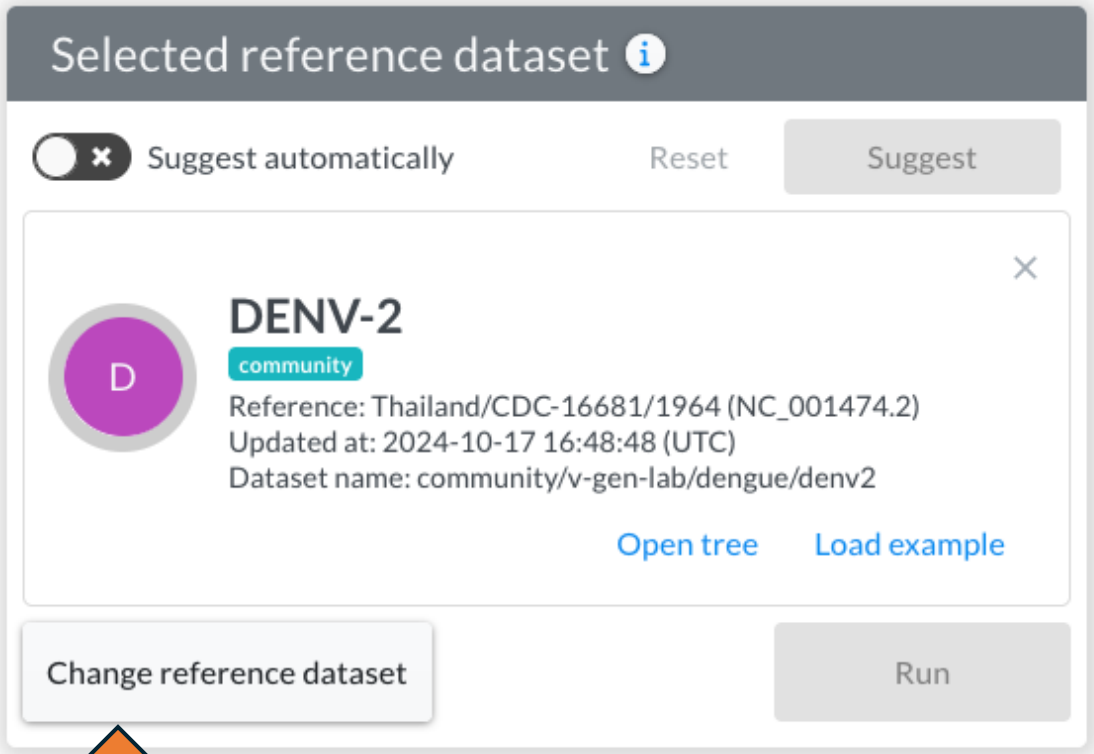
- Go to XXXXXX
- Download your data and a country specific dataset
- Launch Nextclade.org

Data input here

Nextclade v3.10.0

Clade assignment, mutation calling, and sequence quality checks

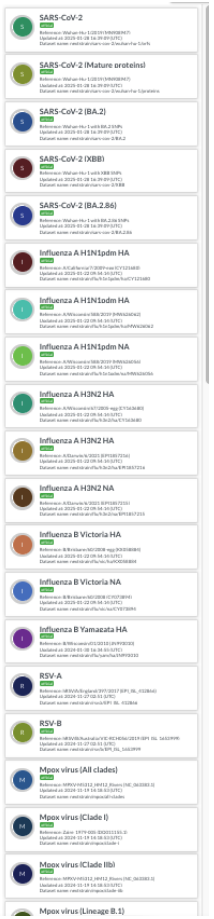




Select reference dataset here

Nextclade supports many viruses

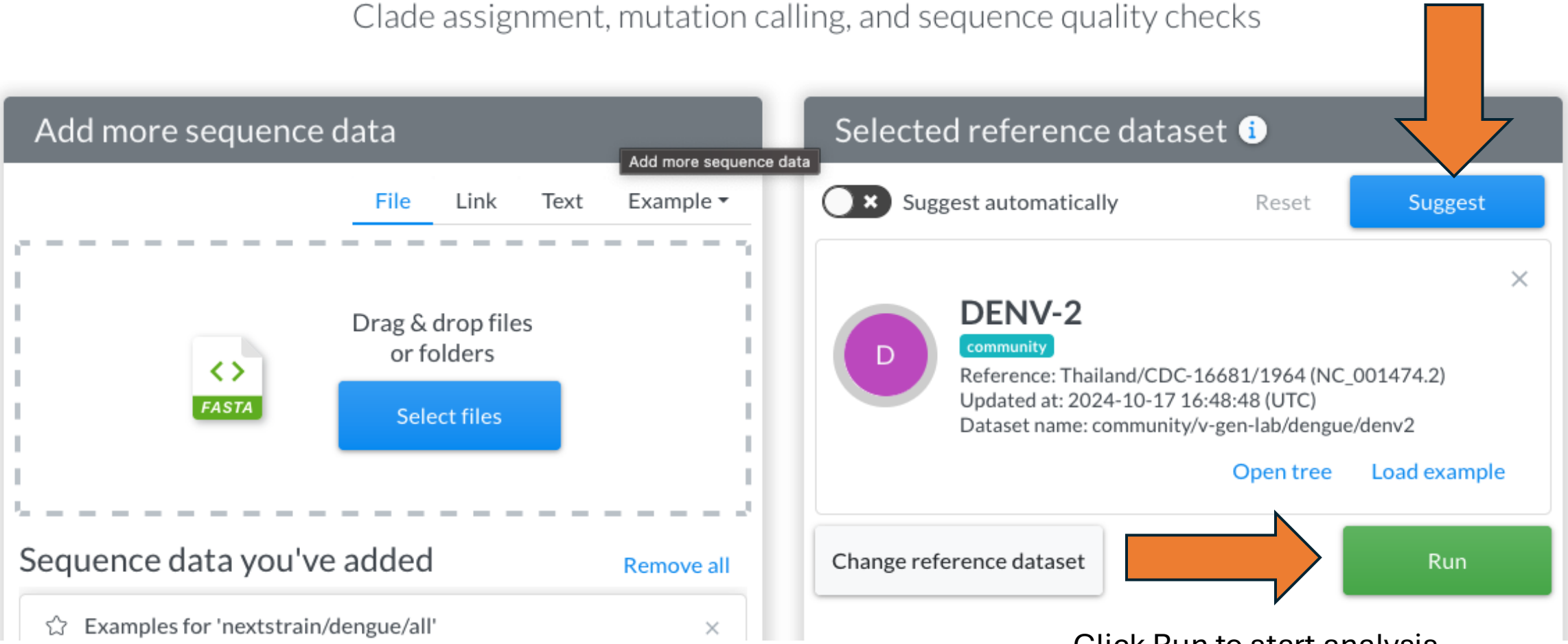
Change reference dataset



Nextclade v3.10.0

Clade assignment, mutation calling, and sequence quality checks

If unsure which ref dataset, click 'Suggest'



Click Run to start analysis

Nextclade | Start | Dataset | Results | Tree | Export | Done. Total sequences: 165. Succeeded: 165 | Settings | About | Citation | Docs | CLI | EN

Show tree

Download results

Filter results

Select gene

Bad sequence

Stretches of Ns

Frameshift

#	i	Sequence name	QC	Clade	Pangolin lineage (Nextclade)	Unaliased	Immune escape	ACE2 binding	Mut.	non-ACGTN	Ns	Cov.	Gaps	Ins.	FS	SC
0	1	USA/CA-LACPHL-AY03266/2023	N M P C F S	23F	EG.5.1.13	XBB.1.9.2.5.1.	0.39	-0.27	30	0	1229	95.1%	39	0	0	0 (1)
1	4	USA/CA-LACPHL-AY03267/2023	N M P C F S	23F	EG.5.1.16	XBB.1.9.2.5.1	0.32	-0.22	20	0	0	99.3%	30	0	0	0 (1)
2	0	OY754687								2	160	99.5%	56	0		
3	3	OY754651	N M P C F S	23F	EG.5.1	XBB.1.9.2.5.1	0.32	-0.22	22	2	3400	88.6%	9	0		
4	2	OY754528	N M P C F S	23E	GE.1	XBB.2.3.10.1	0.05	0.46	13	14	11269	62.3%	0	0	0	0
5	6	USA/WA-UW-23[illegible]46451/2023	N M P C F S	recombinant	XCR	XCR	0.37	-0.77	24	0	0	98.8%	30	0	0	0 (1)
6	7	USA/WA-UW-23102330989/2023	N M P C F S	23F	HV.1	XBB.1.9.2.5.1.	0.42	-0.26	22	0	0	98.8%	30	0	0	0 (1)
7	8	USA/CA-LACI[illegible]	[illegible]G		XBB.1.5.70	XBB.1.5.70	0.37	-0.77	18	0	756	96.7%	30	0	0	0 (1)
8	5	OY754526	[illegible]23E		GJ.2	XBB.2.3.3.2	0.04	0.55	19	0	1944	93.5%	9	0	0	0
9	9	OY754632	N M P C F S	23I	JN.3	BA.2.86.1.3	0.64	-0.09	67	2	185	99.4%	3	0	0	0
10	12	OY754626	N M P C F S	23F	EG.5.1.3	XBB.1.9.2.5.1.	0.32	-0.22	20	2	143	99.5%	9	0	0	0 (1)
11	10	USA/CA-LACPHL-AY03247/2023	N M P C F S	23B	XBB.1.16.6	XBB.1.16.6	0.34	-0.37	25	0	129	98.8%	32	0	0 (1)	0 (1)
12	11	USA/CA-LACPHL-AY03264/2023	N M P C F S	23A	HR.1.1	XBB.1.5.77.1.1	0.05	0.46	26	0	1	99.2%	30	0	0	0 (1)
13	14	OY754681	N M P C F S	23B	JF.1	XBB.1.16.6.1	0.39	-0.92	25	2	83	99.7%	56	0	0	0 (1)
14	13	USA/CA-LACPHL-AY03246/2023	N M P C F S	23C	DV.7.1.4	BA.2.75.3.4.1.	0.80	-1.46	50	0	1398	94.6%	27	0	0	0
15	15	OY754673	N M P C F S	recombinant	XCH.1	XCH.1	0.37	-0.77	27	5	142	99.5%	9	0	0	0 (1)
16	16	USA/LA-EVTL20384/2023	N M P C F S	23A	JD.1.1	XBB.1.5.102.1	0.39	-1.49	24	0	0	99.9%	21	0	0	0 (1)
17	18	OY754523	N M P C F S	23F	EG.5.1.3	XBB.1.9.2.5.1.	0.32	-0.22	21	2	143	99.5%	9	0	0	0 (1)
18	19	USA/CA-LACPHL-AY03248/2023	N M P C F S	23D	FL.2.5	XBB.1.9.1.2.5	0.03	0.61	22	0	1	99.2%	30	0	0	0 (1)
19	17	USA/NY-PBRI-NYC2O026/2023	N M P C F S	23F	EG.5.1.4	XBB.1.9.2.5.1.	0.58	0.00	25	0	0	99.6%	56	0	0	0 (1)
20	20	OY754683	N M P C F S	23I	BA.2.86.1	BA.2.86.1	0.64	-0.09	61	2	182	99.4%	3	0	0	0
21	21	USA/CA-LACPHL-AY03262/2023	N M P C F S	23F	HK.31	XBB.1.9.2.5.1.	0.32	-0.22	24	0	0	99.3%	30	0	0	0 (1)
22	22	OY754698	N M P C F S	23D	EG.9.1	XBB.1.9.2.9.1	0.10	-0.01	21	1	331	98.9%	9	0	0	0 (2)
23	23	USA/MN-MDH-37211/2023	N M P C F S	23B	HF.1	XBB.1.16.13.1	0.05	0.46	27	0	0	99.5%	56	0	0	0 (1)
24	24	OY754610	N M P C F S	23F	EG.5.1.6	XBB.1.9.2.5.1.	0.32	-0.22	20	3	142	99.5%	9	0	0	0 (1)
25	26	OY754688	N M P C F S	23I	JN.3.1	BA.2.86.1.3.1	-0.00	0.00	69	2	79	99.7%	68	0	1 (2)	0
26	25	OY754532	N M P C F S	23B	XBB.1.16.15	XBB.1.16.15	0.06	0.40	21	0	1596	94.7%	9	0	0	0 (1)
27	29	USA/CA-LACPHL-AY03275/2023	N M P C F S	23G	GK.1.10	XBB.1.5.70.1.1	0.47	-1.38	32	0	4	99.2%	27	0	0	0 (1)
28	27	USA/NY-PBRI-NYC2O027/2023	N M P C F S	23F	HV.1.6	XBB.1.9.2.5.1	0.69	-0.19	29	0	19	99.6%	56	0	0	0 (1)

Genome annotation ?

Nucleotide sequence

5000 10000 15000 20000 25000

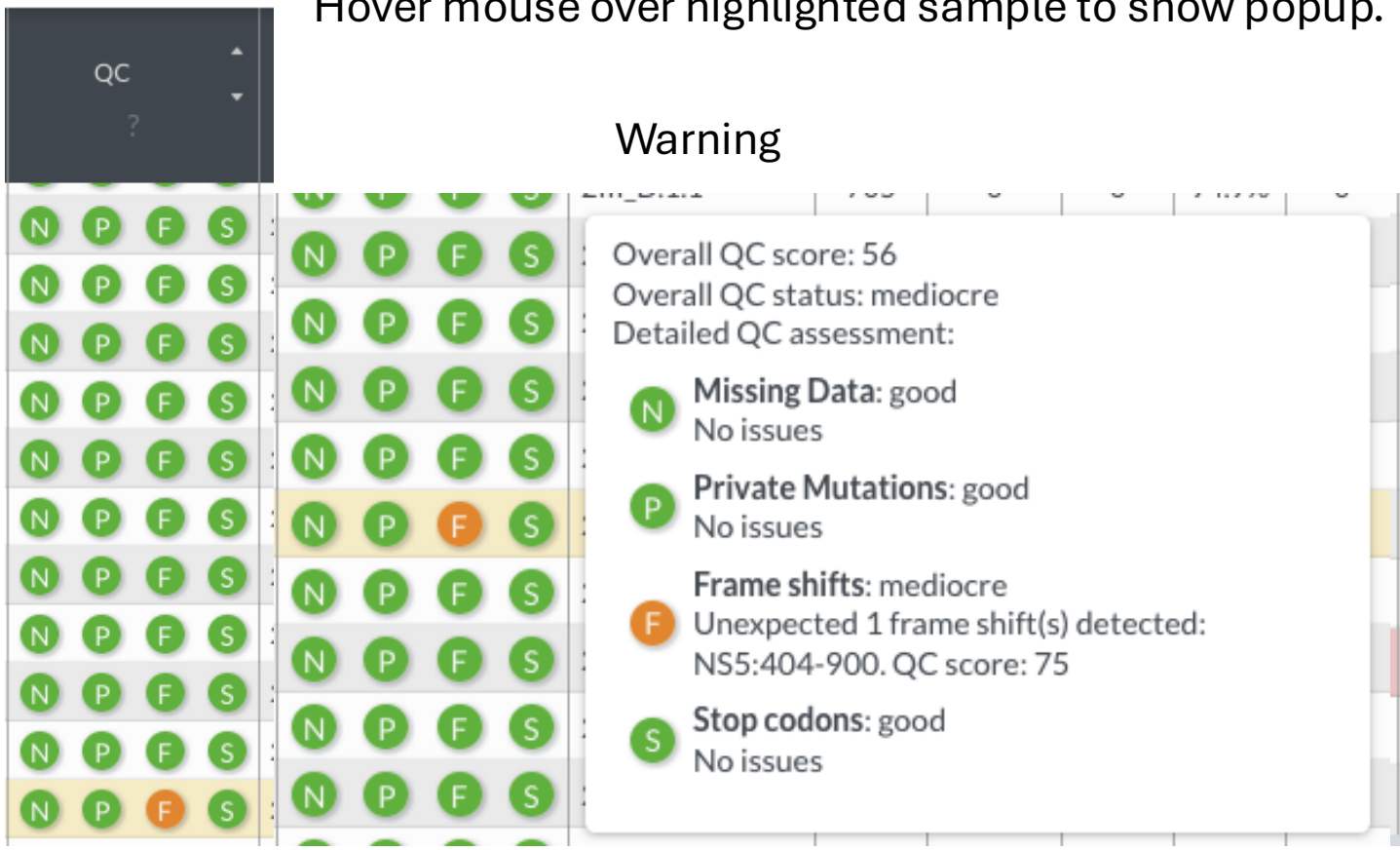
version 3.0.0-alpha.0 (commit: 4deeba5, branch: master)

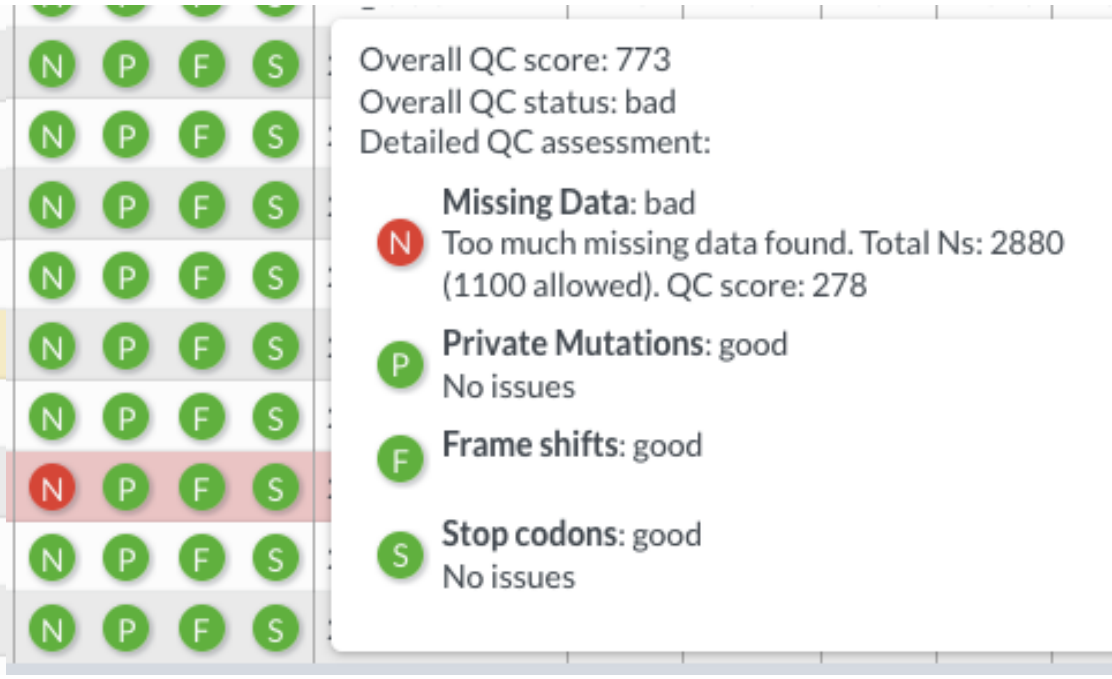
Sequence name	QC	Clade	Mut.	non-ACGTN	Ns	Cov.	Gaps	Ins.	FS	SC
?	?	?	?	?	?	?	?	?	?	?

- “Mut.”: number of mutations with respect to the reference sequence
- “non-ACGTN”: number of ambiguous nucleotides that are not *N*
- “Ns”: number of missing nucleotides indicated by *N*
- “Gaps”: number of nucleotides that are deleted with respect to the reference sequence
- “Ins.”: number of nucleotides that are inserted with respect to the reference sequence
- “FS”: Number of uncommon frame shifts (total number, including common frame shifts are in parentheses)
- “SC”: Number of uncommon premature stop codons (total number, including common premature stops are in parentheses)

QC Field

Hover mouse over highlighted sample to show popup.





N	P	F	S	2III_C.1.1	508	0	2880	68.0%	0	0	0	0

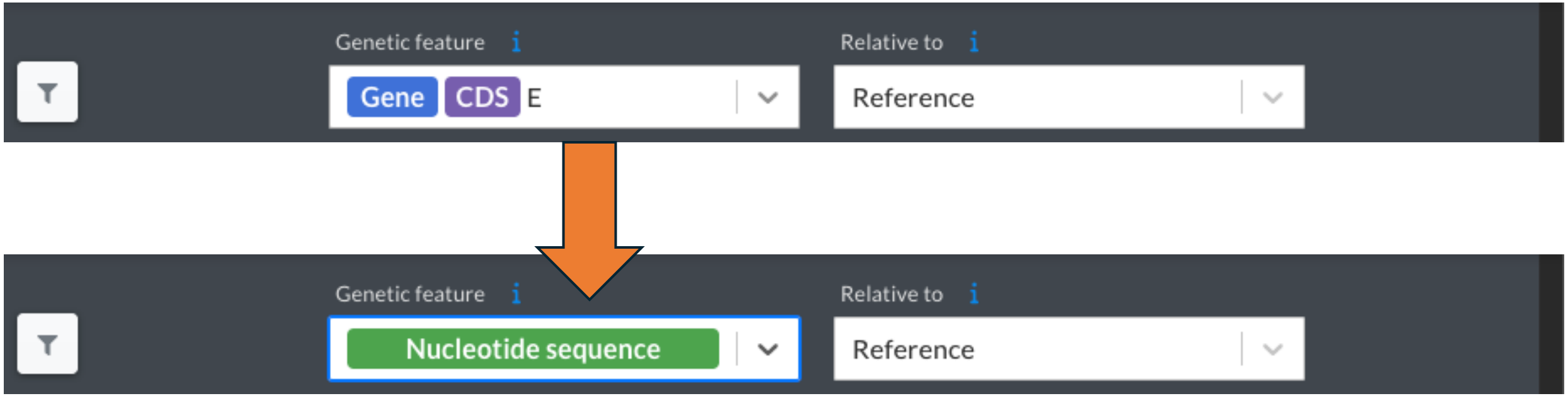
Why was it marked as fail? Too much missing data (N or -)

QC Metrics

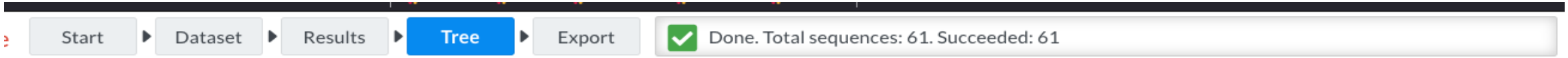
- N Missing Data – threshold 1000
- P Private Mutations – cutoff 216, typical 72
- F Frame Shift
- S Stop Codon

- Nextclade implements a variety of quality control metrics to quickly spot problems in your sequencing/assembly pipeline.
- Bad sequences are colored red, mediocre ones yellow and good ones white. You can view detailed results of the QC metrics by hovering your mouse over a sequences QC entry:

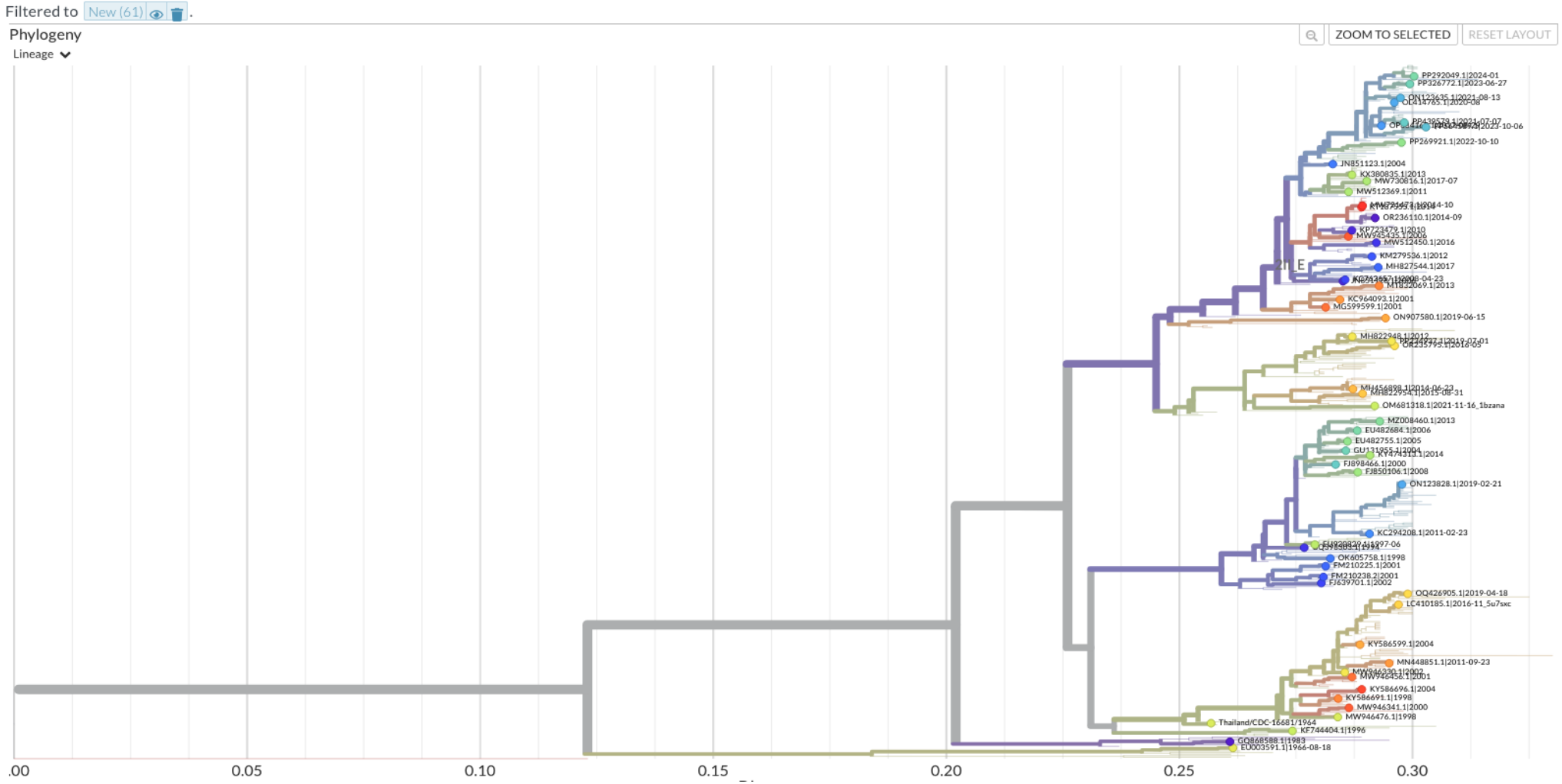
Change Genetic Feature to Nucleotide to show full genome



Note: In Dengue the E gene is routinely sequenced as it's genetically distinct between serotypes and provides sufficient resolution for clade information and possible lineage information.



Click on the "Tree" button to see where your sequences have been placed.
The tree is nearly identical to Next*strain* tree, so interact with it.



Nextclade Demonstration

- Loading consensus sequences
- Selecting datasets
- Interpreting results
- Exporting results

Nextclade Practice

Use the amplicon data you generate in the lab to answer these questions:

1. What is the clade/lineage of each sample?
2. Check each samples QC metrics, which would you use for:
   - Clade/Lineage calling?
   - Phylogenetic analysis?
   - Speciation?
3. Check for odd genomic features (premature stop, gaps, etc).
4. Which samples would you pass/fail?

Questions? + Resources

- Nextclade is also a CLI tool and is avaliable in Galaxy.
  - https://docs.nextstrain.org/projects/nextclade/en/stable/user/nextclade-cli/index.html
- The web interface is excellent, provides all the features of the CLI and even more.
- It's possible to create your own reference datasets for Nextclade but it requires knowledge of the organism and advanced bioinformatic skills.