Creating a Nextclade dataset

- Bedford Lab Meeting -

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

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

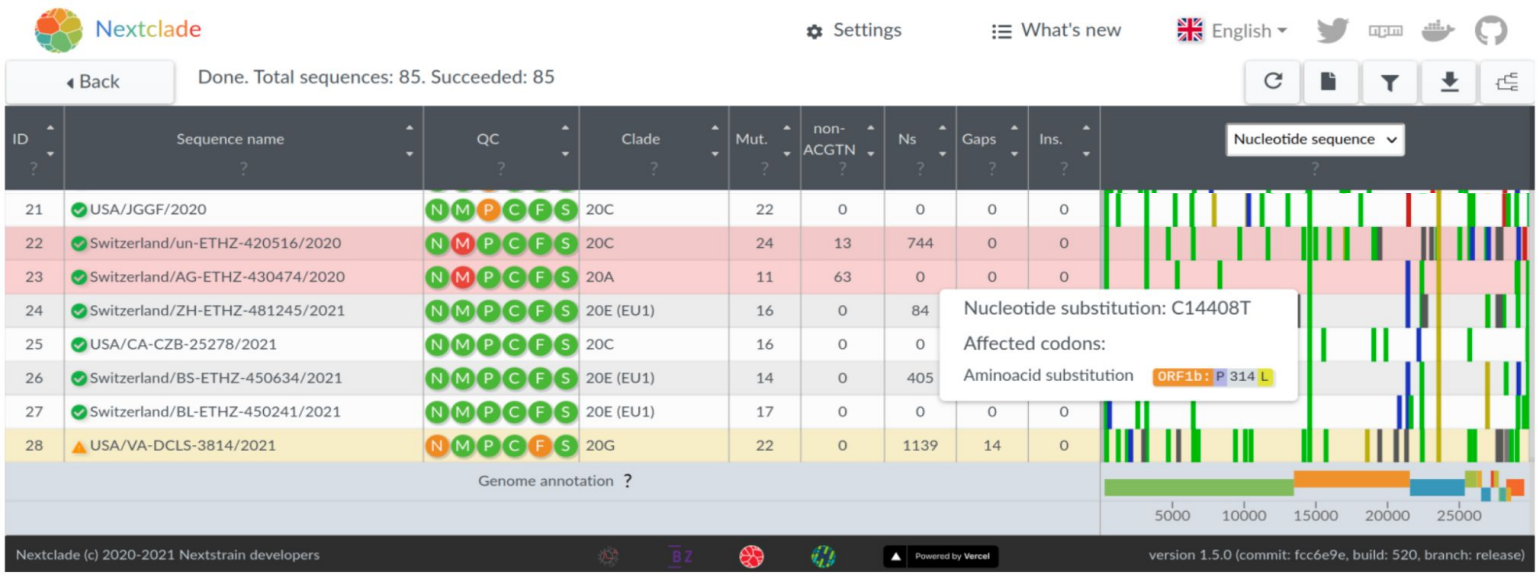


Figure 1: Overview of the results page with clade assignments, QC metrics, and the nucleotide mutation view. The results can be explored interactively and exported in standard tabular file formats.

Aksamentov et al., JOSS 2021

What is a Nextclade dataset?

Nextclade datasets

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

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?

Nextclade datasets

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

nextstrain / nextclade_data Public

Edit Pins | Watch 3 | Fork 14 | Star 19

Code | Issues 5 | Pull requests 2 | Zenhub | Discussions | Actions | Projects | Security | Insights | Settings

Code

master

Go to file t

- .github
- data
  - datasets
    - MPXV
    - flu_h1n1pdm_ha
      - references
        - CY121680
        - MW626062
          - versions
            - 2022-12-07T08:35:53Z/fil...
              - genemap.gff
              - primers.csv
              - qc.json
              - reference.fasta
              - sequences.fasta
              - tag.json
              - tree.json
              - virus_properties.json
            - 2023-01-19T12:00:00Z/fil...
            - 2023-01-27T12:00:00Z/fil...
            - 2023-04-02T12:00:00Z/fi...
          - datasetRef.json
          - dataset.json

github.com/nextstrain/nextclade_data/tree/ma...

nextclade_data / data / datasets /

Add file

corneliusroemer Add placementMaskRanges and new labeled muts (23B) bcdfeca · 4 days ago History

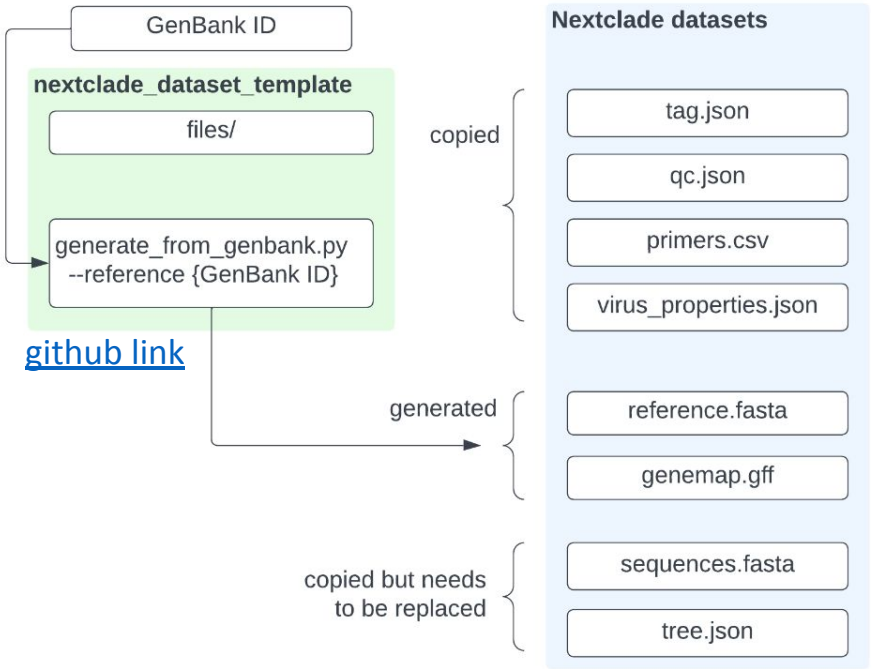
Name	Last commit message	Last commit date
..		
MPXV	Monkeypox dataset update 2023-01-26	4 months ago
flu_h1n1pdm_ha	update flu datasets and fix B/vic annotation	last month
flu_h1n1pdm_na	update flu datasets and fix B/vic annotation	last month
flu_h3n2_ha	update flu datasets and fix B/vic annotation	last month
flu_h3n2_na	update flu datasets and fix B/vic annotation	last month
flu_vic_ha	update flu datasets and fix B/vic annotation	last month
flu_vic_na	update flu datasets and fix B/vic annotation	last month
flu_yam_ha	fix: date in tag wrong month	10 months ago
hMPXV	Monkeypox dataset update 2023-01-26	4 months ago
hMPXV_B1	Monkeypox dataset update 2023-01-26	4 months ago
rsv_a	rsv: update data sets	3 months ago
rsv_b	rsv: update data sets	3 months ago
sars-cov-2-21L	Add placementMaskRanges and new labeled muts (23B)	4 days ago
sars-cov-2-no-recomb	sc2: Update datasets, now containing placement_prior	2 months ago
sars-cov-2	Add placementMaskRanges and new labeled muts (23B)	4 days ago

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



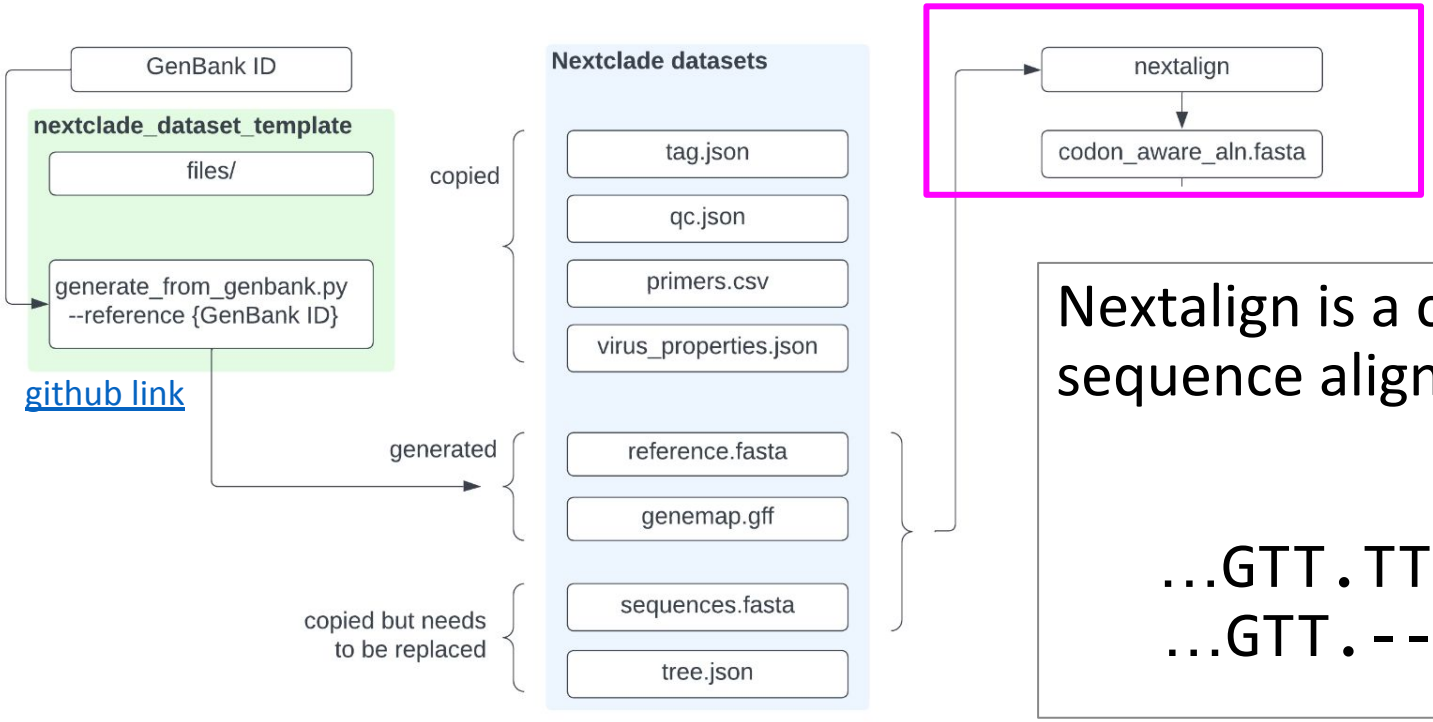
Pick a reference that is close to the base of the tree.

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 | ...
```

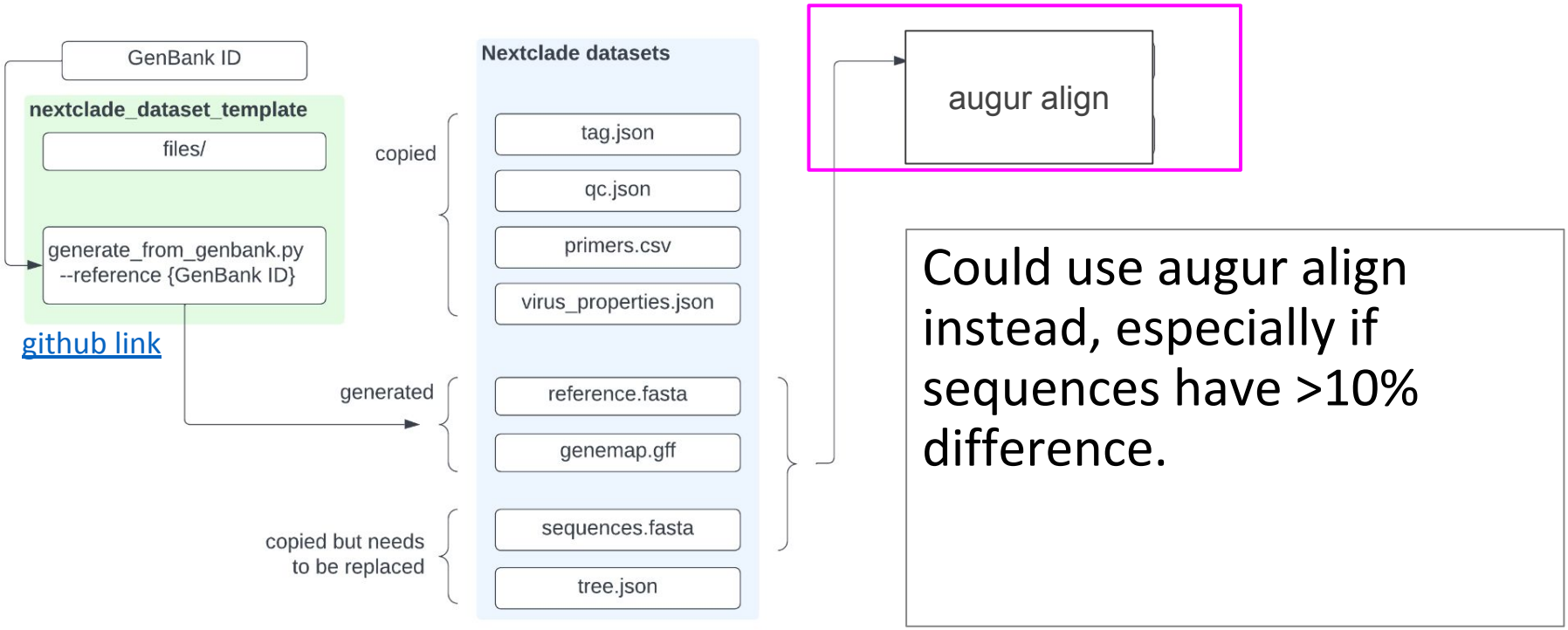
If a genome annotation is provided, Nextclade will use a lower gap-open-penalty at the beginning of a codon (delimited by the `|` characters in the schema above), thereby locking a gap in-frame if possible. Similarly, nextalign preferentially places gaps outside of genes in case of ambiguities.

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

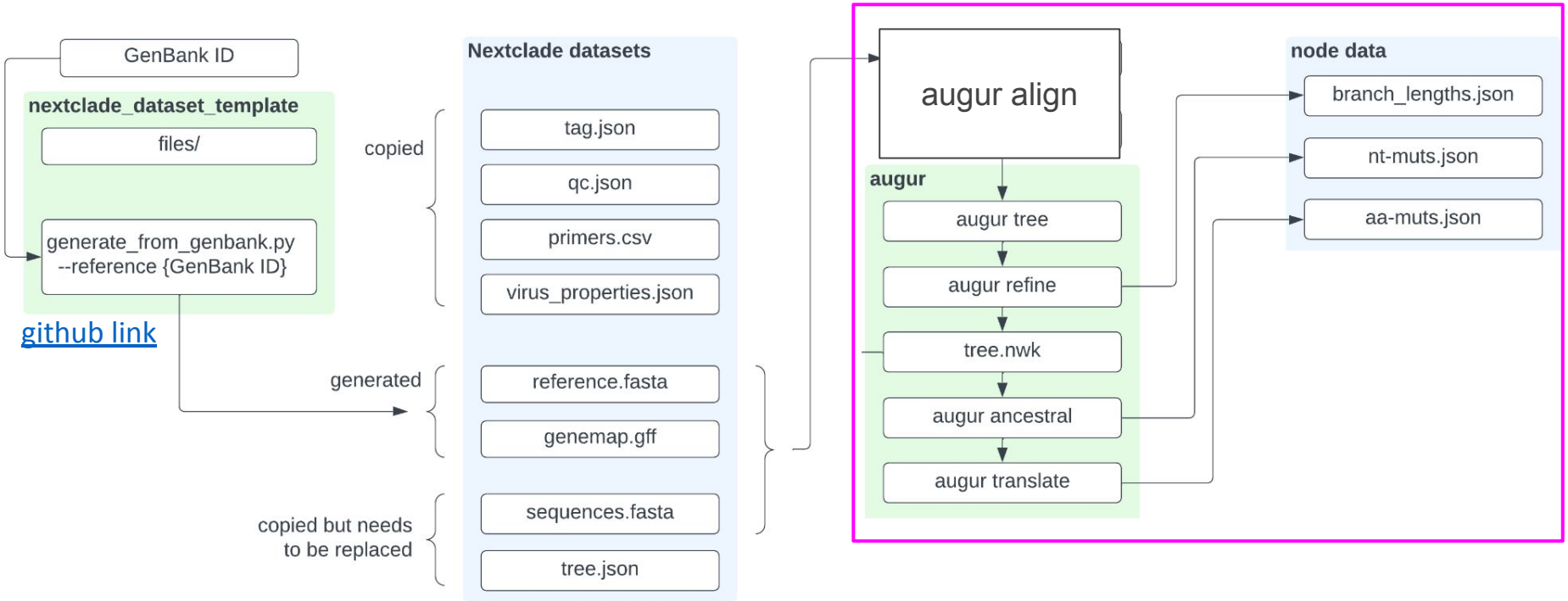
Nextalign or augur align



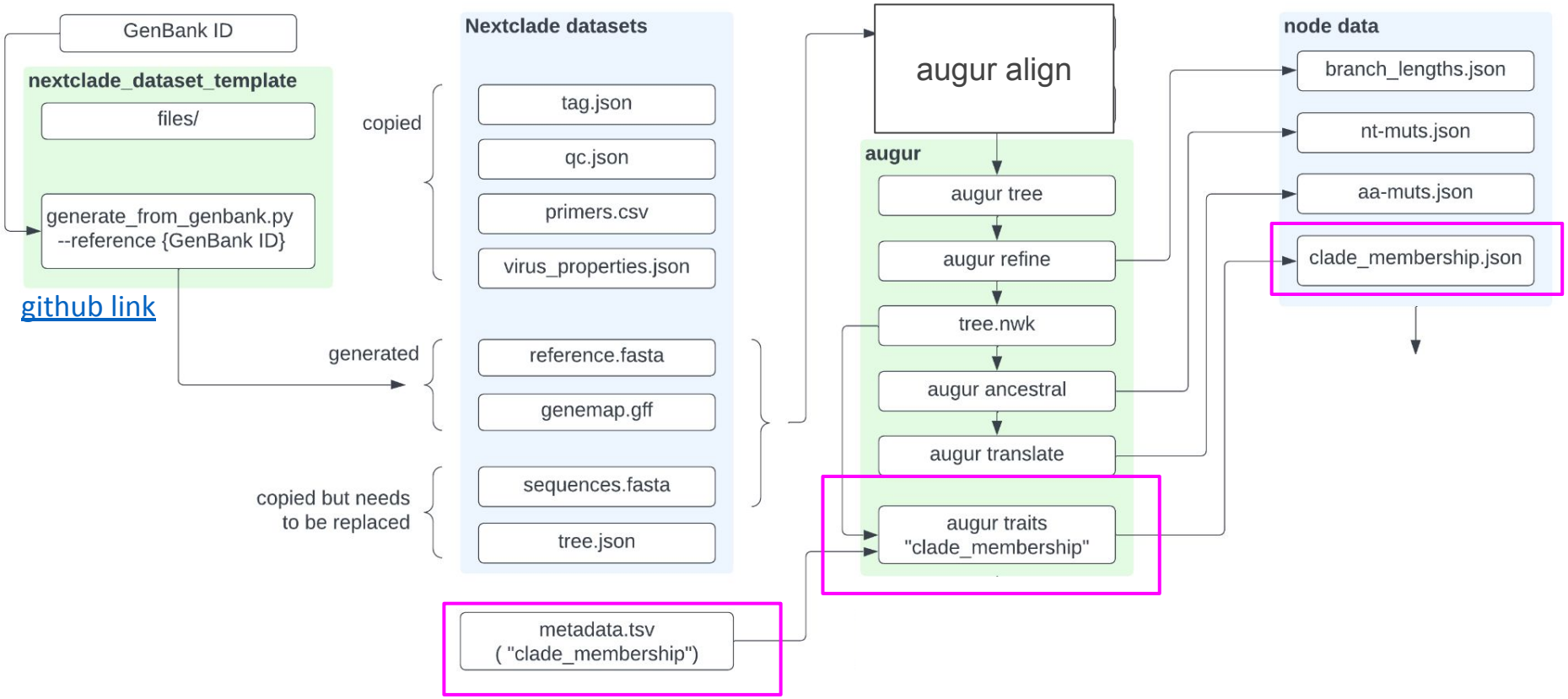
Could use augur align instead, especially if sequences have >10% difference.

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

Standard augur commands

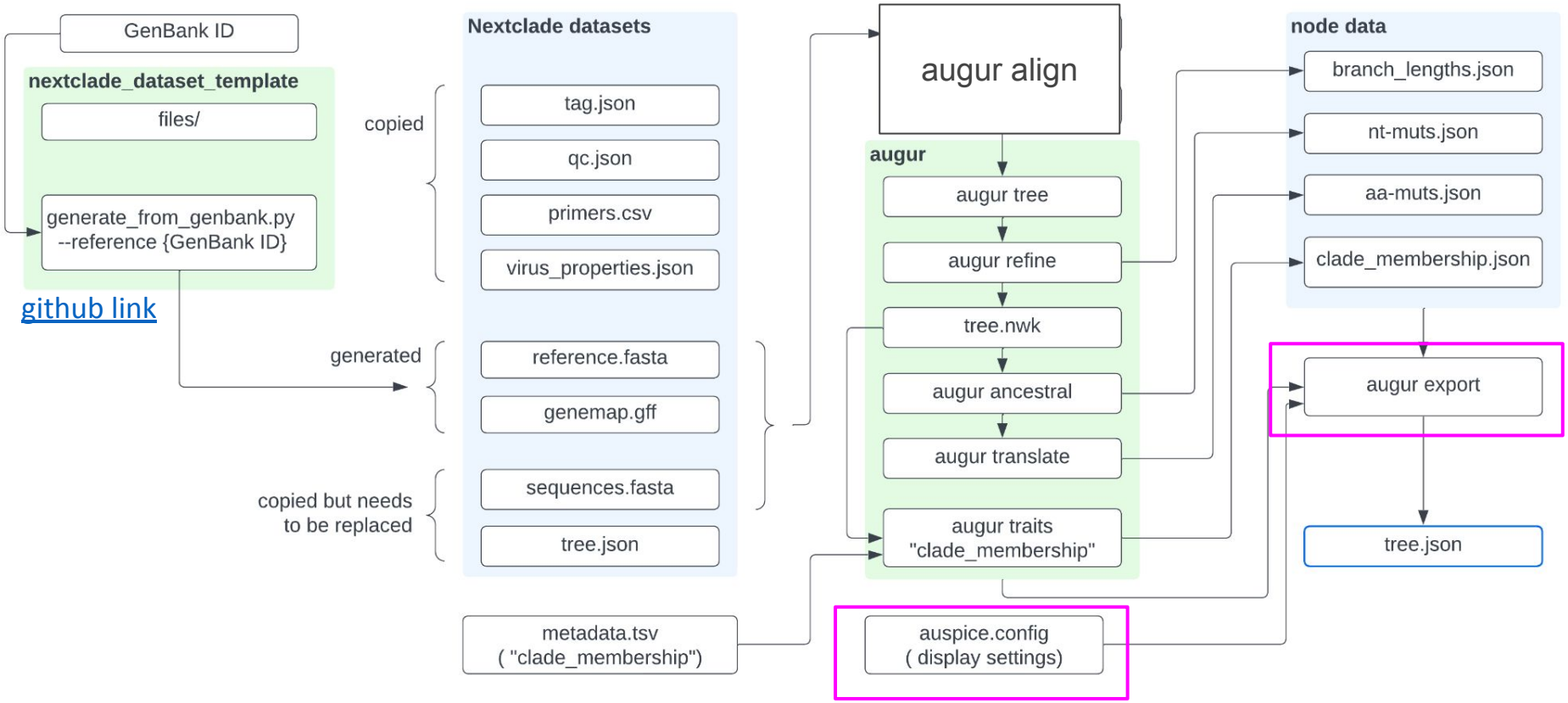


Clade membership

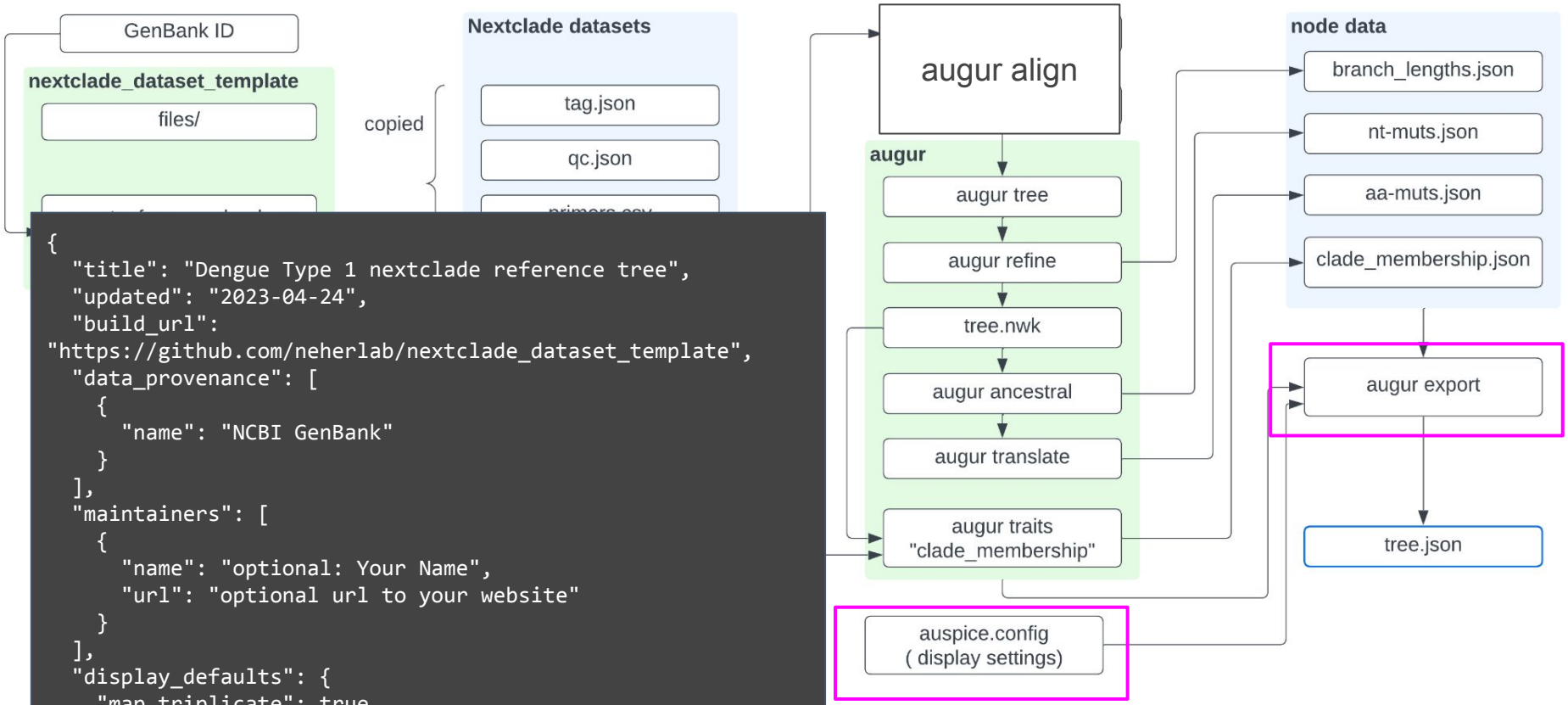


metadata.tsv should have "strain" and "clade_membership" columns

Augur export

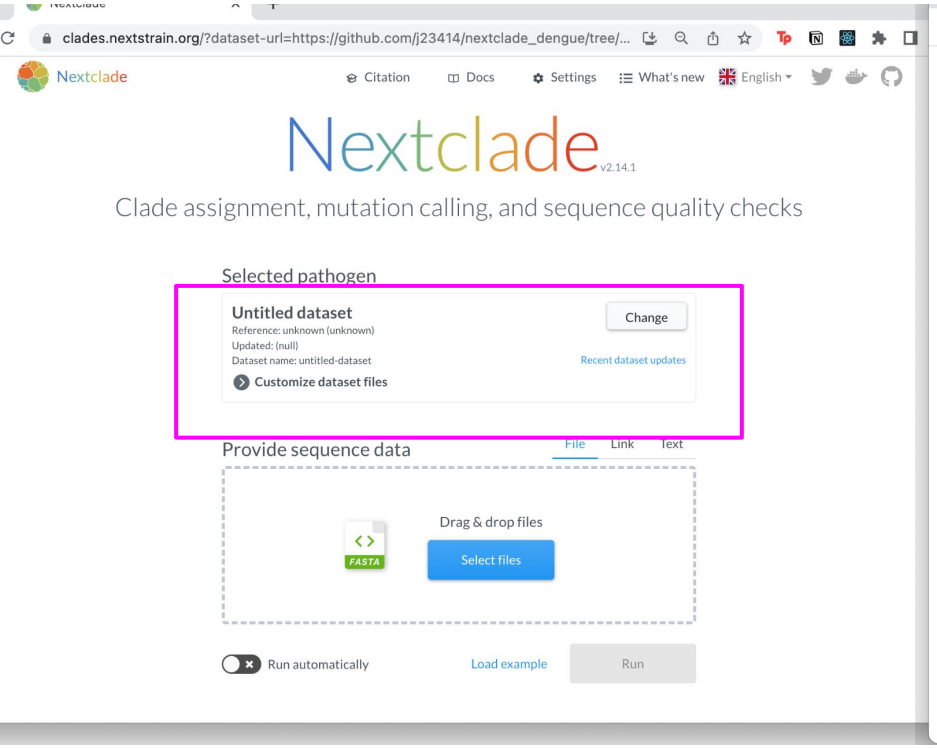


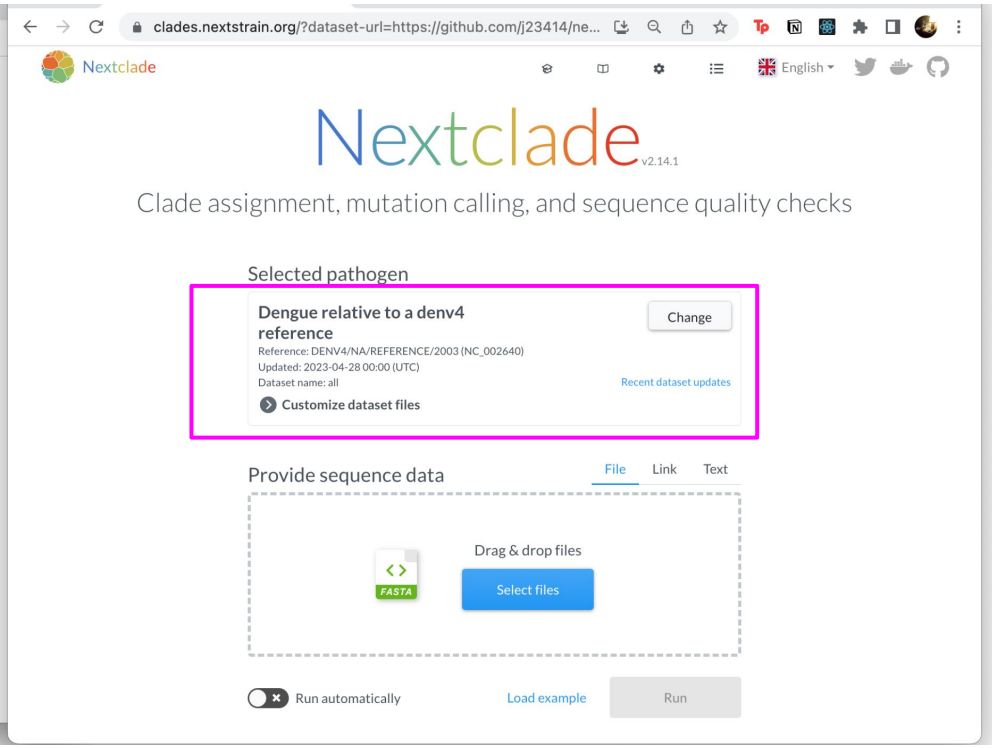
Augur export




```
{
  "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

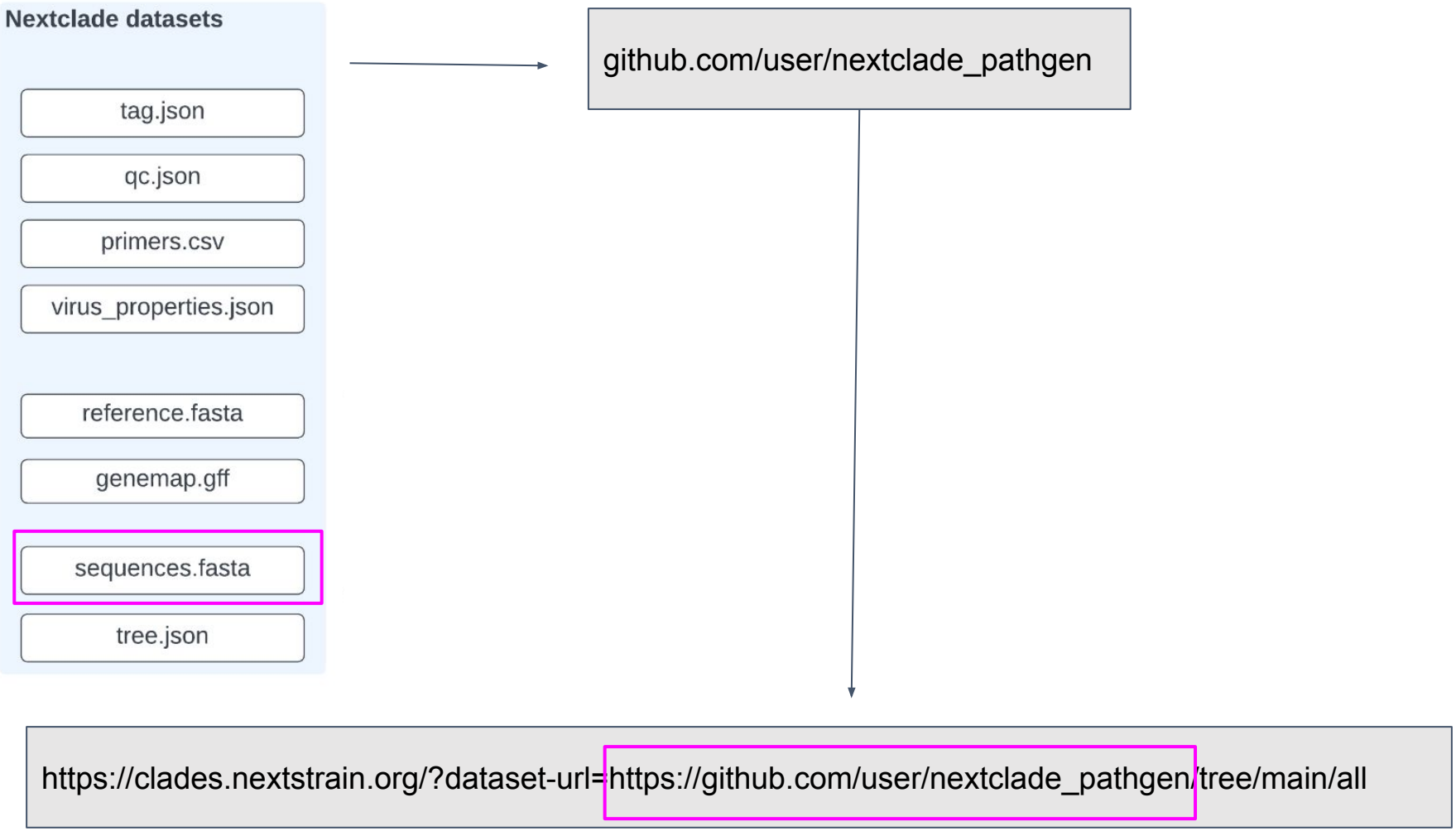




Agenda

- 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 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

j23414 / nextclade_dengue Public

Pin | Unwatch 1

Code | Issues 2 | Pull requests | Zenhub | Actions | Projects | Wiki | Security | Insights

main | 3 branches | 0 tags | Go to file | Add file | Code

Your main branch isn't protected
Protect this branch from force pushing or deletion, or require status checks before merging. Learn more

Protect this branch

j23414 Mask regions other than Env ... 1116eb5 6 hours ago 5 commits

all	Mask regions other than Env	6 hours ago
denv1	Mask regions other than Env	6 hours ago
denv2	Mask regions other than Env	6 hours ago
denv3	Mask regions other than Env	6 hours ago
denv4	Mask regions other than Env	6 hours ago
01_run_Nextclade_template.sh	Generate Nextclade dataset template from the following commands:	9 hours ago
02_create_reference_tree.sh	Changes for Dengue dataset	8 hours ago
LICENSE	Initial commit	3 weeks ago
README.md	Generate Nextclade dataset template from the following commands:	9 hours ago

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

Validate Nextclade Dengue - denv1

clades.nextstrain.org/results

Nextclade

Citation | Docs | Settings | What's new | English

Back

Done. Total sequences: 50. Succeeded: 50

#	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/AILANDKPPKDV08	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

Nucleotide sequence

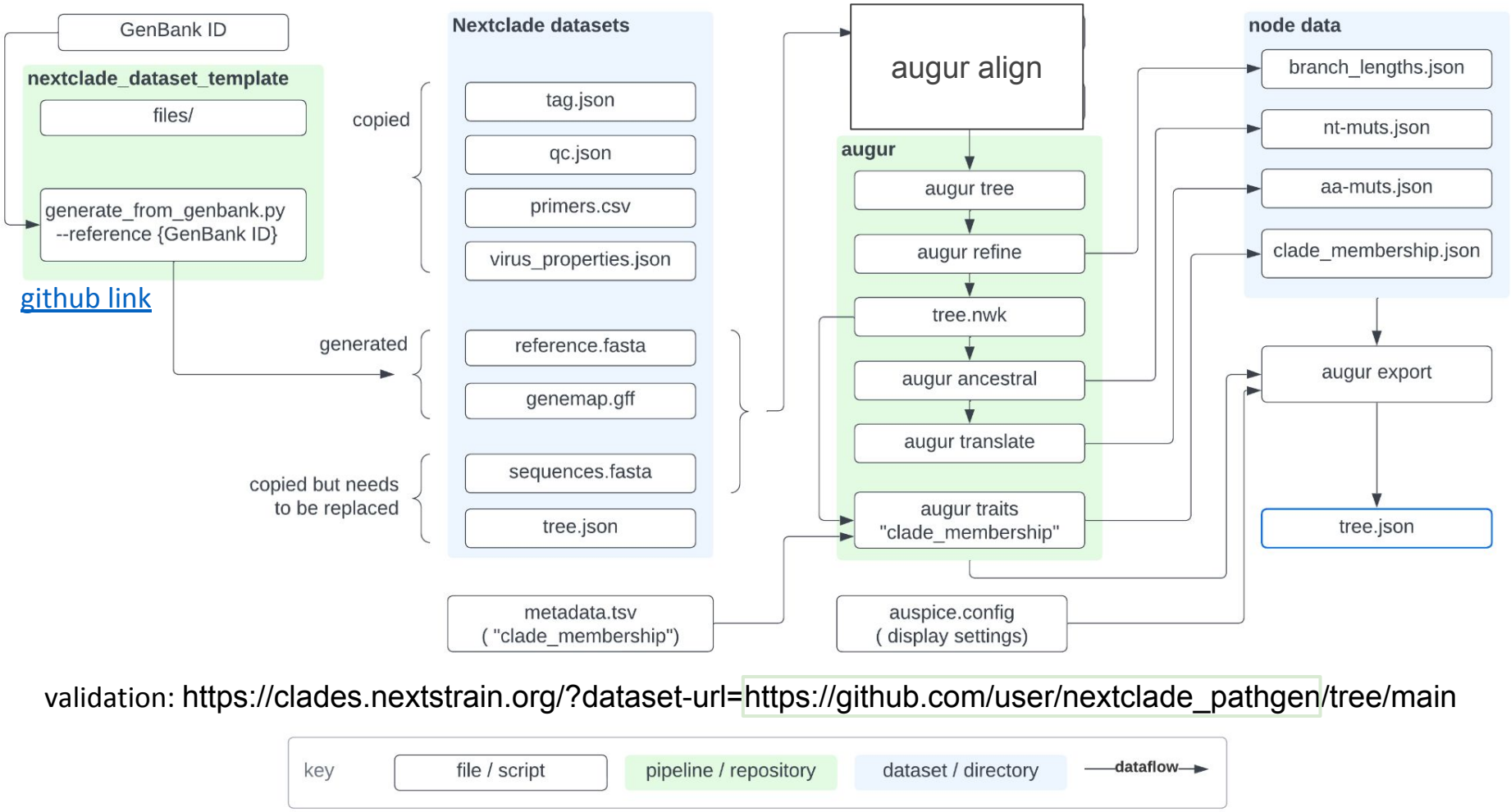
Genome annotation

2000 4000 6000 8000 10000

Nextclade (c) 2020-2023 Nextstrain developers

version 2.14.1 (commit: 85e00e8, branch: release)

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.