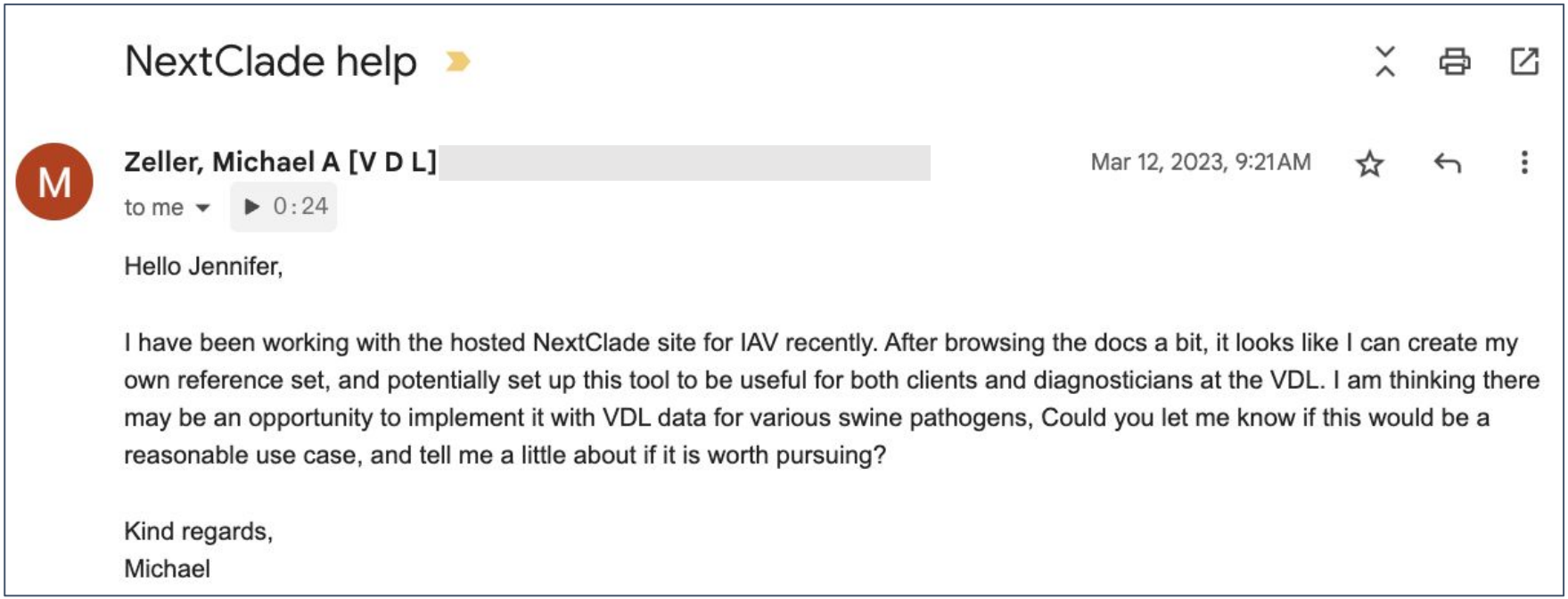
Creating a Nextclade dataset

- Bedford Lab Meeting -

Jennifer Chang, Ph.D.
Bioinformatic Analyst III
Fred Hutchinson Cancer Center

Motivations



NextClade help

Zeller, Michael A [V D L] Mar 12, 2023, 9:21AM

to me ▶ 0:24

Hello Jennifer,

I have been working with the hosted NextClade site for IAV recently. After browsing the docs a bit, it looks like I can create my own reference set, and potentially set up this tool to be useful for both clients and diagnosticians at the VDL. I am thinking there may be an opportunity to implement it with VDL data for various swine pathogens, Could you let me know if this would be a reasonable use case, and tell me a little about if it is worth pursuing?

Kind regards,
Michael

Motivations

NextClade help

Zeller, Michael A [V D L] Mar 12, 2023, 9:21AM

to me ▶ 0:24

Hello Jennifer,

I have been working with the hosted NextClade site for IAV recently. After browsing the docs a bit, it looks like I can create my own reference set, and potentially set up this tool to be useful for both clients and diagnosticians at the VDL. I am thinking there may be an opportunity to implement it with VDL data for various swine pathogens, Could you let me know if this would be a

Jennifer Chang Mar 12, 2023, 9:36 AM

to Michael, bcc: hello ▶ 0:19

Hi Michael,

Yes, that sounds like a reasonable use case! It's been a while since I looked at the Nextclade docs, but yes you'd need to create your own reference set.
Also, Richard Neher spec'd out a bootstrap script to take a reference to build a nextclade dataset in case it's helpful

- https://github.com/nextstrain/nextclade_dataset_template

Let me know if you hit any challenges, and we can probably loop someone in to answer questions,

Jennifer

Generalize Ingest: dengue ingest ;
Discussion Nextclade Dengue

Agenda

- What is Nextclade?
  - Statement of need
  - What is a Nextclade dataset?
- How do we create a Nextclade dataset?
  - Nextclade template script
  - Nextalign vs Augur align
- Checking a Nextclade dataset
  - validation and basic checks
- Conclusions

What is Nextclade?

(1) Assess the quality of the sequence
(2) Assign it to a known clade or type
(3) Compare it to a reference sequence to detect evolutionary changes

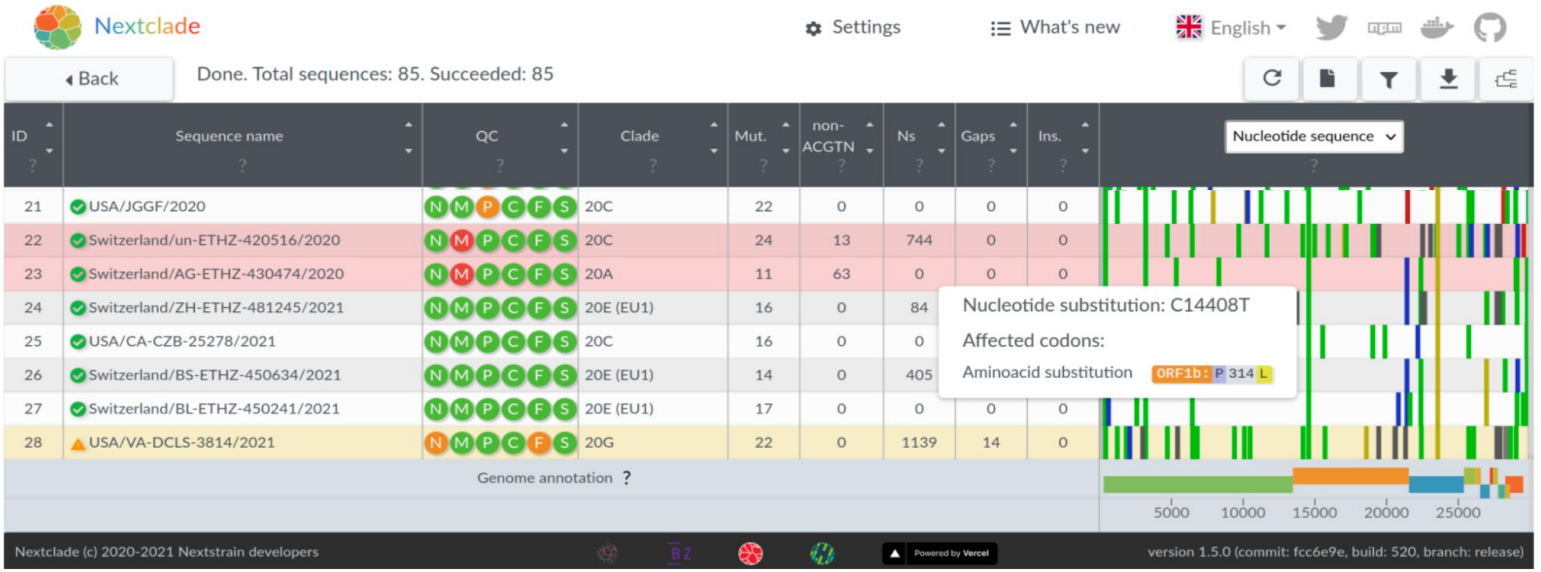
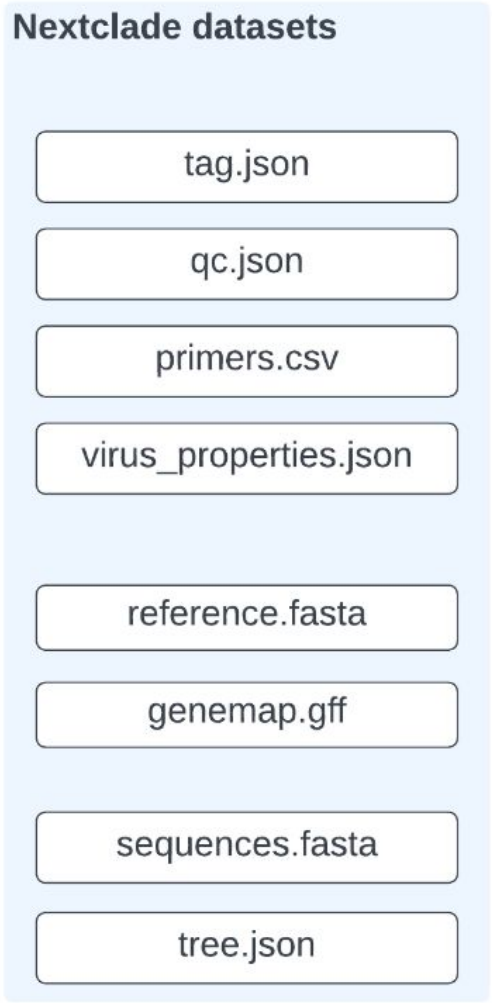


Figure 1: Overview of the results page with clade assignments, QC metrics, and the nucleotide mutation view. The results can be explored interactively and exported in standard tabular file formats.

Aksamentov et al., JOSS 2021

What is a Nextclade dataset?



Nextclade datasets

Nextclade dataset is a set of input data files required for Nextclade to run the analysis:

- reference (root) sequence (`reference.fasta`)
- reference tree (`tree.json`)
- quality control configuration (`qc.json`)
- gene map (`genemap.gff`)
- PCR primers (`primers.csv`)
- virus properties (`virus_properties.json`)

See also: Input files

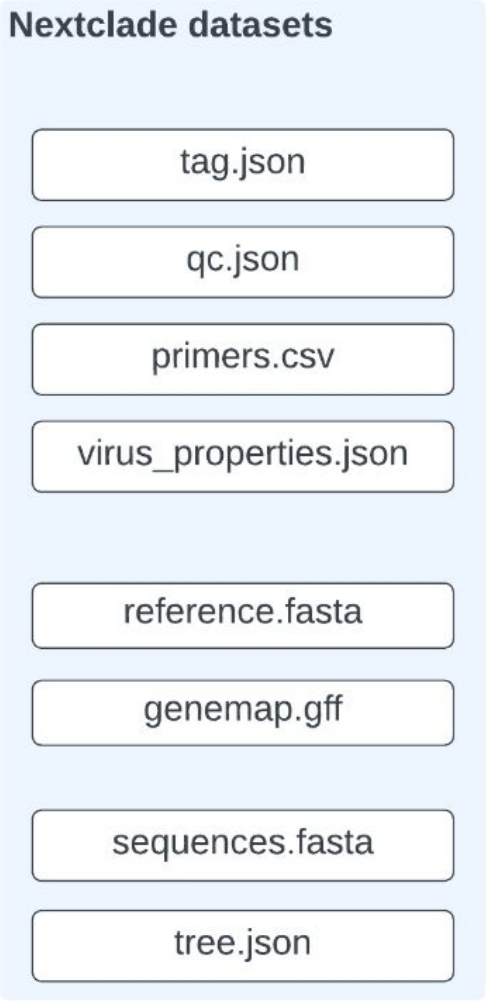
Dataset might also include example sequence data (`sequences.fasta`) - typically a diverse set of query sequences that represents major clades, used for demonstration and highlights analysis features of Nextclade. Most of the time you want to analyze your own sequence data.

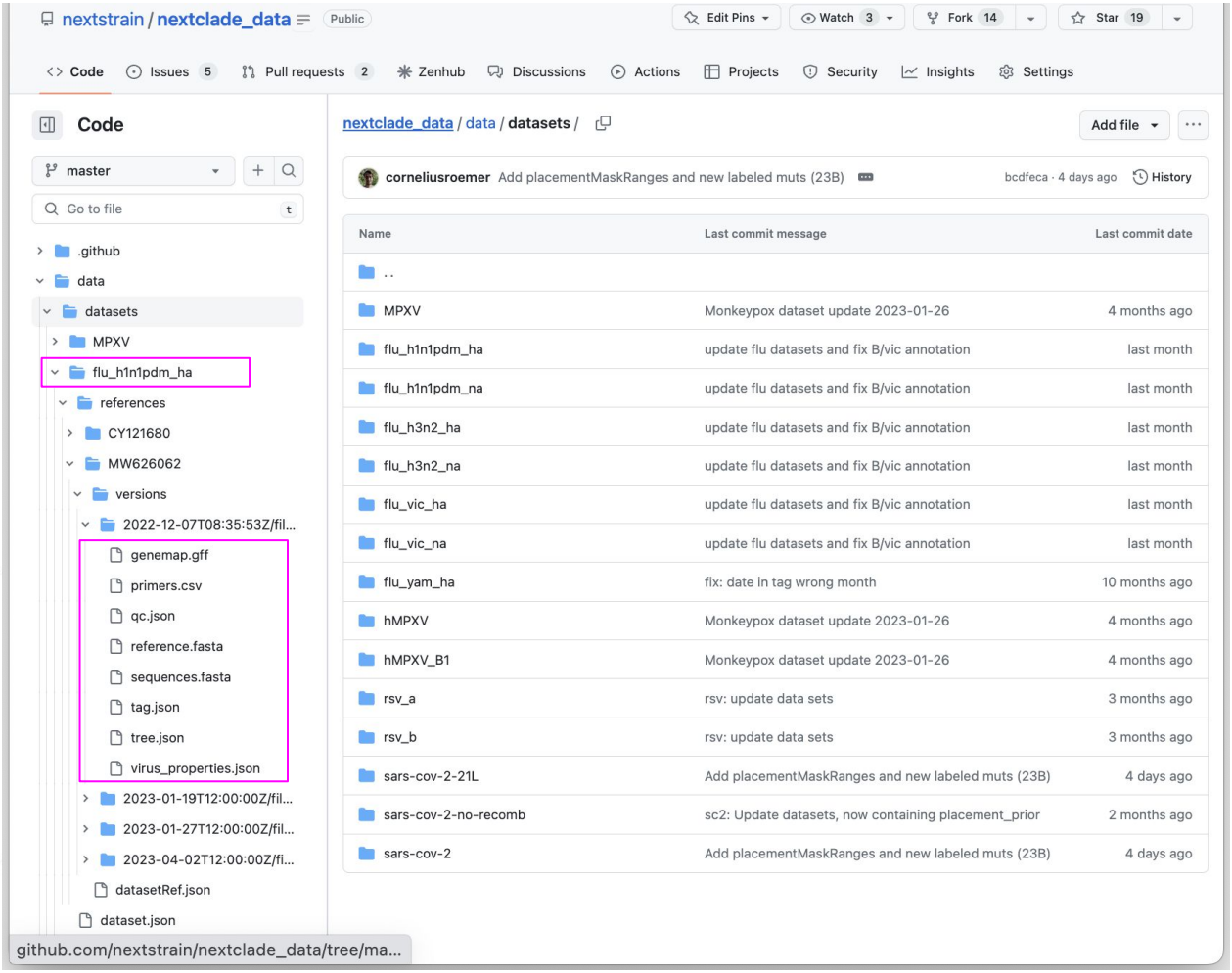
Dataset also includes a file `tag.json` which contains version tag and other properties of the dataset. This file is currently not used by Nextclade and serves only for informational purposes.

An instance of a dataset is a directory containing the dataset files or an equivalent zip archive.

Nextclade docs: datasets

What is a Nextclade dataset?



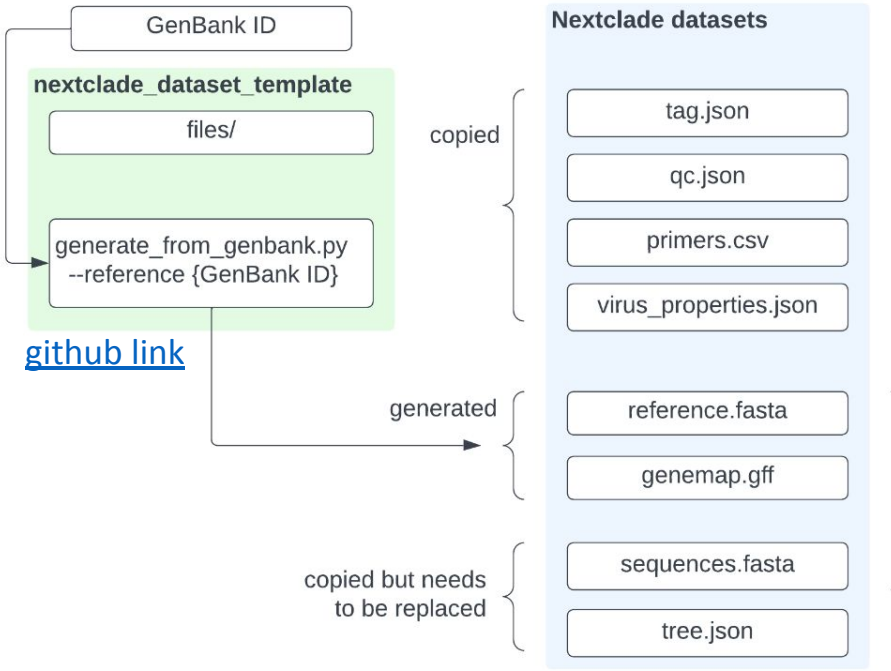


https://github.com/nextstrain/nextclade_data/tree/master/data/datasets

Agenda

- What is Nextclade?
  - Statement of need
  - What is a Nextclade dataset?
- **How do we create a Nextclade dataset?**
  - Nextclade template script
  - Nextalign vs Augur align
- Checking a Nextclade dataset
  - validation and basic checks
- Conclusions

Picking a reference

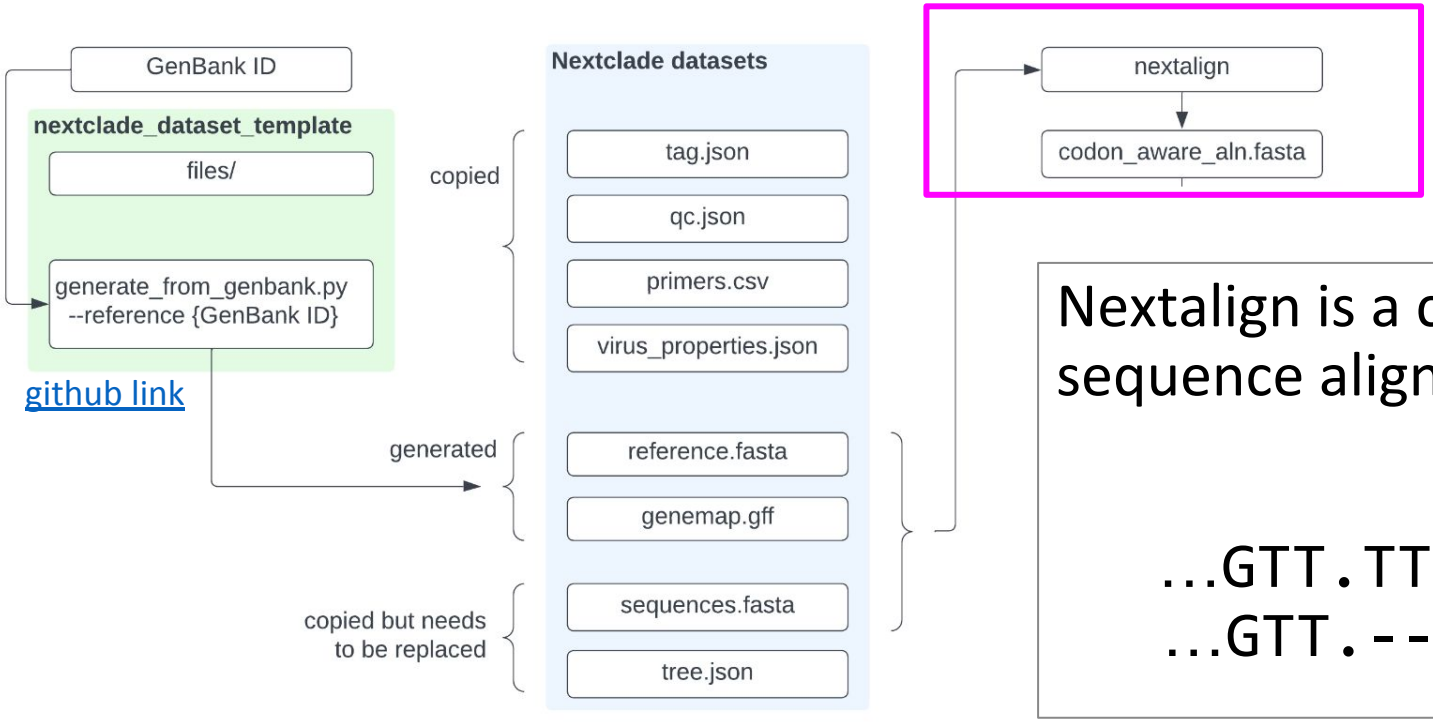


Pick a reference that is close to the base of the tree.

Refseq reference
(e.g. NCBI Virus: Dengue)

```
git clone https://github.com/nextstrain/nextclade_dataset_template.git
cd nextclade_dataset_template

python generate_from_genbank.py \
  --reference NC_001477 \
  --output-dir denv1_dataset
```

Nextalign or augur align



Nextalign is a codon-aware sequence aligner

```
...GTT.TTT.TAC...
...GTT.---.TAC...
```

```
cat sequences.fasta \
 | nextalign run \
  --jobs=`nproc` \
  --reference reference.fasta \
  --genemap genemap.gff \
  --output-translations translations_{gene}.txt \
  --output-fasta aln.fasta
```

Nextclade docs: nextalign-cli

Nextalign or augur align

Nextclade can use a genome annotation to make the alignment more interpretable. Sometimes, the placement of a sequence deletion or insertion is ambiguous as in the following example. The gap could be moved forward or backward by one base with the same number of matches:

```
Reference  : ... | GTT | TAT | TAC | ...
Alignment 1: ... | GTT | --- | TAC | ...
Alignment 2: ... | GT- | --T | TAC | ...
Alignment 3: ... | GTT | T-- | -AC | ...
```

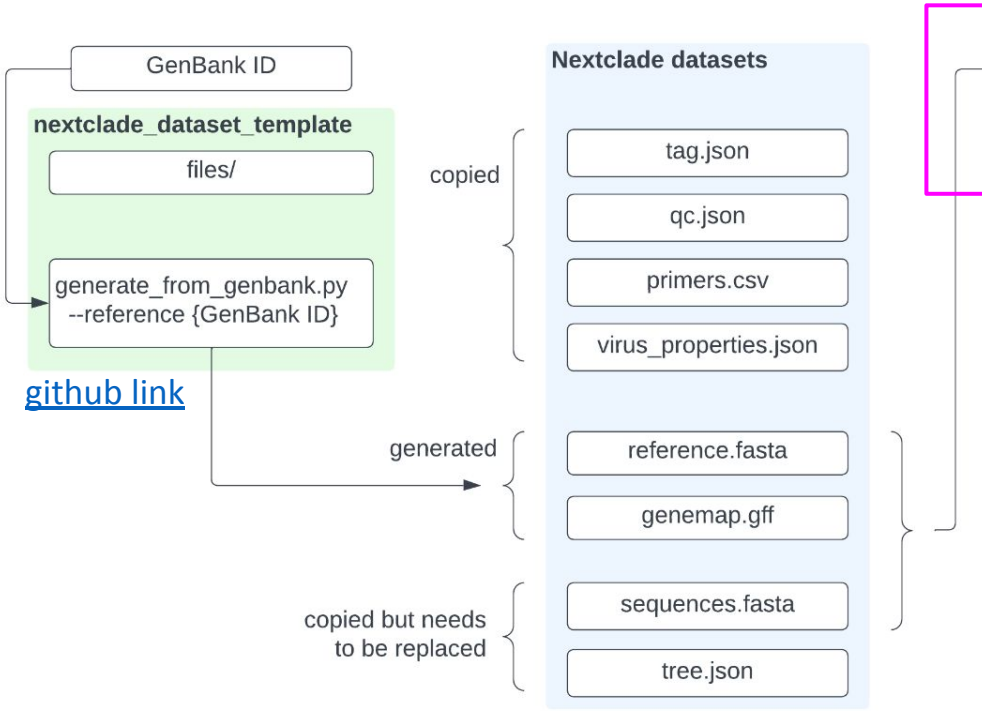
If a genome annotation is provided, Nextclade will use a lower gap-open-penalty at the beginning of a codon (delimited by the `|` characters in the schema above), thereby locking a gap in-frame if possible. Similarly, nextalign preferentially places gaps outside of genes in case of ambiguities.

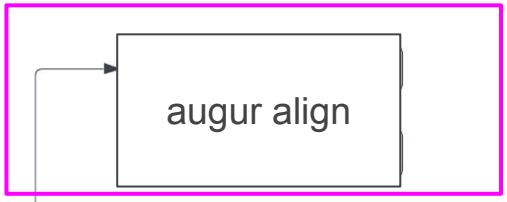
denv1_dataset > files > genemap.gff

```
1  ##gff-version 3
2  ##sequence-region NC_001477.1 1 10735
3  NC_001477.1 feature gene  95  10273 . + . codon_start=1;gene=POLY;gene_name=POLY
4  NC_001477.1 feature gene  710 934 . + . codon_start=1;gene=M;gene_name=M
5  NC_001477.1 feature gene  935 2419  . + . codon_start=1;gene=E;gene_name=E
```

Nextclade docs: algorithm/01-sequence-alignment.html

Nextalign or augur align

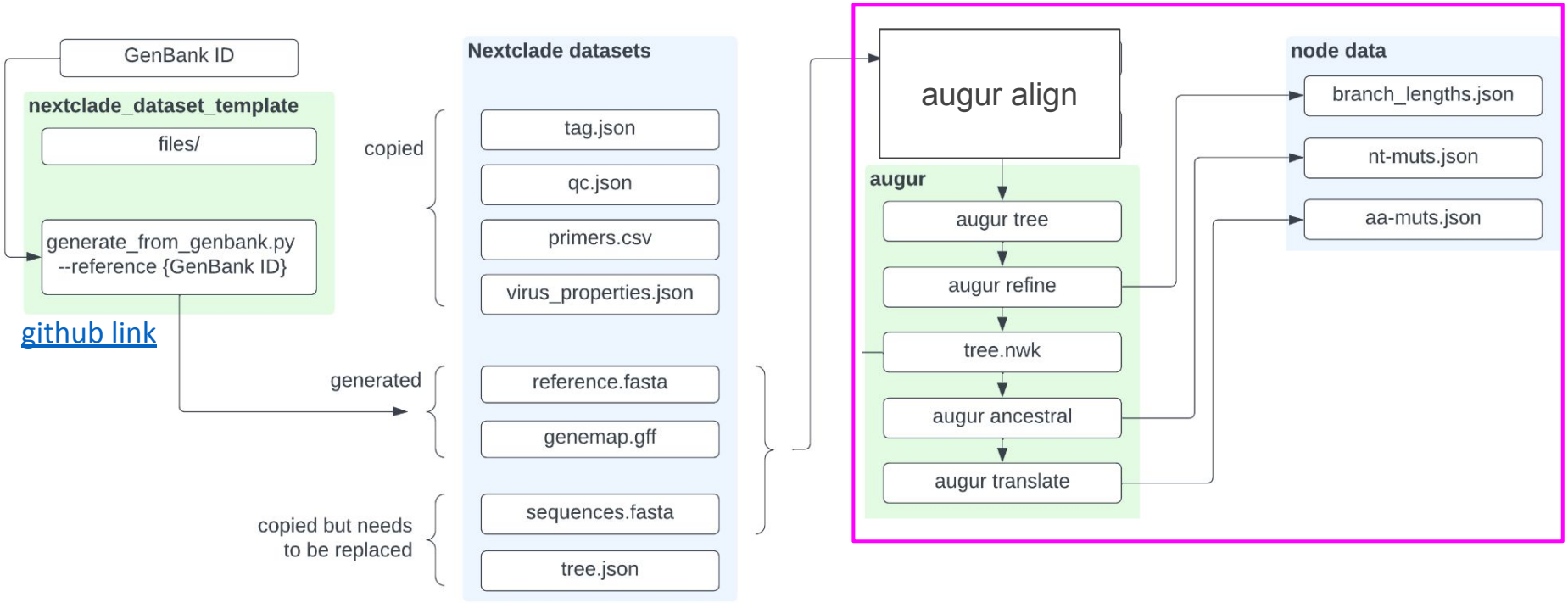




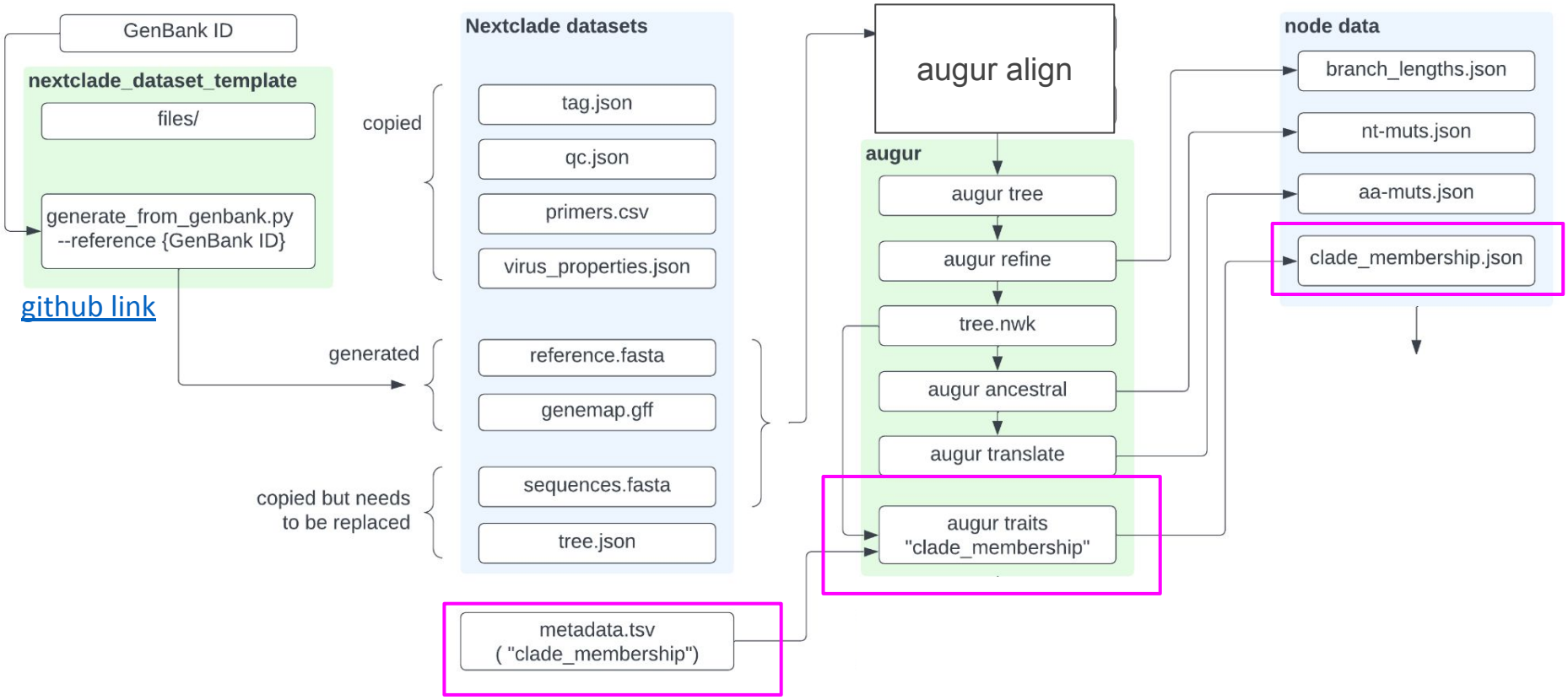
Could use augur align instead, especially if sequences have >10% difference.

```
augur align \
  --sequences sequences.fasta \
  --reference-sequence reference.fasta \
  --output results/aln.fasta \
  --fill-gaps \
  --nthreads `nproc`
```

Standard augur commands

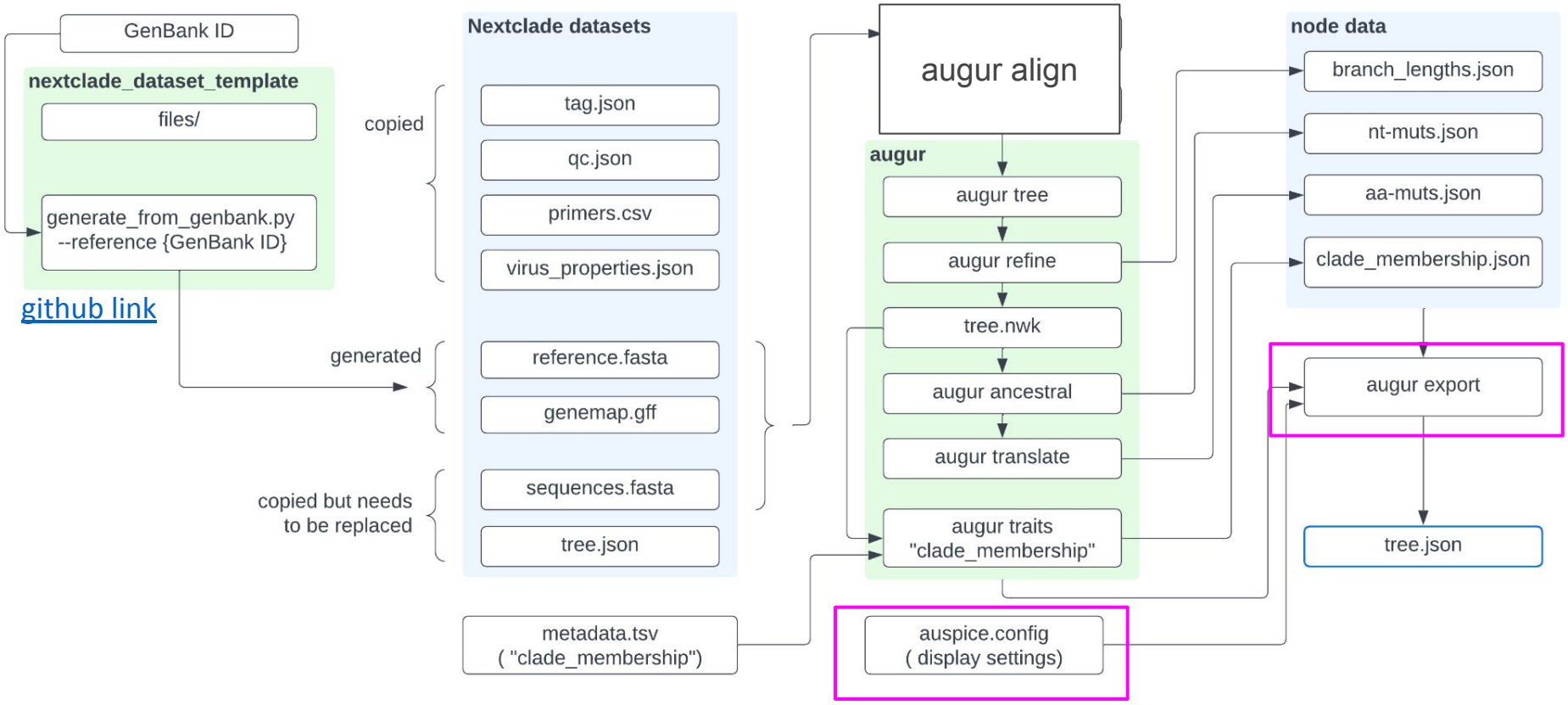


Clade membership

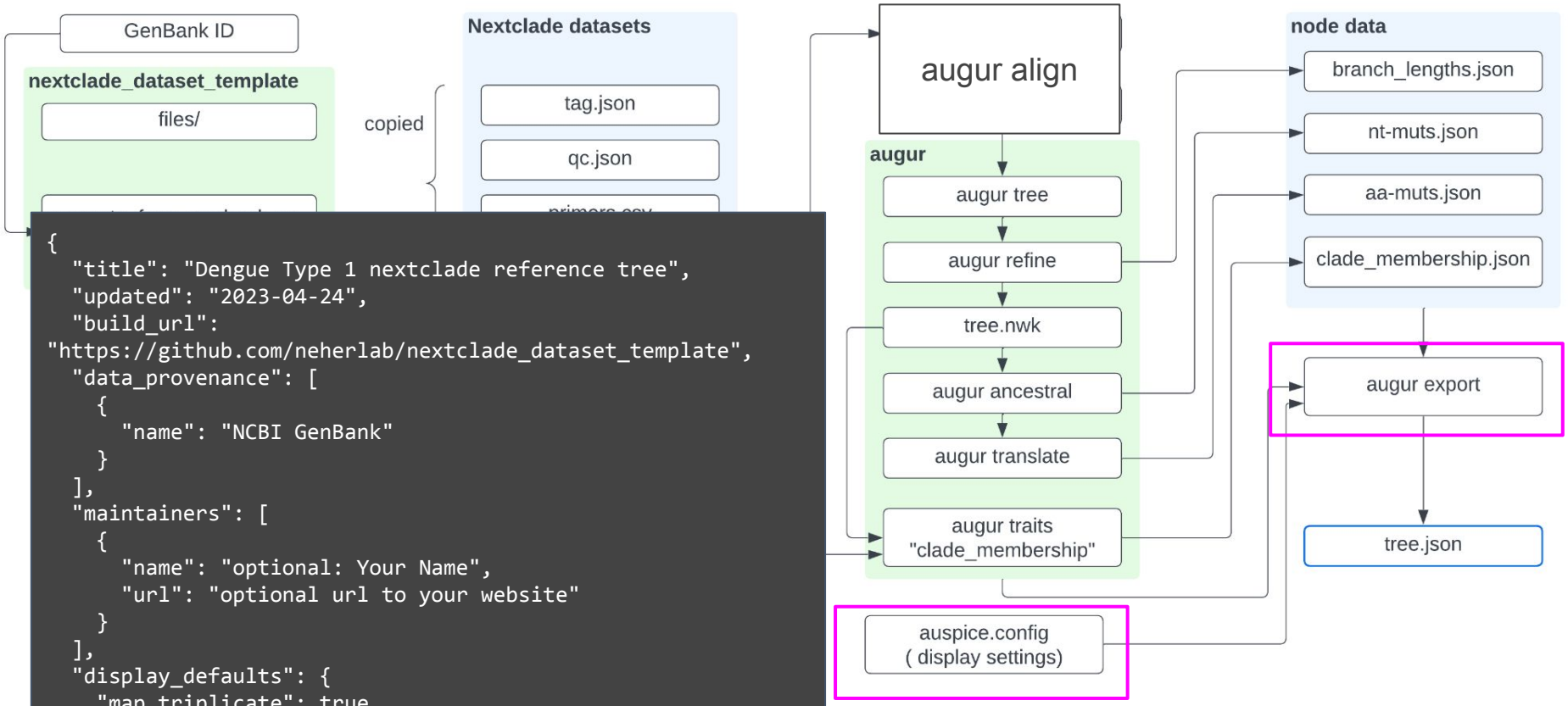


metadata.tsv should have "strain" and "clade_membership" columns

Augur export

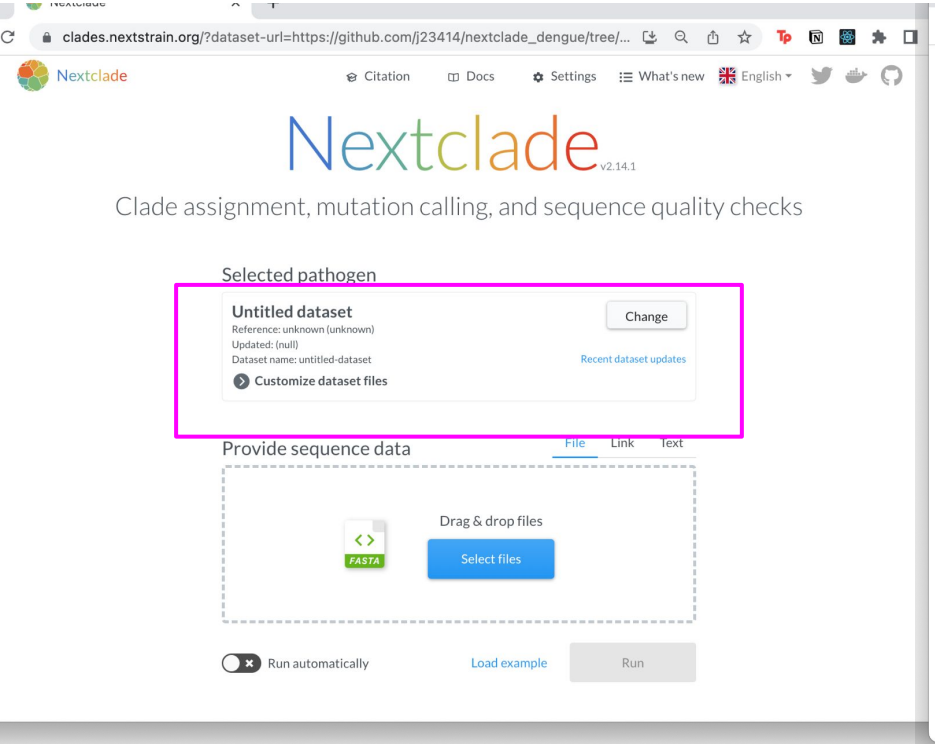


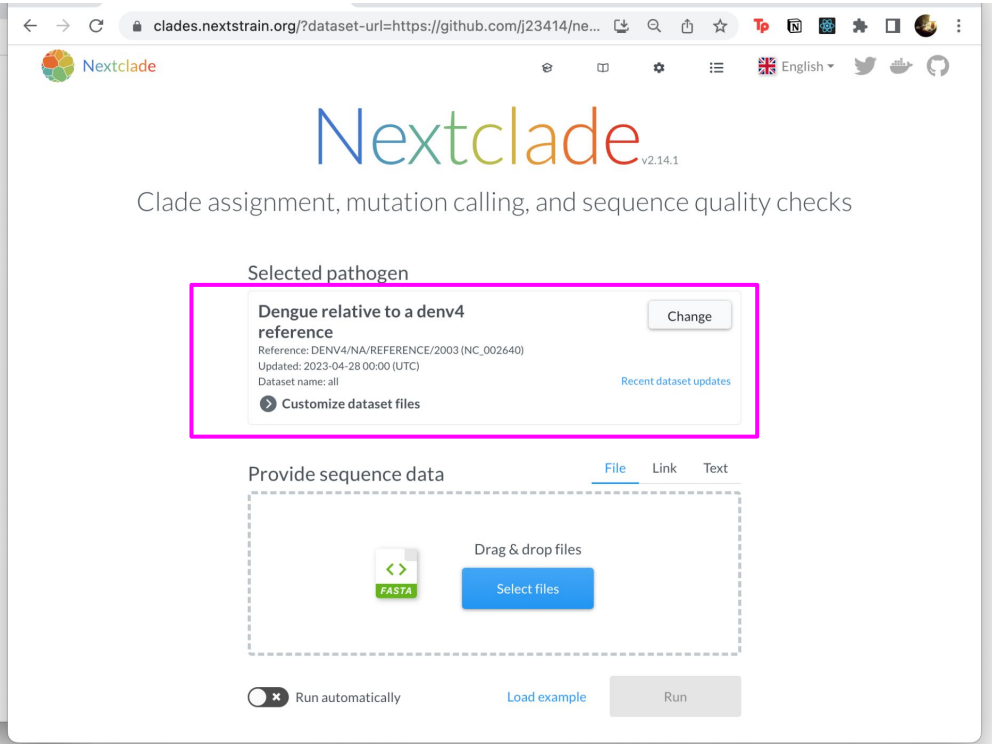
Augur export




```
{
  "title": "Dengue Type 1 nextclade reference tree",
  "updated": "2023-04-24",
  "build_url":
"https://github.com/neherlab/nextclade_dataset_template",
  "data_provenance": [
    {
      "name": "NCBI GenBank"
    }
  ],
  "maintainers": [
    {
      "name": "optional: Your Name",
      "url": "optional url to your website"
    }
  ],
  "display_defaults": {
    "map_triplicate": true,
    "color_by": "clade_membership",
    "distance_measure": "div"
  }
}
```

Augur export

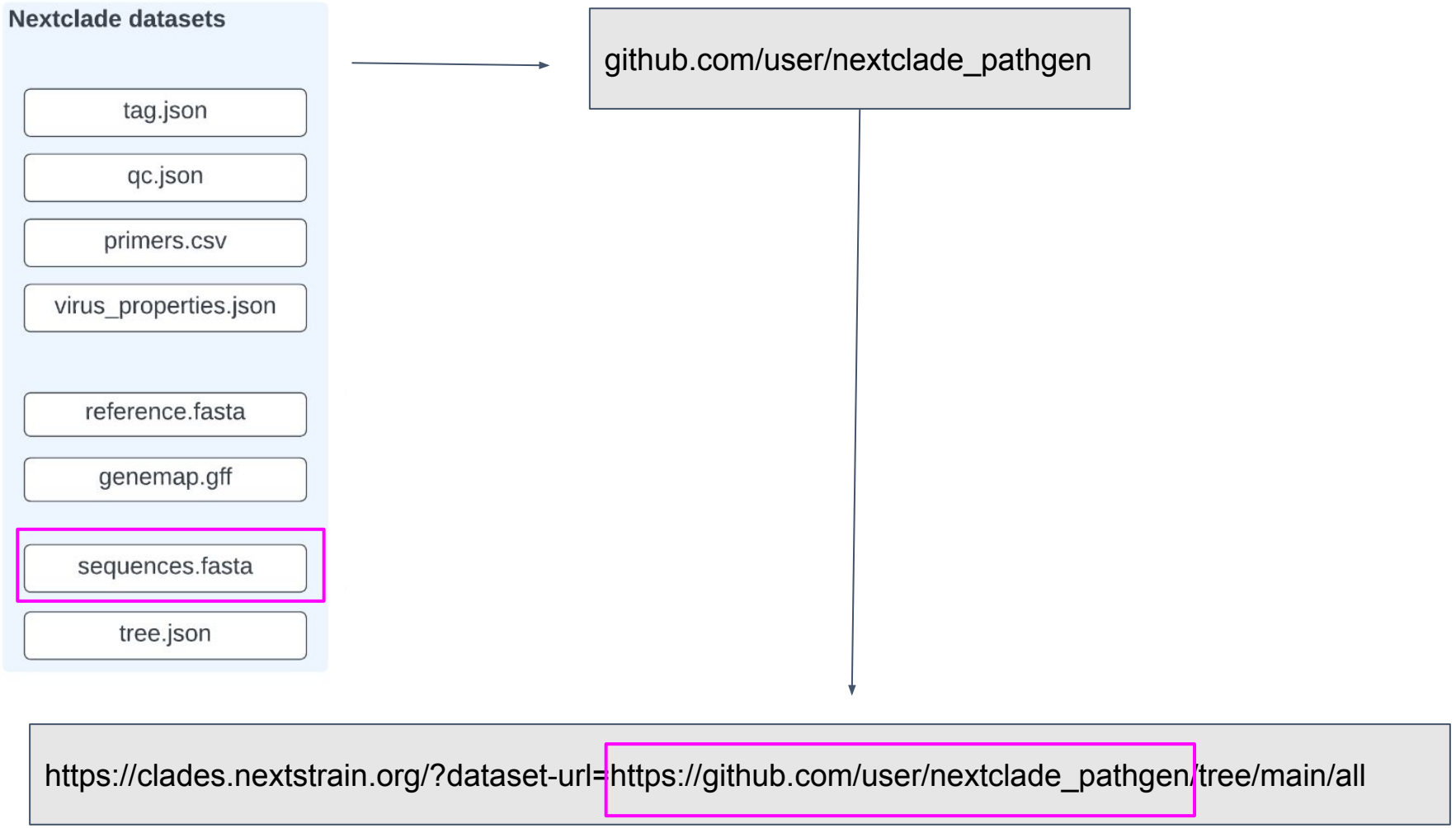




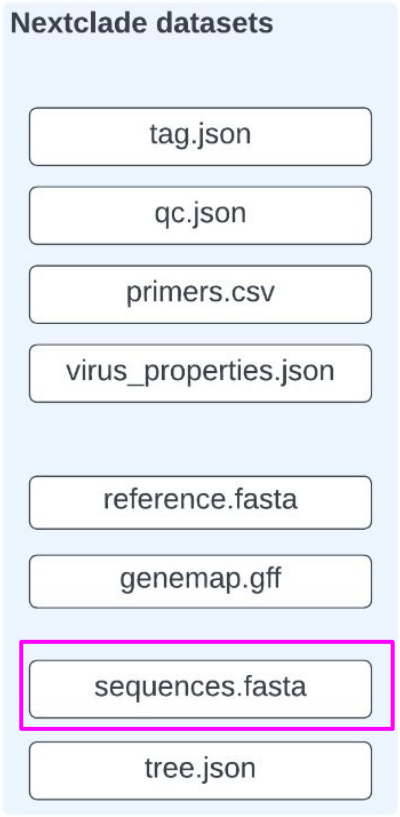
Agenda

- What is Nextclade?
  - Statement of need
  - What is a Nextclade dataset?
- **How do we create a Nextclade dataset?**
  - Nextclade template script
  - Nextalign vs Augur align
- Checking a Nextclade dataset
  - validation and basic checks
- Conclusions

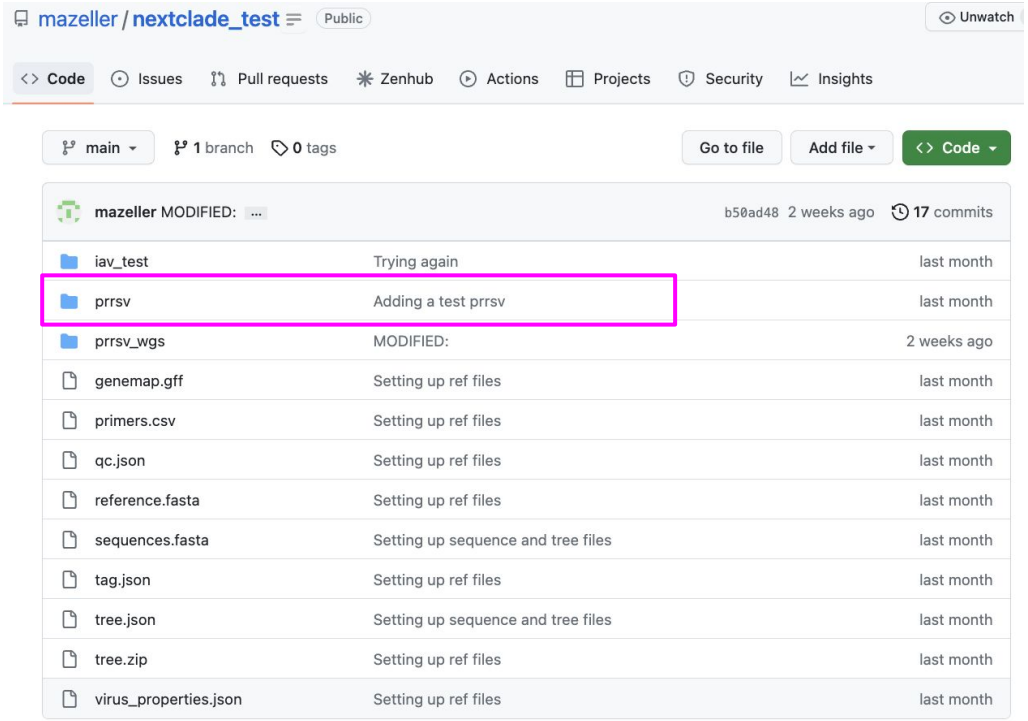
Validate Nextclade dataset



Validate Nextclade PRRS - env protein

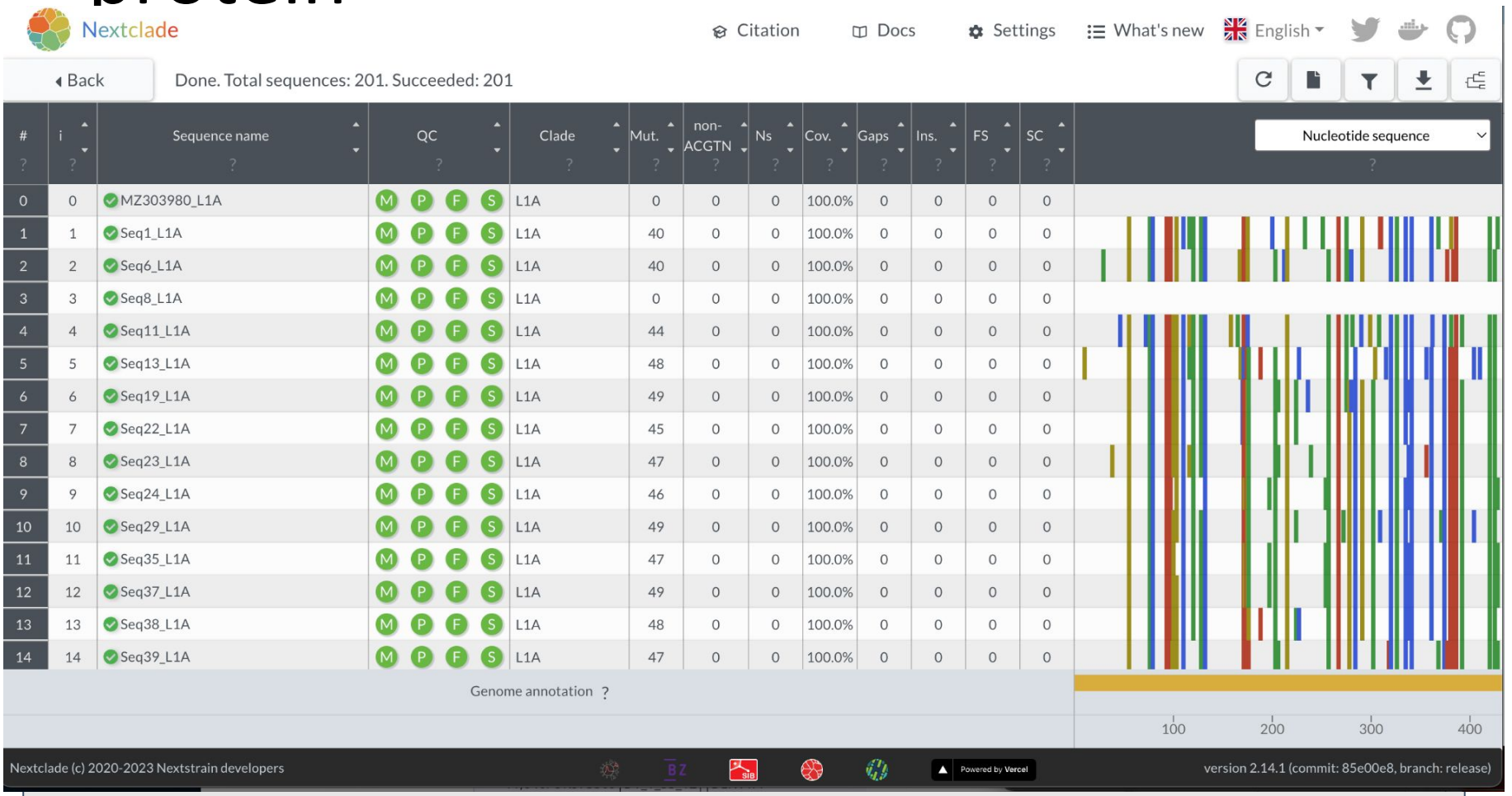


https://github.com/mazeller/nextclade_test



https://clades.nextstrain.org/?dataset-url=https://github.com/mazeller/nextclade_test/tree/main/prrsv

Validate Nextclade PRRS - env protein



#	i	Sequence name	QC	Clade	Mut.	non-ACGTN	Ns	Cov.	Gaps	Ins.	FS	SC
0	0	MZ303980_L1A	M P F S	L1A	0	0	0	100.0%	0	0	0	0
1	1	Seq1_L1A	M P F S	L1A	40	0	0	100.0%	0	0	0	0
2	2	Seq6_L1A	M P F S	L1A	40	0	0	100.0%	0	0	0	0
3	3	Seq8_L1A	M P F S	L1A	0	0	0	100.0%	0	0	0	0
4	4	Seq11_L1A	M P F S	L1A	44	0	0	100.0%	0	0	0	0
5	5	Seq13_L1A	M P F S	L1A	48	0	0	100.0%	0	0	0	0
6	6	Seq19_L1A	M P F S	L1A	49	0	0	100.0%	0	0	0	0
7	7	Seq22_L1A	M P F S	L1A	45	0	0	100.0%	0	0	0	0
8	8	Seq23_L1A	M P F S	L1A	47	0	0	100.0%	0	0	0	0
9	9	Seq24_L1A	M P F S	L1A	46	0	0	100.0%	0	0	0	0
10	10	Seq29_L1A	M P F S	L1A	49	0	0	100.0%	0	0	0	0
11	11	Seq35_L1A	M P F S	L1A	47	0	0	100.0%	0	0	0	0
12	12	Seq37_L1A	M P F S	L1A	49	0	0	100.0%	0	0	0	0
13	13	Seq38_L1A	M P F S	L1A	48	0	0	100.0%	0	0	0	0
14	14	Seq39_L1A	M P F S	L1A	47	0	0	100.0%	0	0	0	0

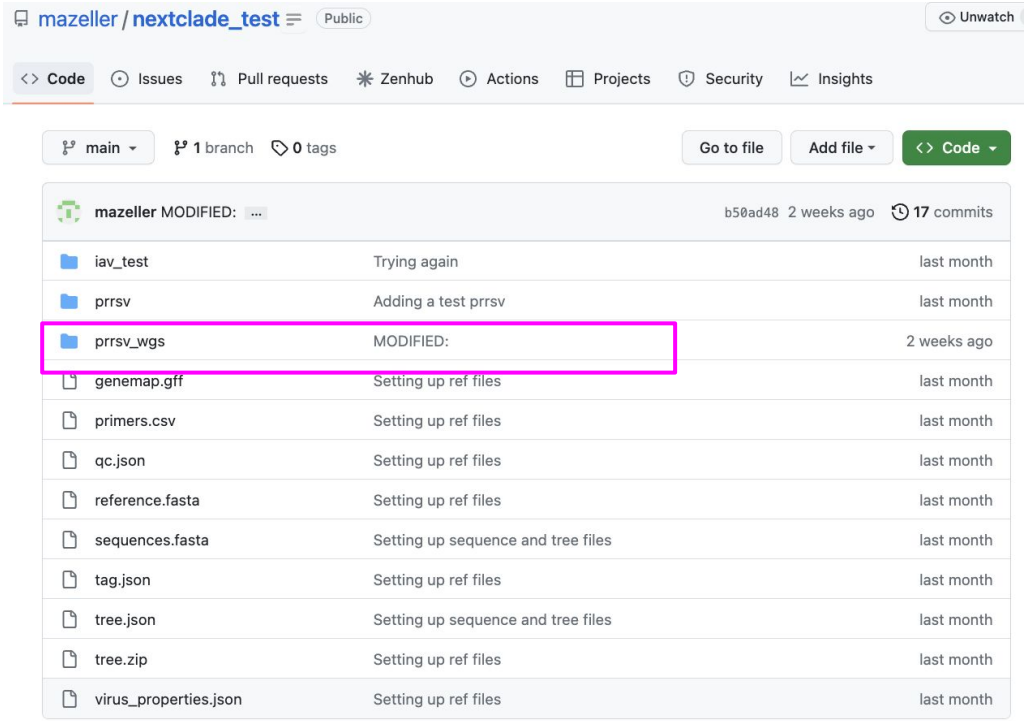
https://clades.nextstrain.org/?dataset-url=https://github.com/mazeller/nextclade_test/tree/main/prrsv

Validate Nextclade PRRS - WGS

Nextclade datasets

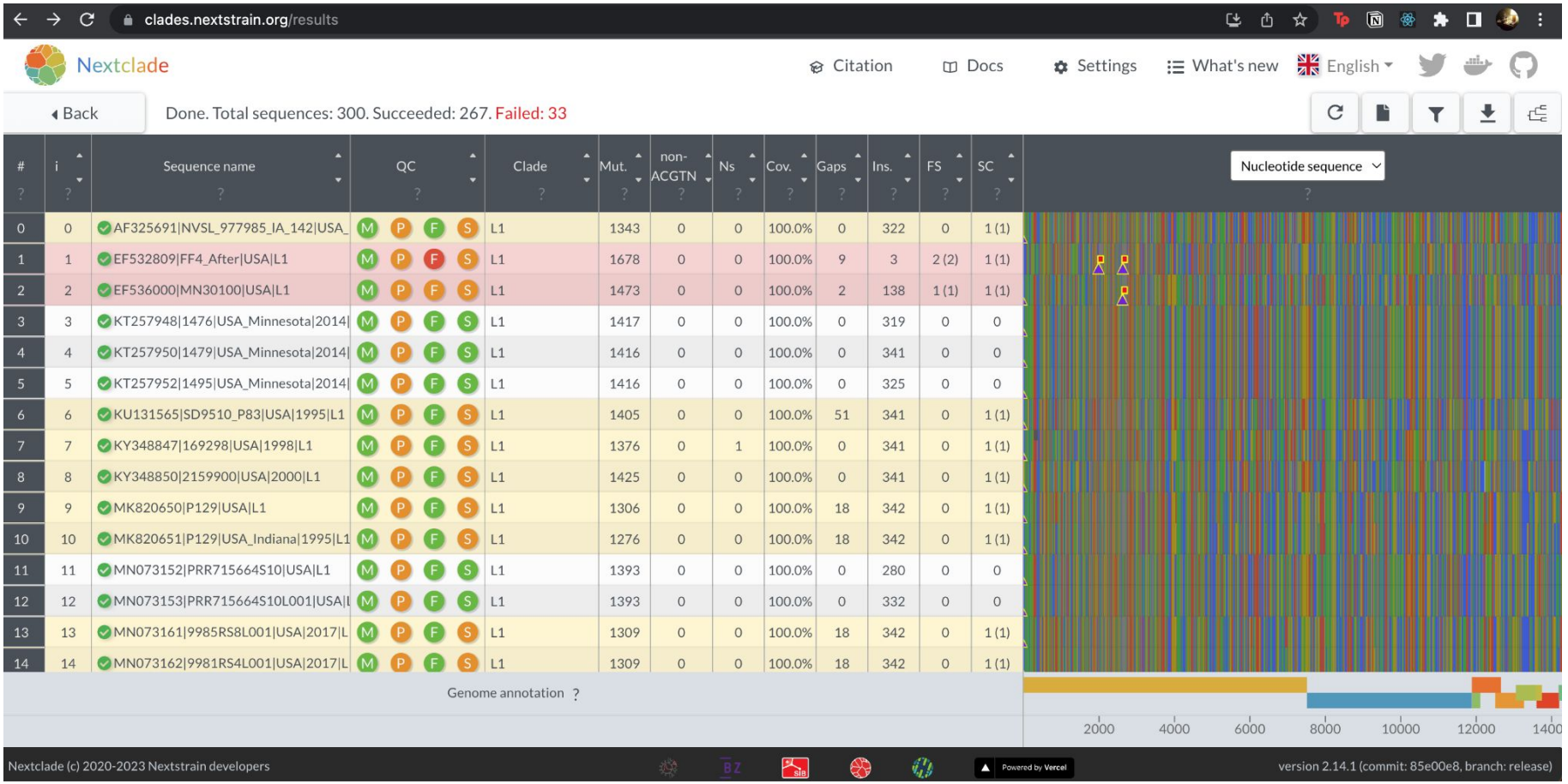
- tag.json
- qc.json
- primers.csv
- virus_properties.json
- reference.fasta
- genemap.gff
- sequences.fasta
- tree.json

github.com/mazeller/nextclade_test



https://clades.nextstrain.org/?dataset-url=https://github.com/mazeller/nextclade_test/tree/main/prrsv_wgs

Validate Nextclade PRRS - WGS



#	i	Sequence name	QC	Clade	Mut.	non-ACGTN	Ns	Cov.	Gaps	Ins.	FS	SC
0	0	AF325691|NVSL_977985_IA_142|USA_	M P F S	L1	1343	0	0	100.0%	0	322	0	1 (1)
1	1	EF532809|FF4_After|USA|L1	M P F S	L1	1678	0	0	100.0%	9	3	2 (2)	1 (1)
2	2	EF536000|MN30100|USA|L1	M P F S	L1	1473	0	0	100.0%	2	138	1 (1)	1 (1)
3	3	KT257948|1476|USA_Minnesota|2014|	M P F S	L1	1417	0	0	100.0%	0	319	0	0
4	4	KT257950|1479|USA_Minnesota|2014|	M P F S	L1	1416	0	0	100.0%	0	341	0	0
5	5	KT257952|1495|USA_Minnesota|2014|	M P F S	L1	1416	0	0	100.0%	0	325	0	0
6	6	KU131565|SD9510_P83|USA|1995|L1	M P F S	L1	1405	0	0	100.0%	51	341	0	1 (1)
7	7	KY348847|169298|USA|1998|L1	M P F S	L1	1376	0	1	100.0%	0	341	0	1 (1)
8	8	KY348850|2159900|USA|2000|L1	M P F S	L1	1425	0	0	100.0%	0	341	0	1 (1)
9	9	MK820650|P129|USA|L1	M P F S	L1	1306	0	0	100.0%	18	342	0	1 (1)
10	10	MK820651|P129|USA_Indiana|1995|L1	M P F S	L1	1276	0	0	100.0%	18	342	0	1 (1)
11	11	MN073152|PRR715664S10|USA|L1	M P F S	L1	1393	0	0	100.0%	0	280	0	0
12	12	MN073153|PRR715664S10L001|USA|L	M P F S	L1	1393	0	0	100.0%	0	332	0	0
13	13	MN073161|9985RS8L001|USA|2017|L	M P F S	L1	1309	0	0	100.0%	18	342	0	1 (1)
14	14	MN073162|9981RS4L001|USA|2017|L	M P F S	L1	1309	0	0	100.0%	18	342	0	1 (1)

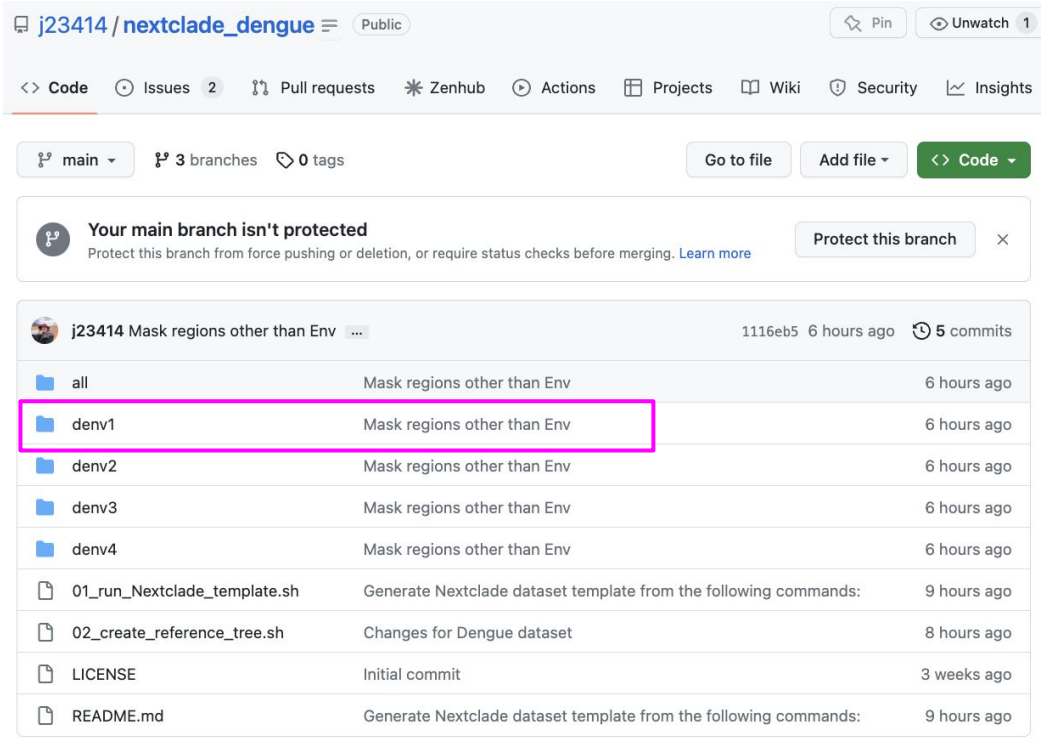
https://clades.nextstrain.org/?dataset-url=https://github.com/mazeller/nextclade_test/tree/main/prrsv_wgs

Validate Nextclade Dengue - denv1

Nextclade datasets

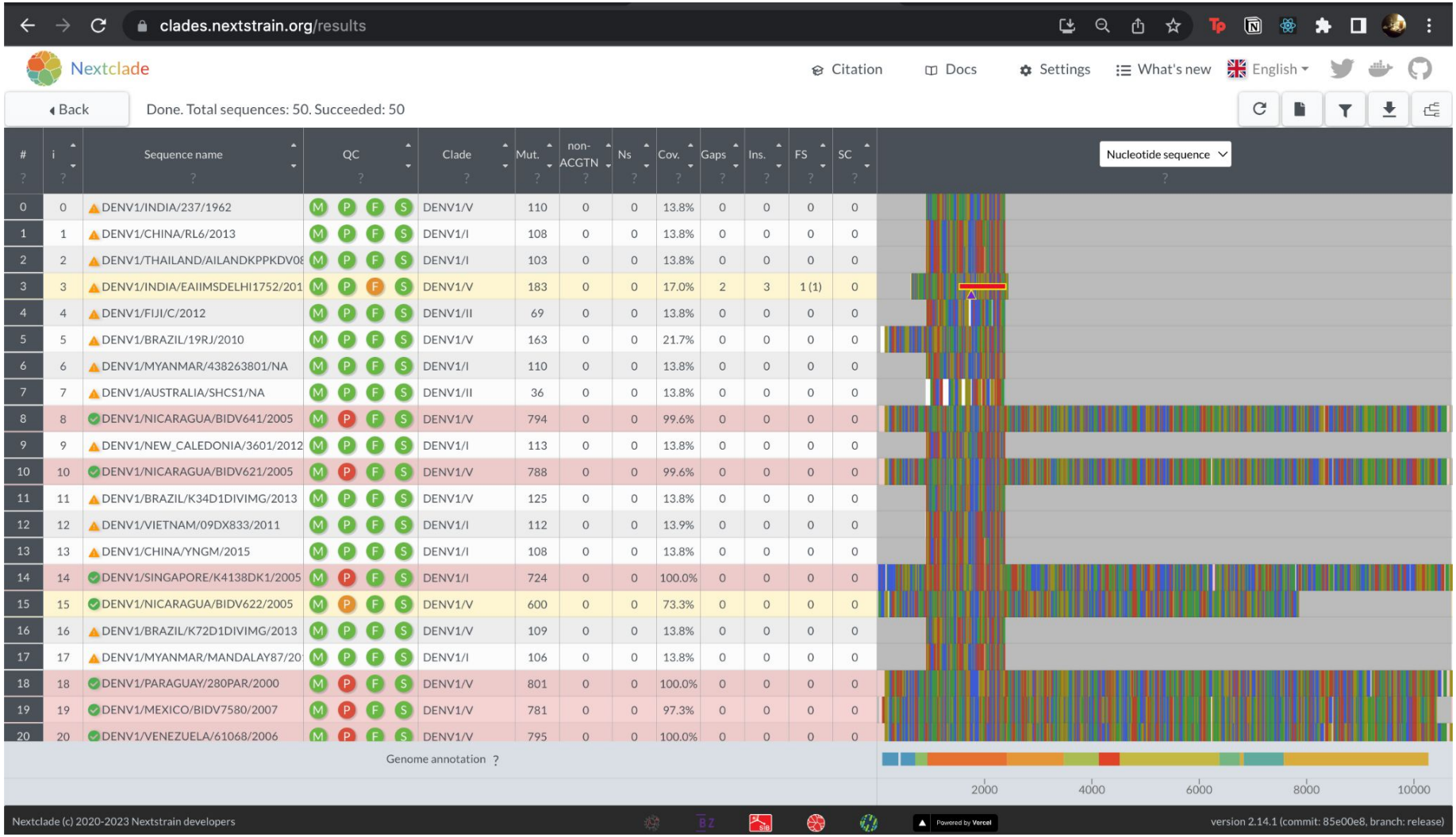
- tag.json
- qc.json
- primers.csv
- virus_properties.json
- reference.fasta
- genemap.gff
- sequences.fasta
- tree.json

github.com/j23414/nextclade_dengue



https://clades.nextstrain.org/?dataset-url=https://github.com/j23414/nextclade_dengue/tree/main/denv1

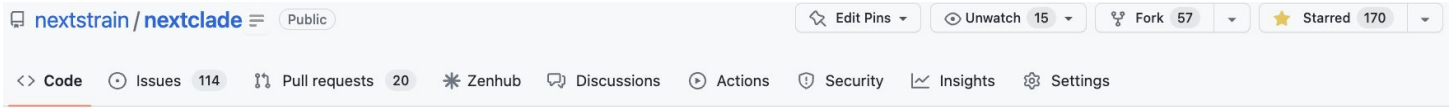
Validate Nextclade Dengue - denv1



#	i	Sequence name	QC	Clade	Mut.	non-ACGTN	Ns	Cov.	Gaps	Ins.	FS	SC
0	0	DENV1/INDIA/237/1962	M P F S	DENV1/V	110	0	0	13.8%	0	0	0	0
1	1	DENV1/CHINA/RL6/2013	M P F S	DENV1/I	108	0	0	13.8%	0	0	0	0
2	2	DENV1/THAILAND/AILANDKPPKDV0	M P F S	DENV1/I	103	0	0	13.8%	0	0	0	0
3	3	DENV1/INDIA/EAIIMSDELHI1752/201	M P F S	DENV1/V	183	0	0	17.0%	2	3	1 (1)	0
4	4	DENV1/FIJI/C/2012	M P F S	DENV1/II	69	0	0	13.8%	0	0	0	0
5	5	DENV1/BRAZIL/19RJ/2010	M P F S	DENV1/V	163	0	0	21.7%	0	0	0	0
6	6	DENV1/MYANMAR/438263801/NA	M P F S	DENV1/I	110	0	0	13.8%	0	0	0	0
7	7	DENV1/AUSTRALIA/SHCS1/NA	M P F S	DENV1/II	36	0	0	13.8%	0	0	0	0
8	8	DENV1/NICARAGUA/BIDV641/2005	M P F S	DENV1/V	794	0	0	99.6%	0	0	0	0
9	9	DENV1/NEW_CALEDONIA/3601/2012	M P F S	DENV1/I	113	0	0	13.8%	0	0	0	0
10	10	DENV1/NICARAGUA/BIDV621/2005	M P F S	DENV1/V	788	0	0	99.6%	0	0	0	0
11	11	DENV1/BRAZIL/K34D1DIVIMG/2013	M P F S	DENV1/V	125	0	0	13.8%	0	0	0	0
12	12	DENV1/VIETNAM/09DX833/2011	M P F S	DENV1/I	112	0	0	13.9%	0	0	0	0
13	13	DENV1/CHINA/YNGM/2015	M P F S	DENV1/I	108	0	0	13.8%	0	0	0	0
14	14	DENV1/SINGAPORE/K4138DK1/2005	M P F S	DENV1/I	724	0	0	100.0%	0	0	0	0
15	15	DENV1/NICARAGUA/BIDV622/2005	M P F S	DENV1/V	600	0	0	73.3%	0	0	0	0
16	16	DENV1/BRAZIL/K72D1DIVIMG/2013	M P F S	DENV1/V	109	0	0	13.8%	0	0	0	0
17	17	DENV1/MYANMAR/MANDALAY87/20	M P F S	DENV1/I	106	0	0	13.8%	0	0	0	0
18	18	DENV1/PARAGUAY/280PAR/2000	M P F S	DENV1/V	801	0	0	100.0%	0	0	0	0
19	19	DENV1/MEXICO/BIDV7580/2007	M P F S	DENV1/V	781	0	0	97.3%	0	0	0	0
20	20	DENV1/VENEZUELA/61068/2006	M P F S	DENV1/V	795	0	0	100.0%	0	0	0	0

placementMaskRanges

https://github.com/nextstrain/nextclade/releases/tag/2.14.0



Releases / 2.14.0

2.14.0 Latest

Compare

nextstrain-bot released this last week 2.14.0 f4cb934

Nextclade Web 2.14.0, Nextclade CLI 2.14.0 (2023-05-09)

Algorithm & Datasets: enable masked sites for distance calculation

For some viruses, genome sequencing is unreliable in specific parts of the genome or some regions should be ignored for other reasons when calculating distances between nodes for the purpose of placing query sequences on the reference tree. These distances are used to find the optimal (smallest distance) placement of the query sequence on the reference tree and sequence errors in these regions can lead to wrong placement.

Until now, to place query sequences on the reference tree, Nextclade counted all nucleotide differences between query and reference sequence. Moving forward, sequence regions to be ignored for reference tree placement can be defined in datasets' `virus_properties.json`. This is useful for example for SARS-CoV-2, where we will start ignoring the terminal parts of the untranslated regions. Another use case is mpox, where the terminal repeats are intrinsically constrained to be identical. Masking one of the two terminals will avoid double-counting of the same mutations.

PR #1128 adds this feature to Nextclade's algorithm.

Masked ranges are specified in the new field `placementMaskRanges` in datasets' `virus_properties.json`. For example, the terminal 50 nucleotides of SARS-CoV-2 can be ignored for tree placement by adding the following line (positions are 0-based and end-exclusive):

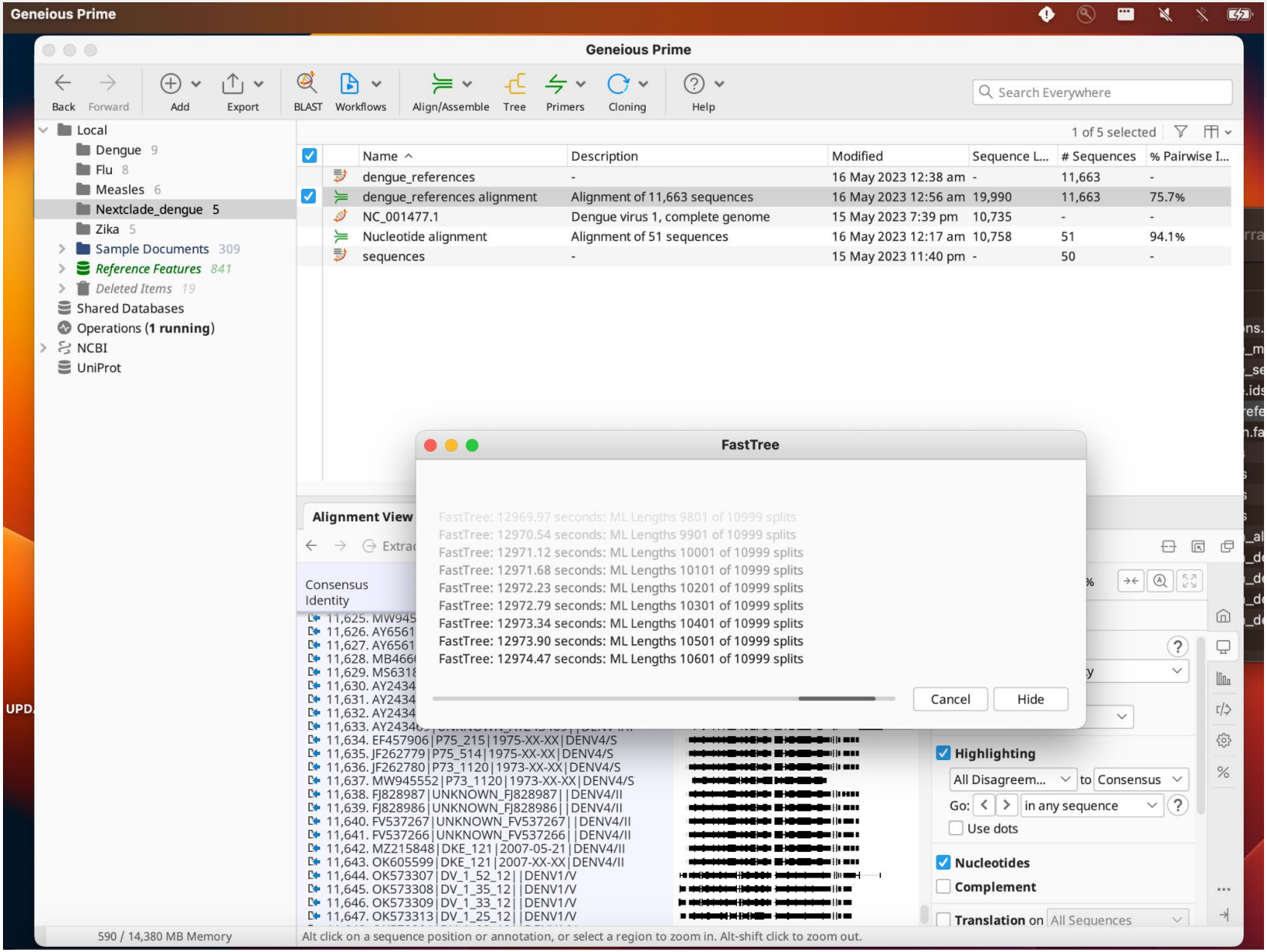
```
"placementMaskRanges":[{"begin":0,"end":50},{"begin":29850,"end":29902}],
```

The changes are backwards compatible, if the field does not exist, Nextclade defaults to the old behavior of counting all nucleotide differences.

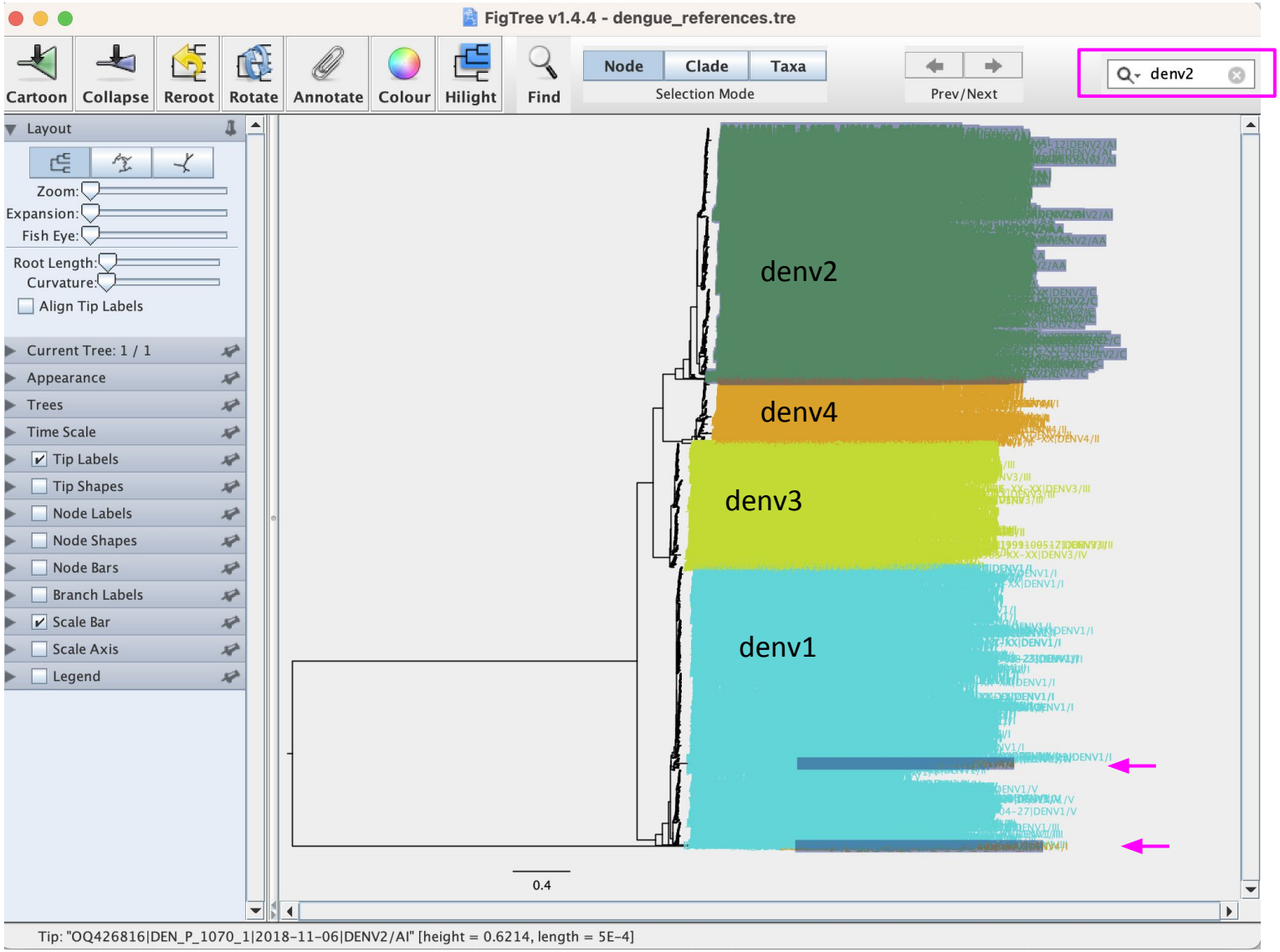
We are planning to shortly release a new version of SARS-CoV-2 datasets making use of this feature. Only a small proportion of sequences (<1%)should be affected, however where there are changes they will be a slight improvement in accuracy.

Try using WGS

```
# Filter by length > 9K nt

augur index \
  --sequences sequences_all.fasta \
  --output index.txt

tsv-filter -H --gt length:9000 a.txt \
  | awk 'NR > 1 {print $1}' > complete.ids


tsv-join -H \
  --filter-file complete.ids \
  --key-fields strain \
  --data-fields accession \
  --allow-duplicate-keys \
  --a 1 \
  metadata_all.tsv \
  > complete_metadata.tsv

smof grep -f complete.ids sequences_all.fasta \
  > complete_sequences.fasta

cat complete_metadata.tsv \
| tsv-select -H -f
accession,strain,date,serotype,subclade \
| tsv-filter -H --not-empty subclade \
>
```

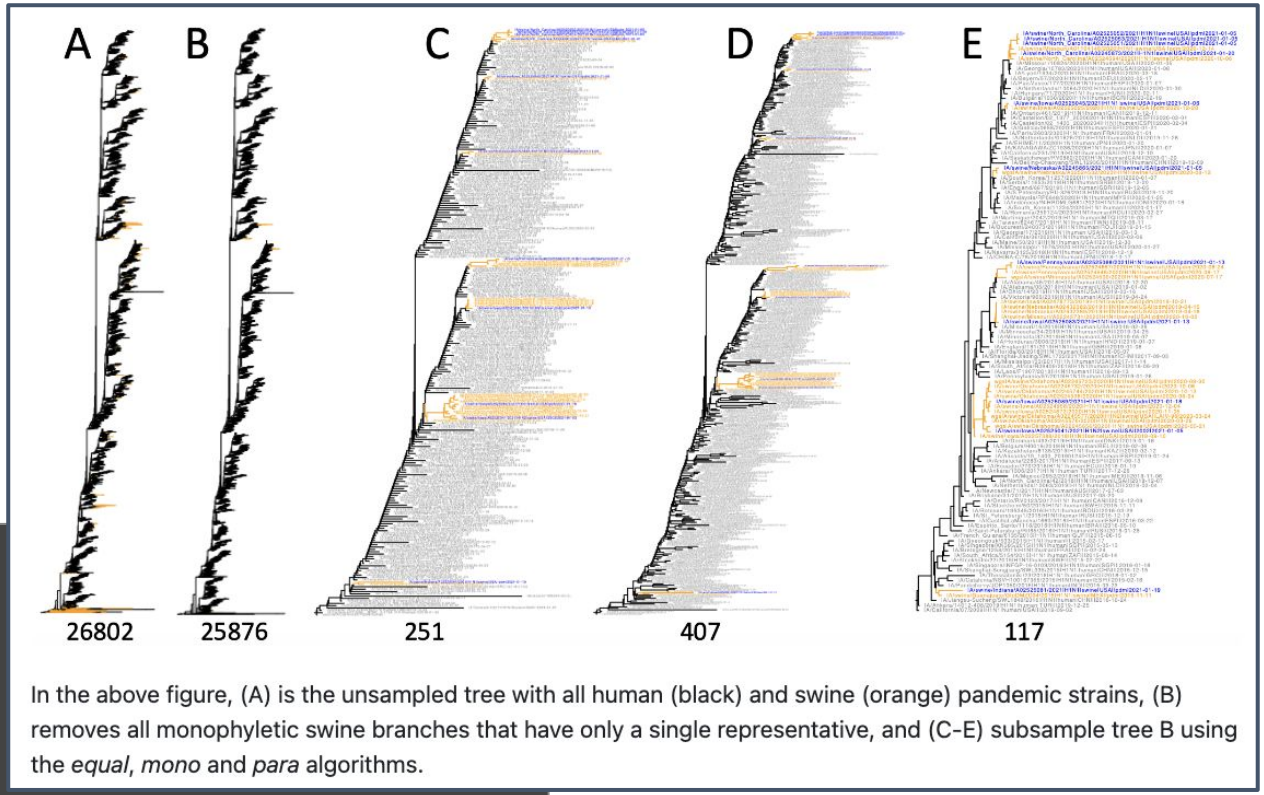

Geneious Prime
Geneious Prime
Back
Forward
Add
Export
BLAST
Workflows
Align/Assemble
Tree
Primers
Cloning
Help
Search Everywhere
Local
Dengue 9
Flu 8
Measles 6
Nextclade_dengue 5
Zika 5
Sample Documents 309
Reference Features 841
Deleted Items 19
Shared Databases
Operations (1 running)
NCBI
UniProt
1 of 5 selected
Name
Description
Modified
Sequence L...
Sequences
% Pairwise I...
dengue_references - 16 May 2023 12:38 am - 11,663 -
dengue_references alignment Alignment of 11,663 sequences 16 May 2023 12:56 am 19,990 11,663 75.7%
NC_001477.1 Dengue virus 1, complete genome 15 May 2023 7:39 pm 10,735 - -
Nucleotide alignment Alignment of 51 sequences 16 May 2023 12:17 am 10,758 51 94.1%
sequences - 15 May 2023 11:40 pm - 50 -
FastTree
FastTree: 12969.97 seconds: ML Lengths 9801 of 10999 splits
FastTree: 12970.54 seconds: ML Lengths 9901 of 10999 splits
FastTree: 12971.12 seconds: ML Lengths 10001 of 10999 splits
FastTree: 12971.68 seconds: ML Lengths 10101 of 10999 splits
FastTree: 12972.23 seconds: ML Lengths 10201 of 10999 splits
FastTree: 12972.79 seconds: ML Lengths 10301 of 10999 splits
FastTree: 12973.34 seconds: ML Lengths 10401 of 10999 splits
FastTree: 12973.90 seconds: ML Lengths 10501 of 10999 splits
FastTree: 12974.47 seconds: ML Lengths 10601 of 10999 splits
Cancel
Hide
Alignment View
Consensus
Identity
11,634. EF457906|P75_215|1975-XX-XX|DENV4/S
11,635. JF262779|P75_514|1975-XX-XX|DENV4/S
11,636. JF262780|P73_1120|1973-XX-XX|DENV4/S
11,637. MW945552|P73_1120|1973-XX-XX|DENV4/S
11,638. FJ828987|UNKNOWN_FJ828987||DENV4/II
11,639. FJ828986|UNKNOWN_FJ828986||DENV4/II
11,640. FV537267|UNKNOWN_FV537267||DENV4/II
11,641. FV537266|UNKNOWN_FV537266||DENV4/II
11,642. MZ215848|DKE_121|2007-05-21|DENV4/II
11,643. OK605599|DKE_121|2007-XX-XX|DENV4/II
11,644. OK573307|DV_1_52_12||DENV1/V
11,645. OK573308|DV_1_35_12||DENV1/V
11,646. OK573309|DV_1_33_12||DENV1/V
11,647. OK573313|DV_1_25_12||DENV1/V
Highlighting
All Disagreem... to Consensus
Go: in any sequence
Use dots
Nucleotides
Complement
Translation on All Sequences
590 / 14,380 MB Memory
Alt click on a sequence position or annotation, or select a region to zoom in. Alt-shift click to zoom out.

dengue_reference.tre



denv2?

dengue_reference.tre - smot



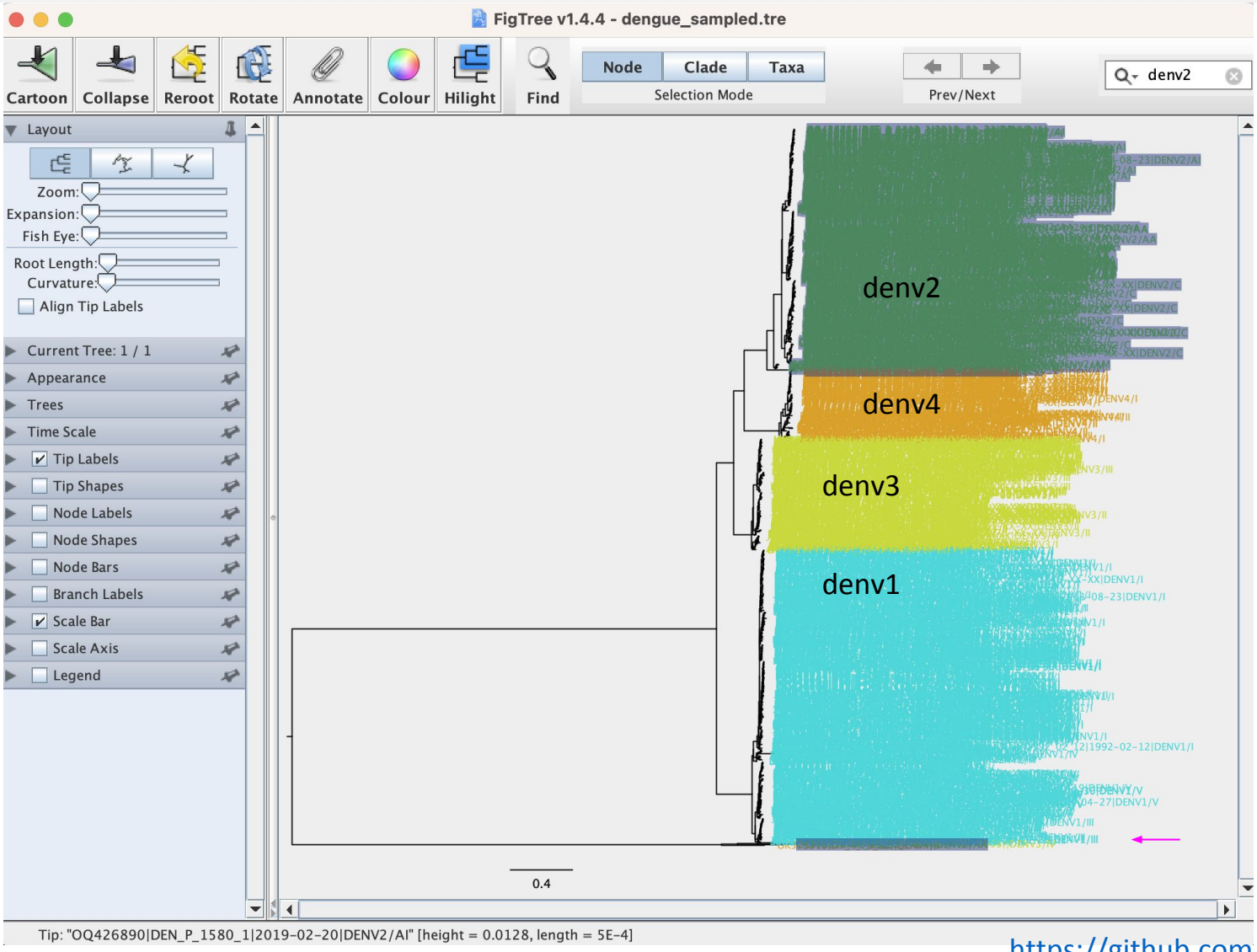
In the above figure, (A) is the unsampled tree with all human (black) and swine (orange) pandemic strains, (B) removes all monophyletic swine branches that have only a single representative, and (C-E) subsample tree B using the *equal*, *mono* and *para* algorithms.

https://github.com/flu-crew/smot

```
smot sample para \
--proportion=0.1 \
dengue_references.tre \
--factor-by-field=4 \
--newick \
> dengue_sampled.tre

smot sample para \
--proportion=0.1 \
dengue_references.tre \
--factor-by-capture="(denv1|denv2|denv3|denv4)" \
--newick \
> dengue_sampled.tre
```

Arendsee et al., JOSS 2022

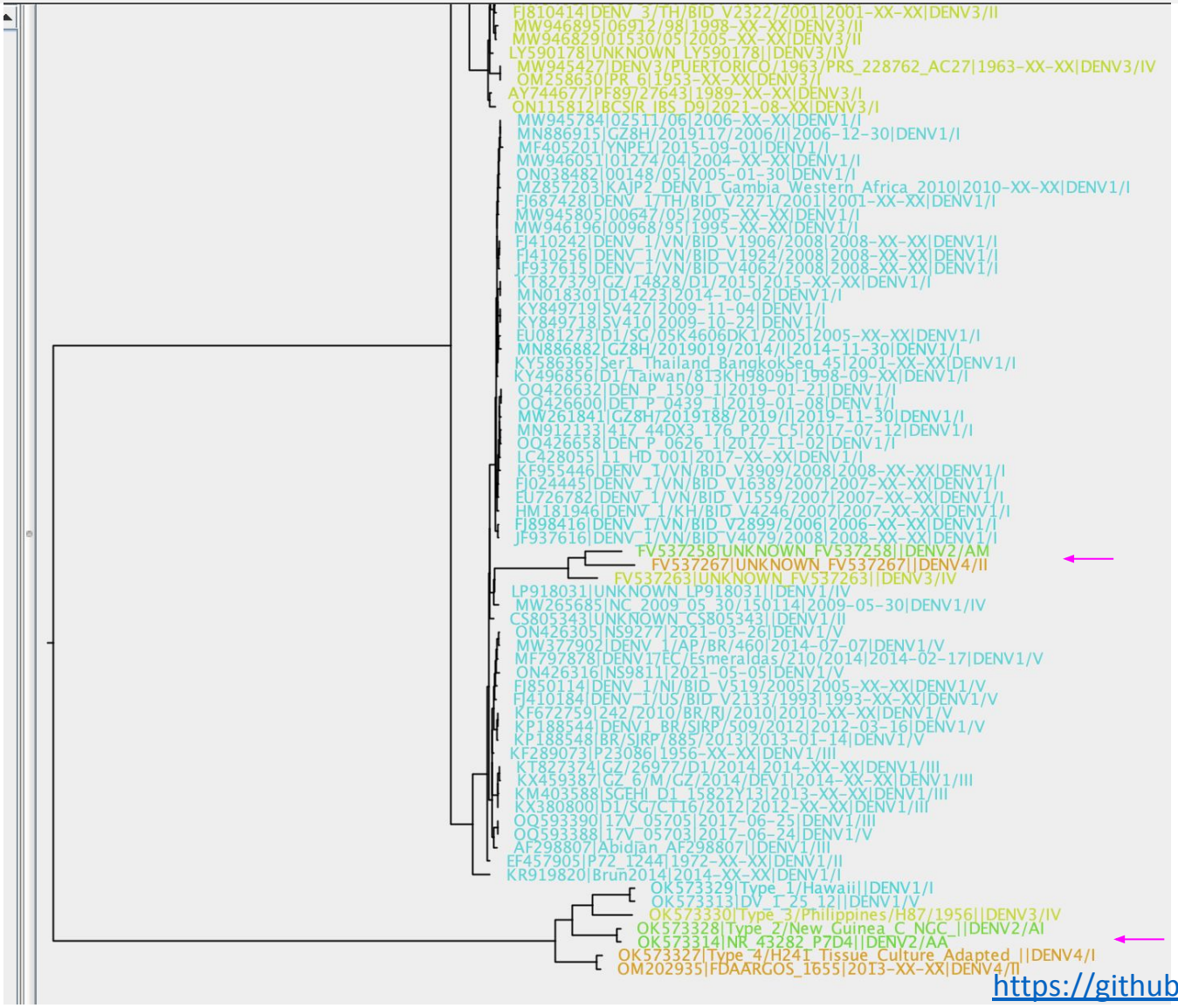
dengue_sampled.tre - smot



less denv2

https://github.com/flu-crew/smot

dengue_sampled.tre - smot



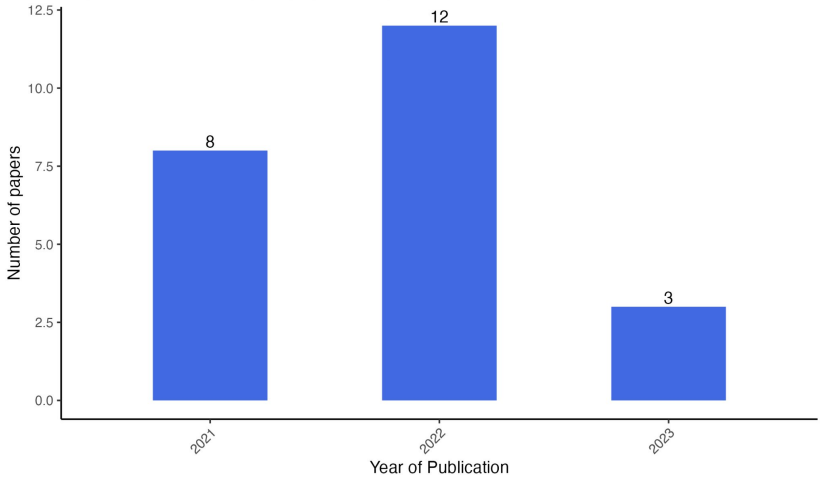
https://github.com/flu-crew/smot

Next steps

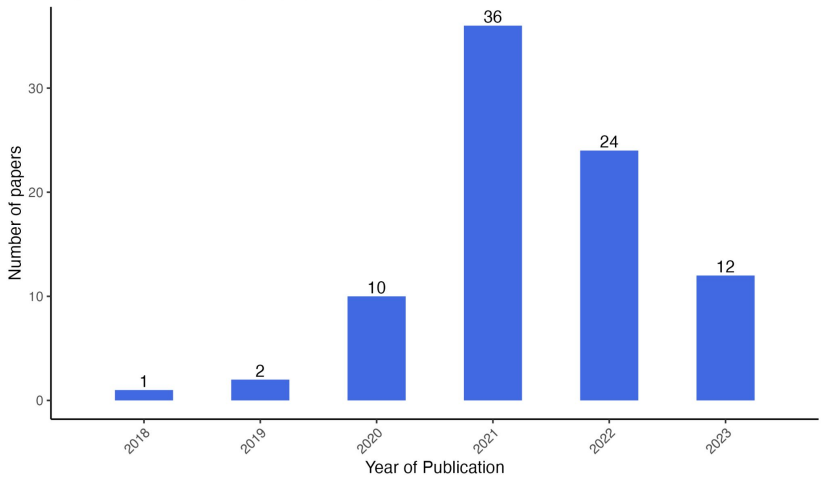
- Investigate and fix or filter out-of-place strains
- Rebuild the Nextclade dengue dataset with sampled tree
- Determine if WGS or single gene origin dataset is better
- Add Nextclade rules back to Dengue Ingest

Trends

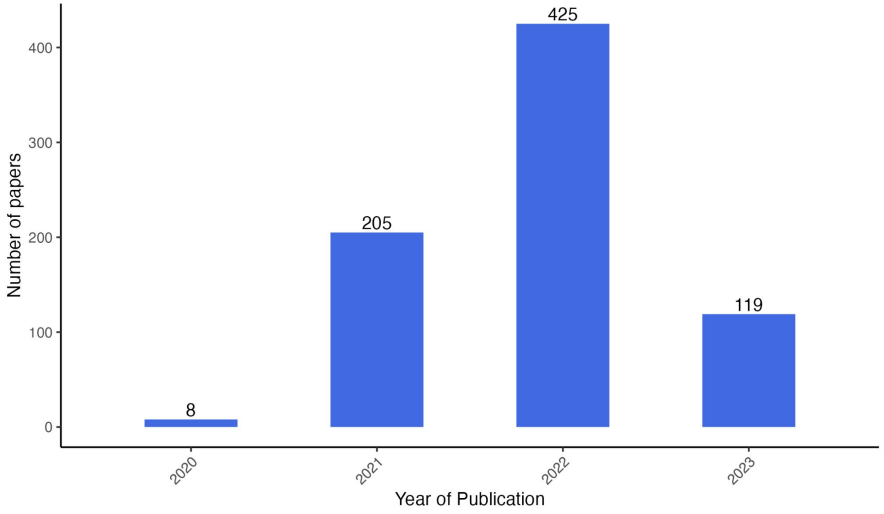
Nextclade - pubmed check (23)



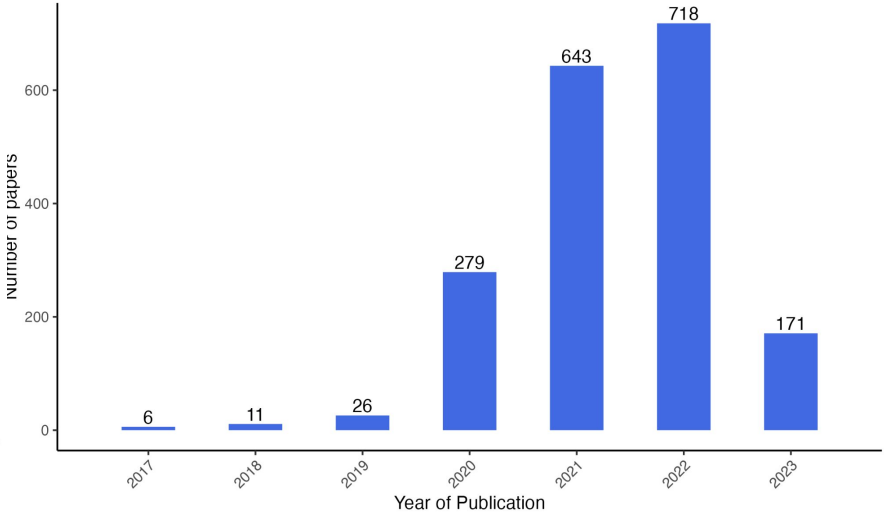
Nextstrain - pubmed_check (85)



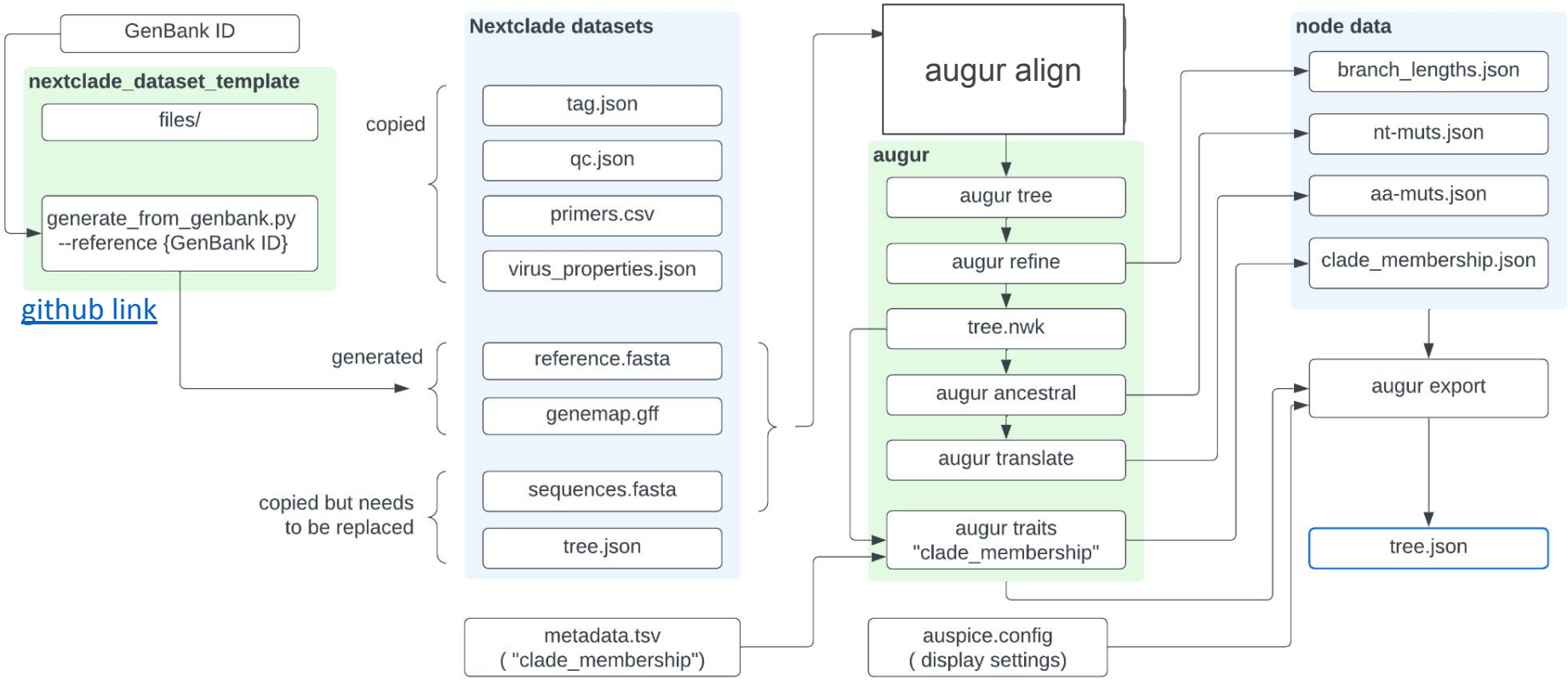
Nextclade - PMC_check (757)



Nextstrain - PMC_check (1854)



In Summary



validation: https://clades.nextstrain.org/?dataset-url=https://github.com/user/nextclade_pathgen/tree/main

https://github.com/j23414/nextclade_dengue

https://github.com/neherlab/nextclade_data_workflows

References

- Aksamentov**, I.**, Roemer, C., Hodcroft, E.B. and Neher, R.A., 2021. Nextclade: clade assignment, mutation calling and quality control for viral genomes. Journal of open source software, 6(67), p.3773.
- Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T. and Neher, R.A., 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics, 34(23), pp.4121-4123.
- Huddleston, J., Hadfield, J., Sibley, T.R., Lee, J., Fay, K., Ilcisin, M., Harkins, E., Bedford, T., Neher, R.A. and Hodcroft, E.B., 2021. Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. Journal of open source software, 6(57).
- Sayers, E.W., Cavanaugh, M., Clark, K., Pruitt, K.D., Schoch, C.L., Sherry, S.T. and Karsch-Mizrachi, I., 2022. GenBank. Nucleic acids research, 50(D1), p.D161.
- Formal specifications of GFF3 format from Sequence Ontology
- Reese, M.G., Moore, B., Batchelor, C., Salas, F., Cunningham, F., Marth, G.T., Stein, L., Flicek, P., Yandell, M. and Eilbeck, K., 2010. A standard variation file format for human genome sequences. Genome biology, 11, pp.1-9.
- Arendsee, Z.W., Baker, A.L.V. and Anderson, T.K., 2022. smot: a python package and CLI tool for contextual phylogenetic subsampling. Journal of Open Source Software, 7(80), p.4193.