Nextstrain gene trees
(for dengue …)

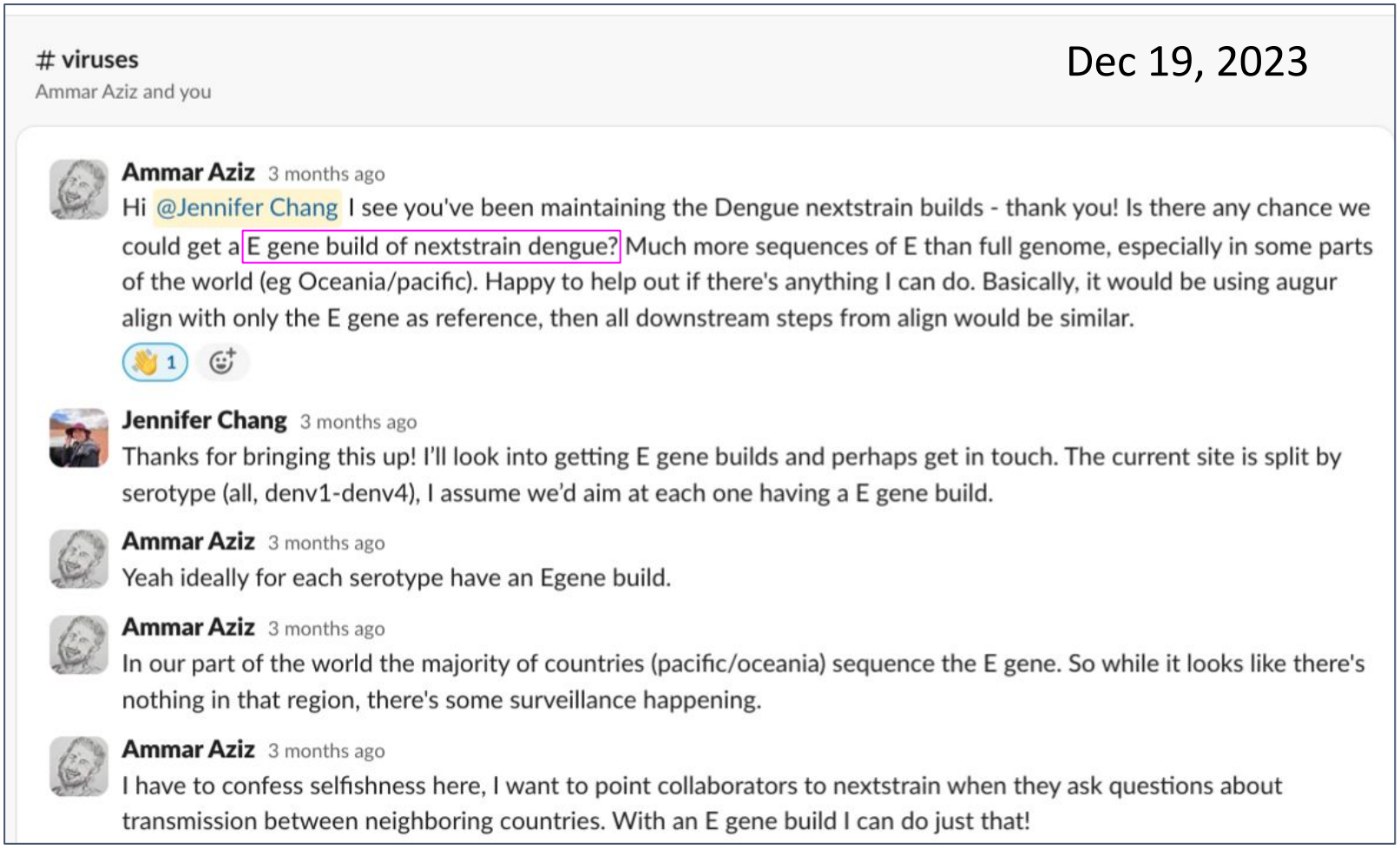
- Bedford Lab Meeting -

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

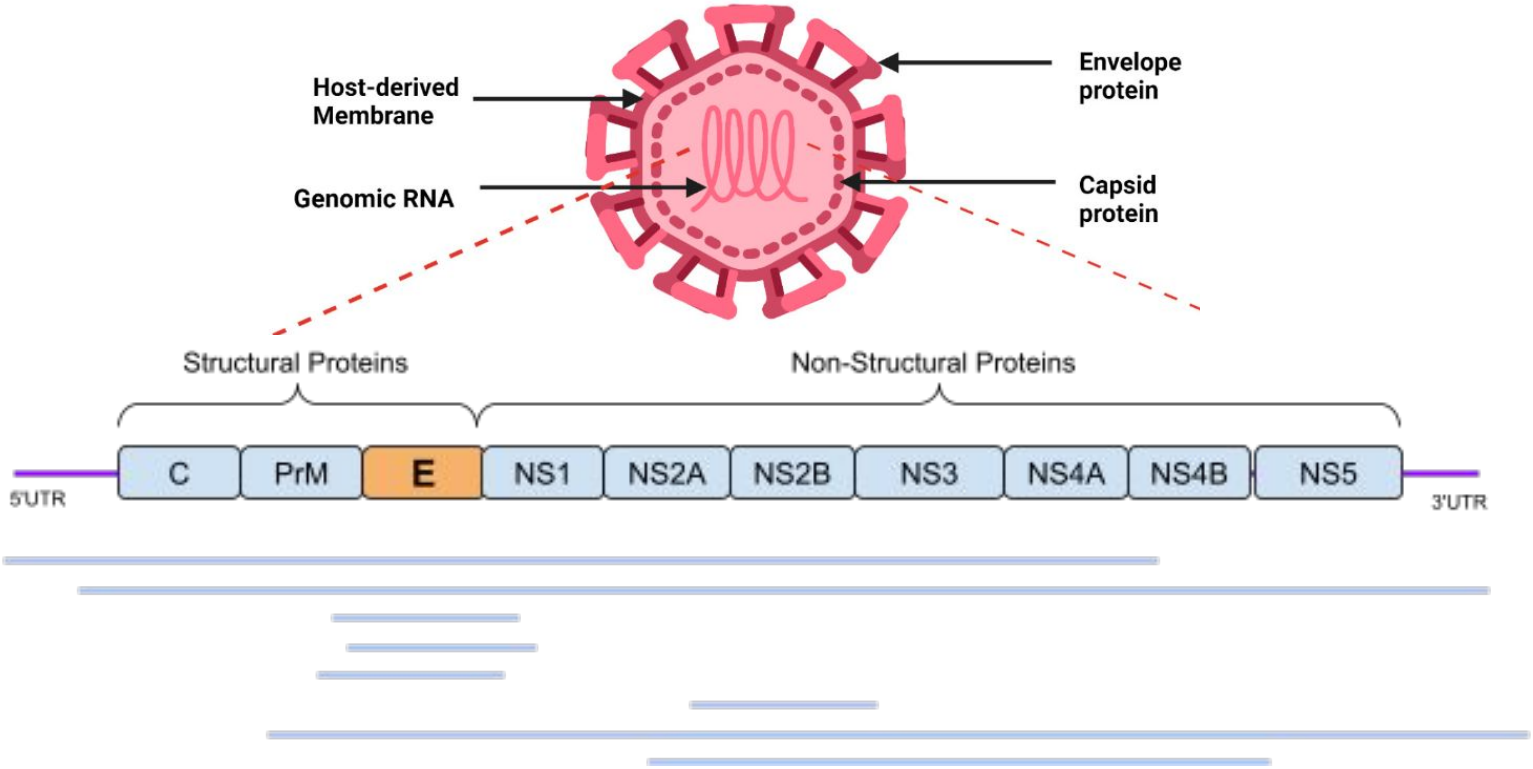
Outline

- **Motivation**
- Overview of modifying the pipeline for "E" gene trees
- Pushing to the live site and future directions

Motivation - user request

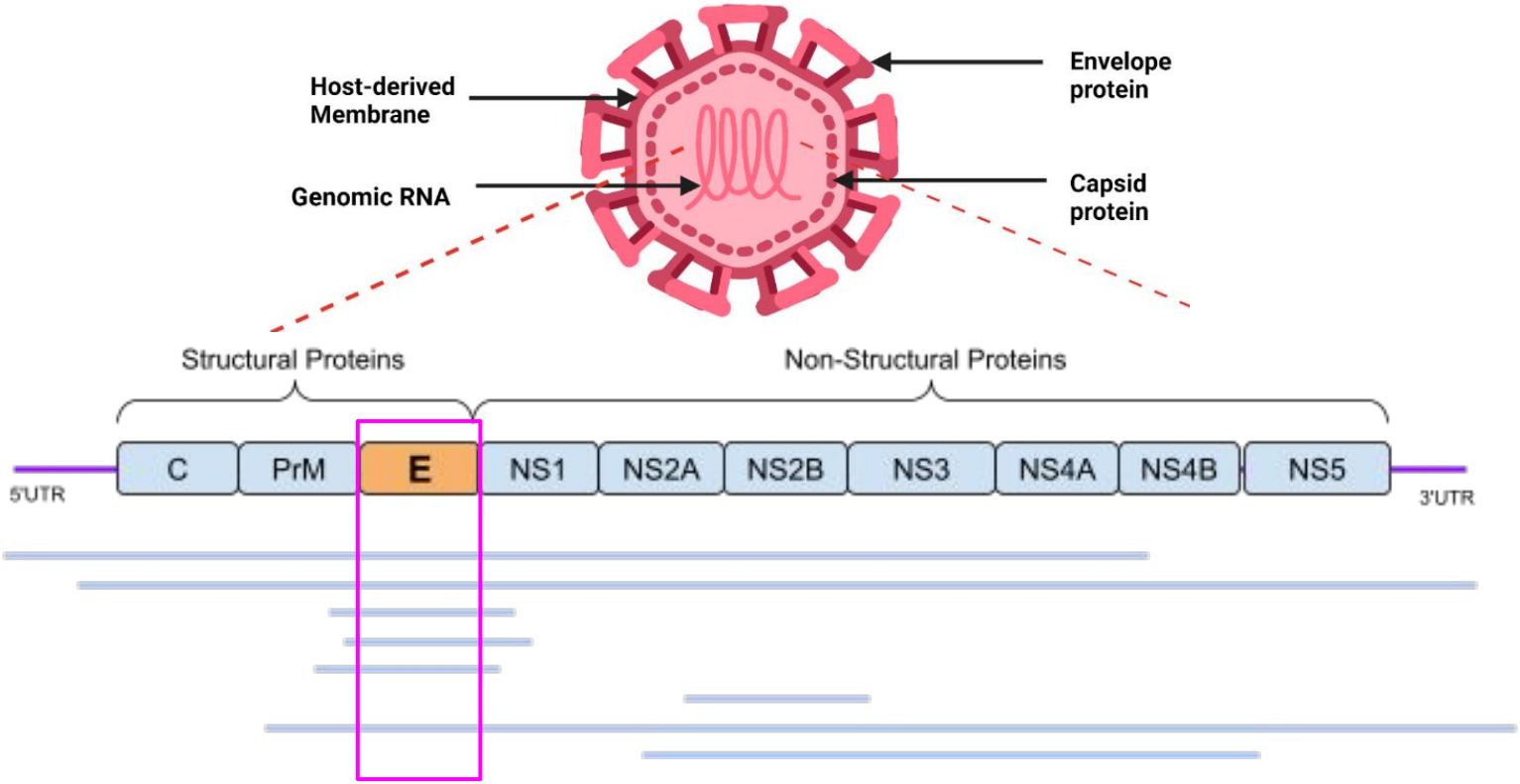


Dengue virus genome



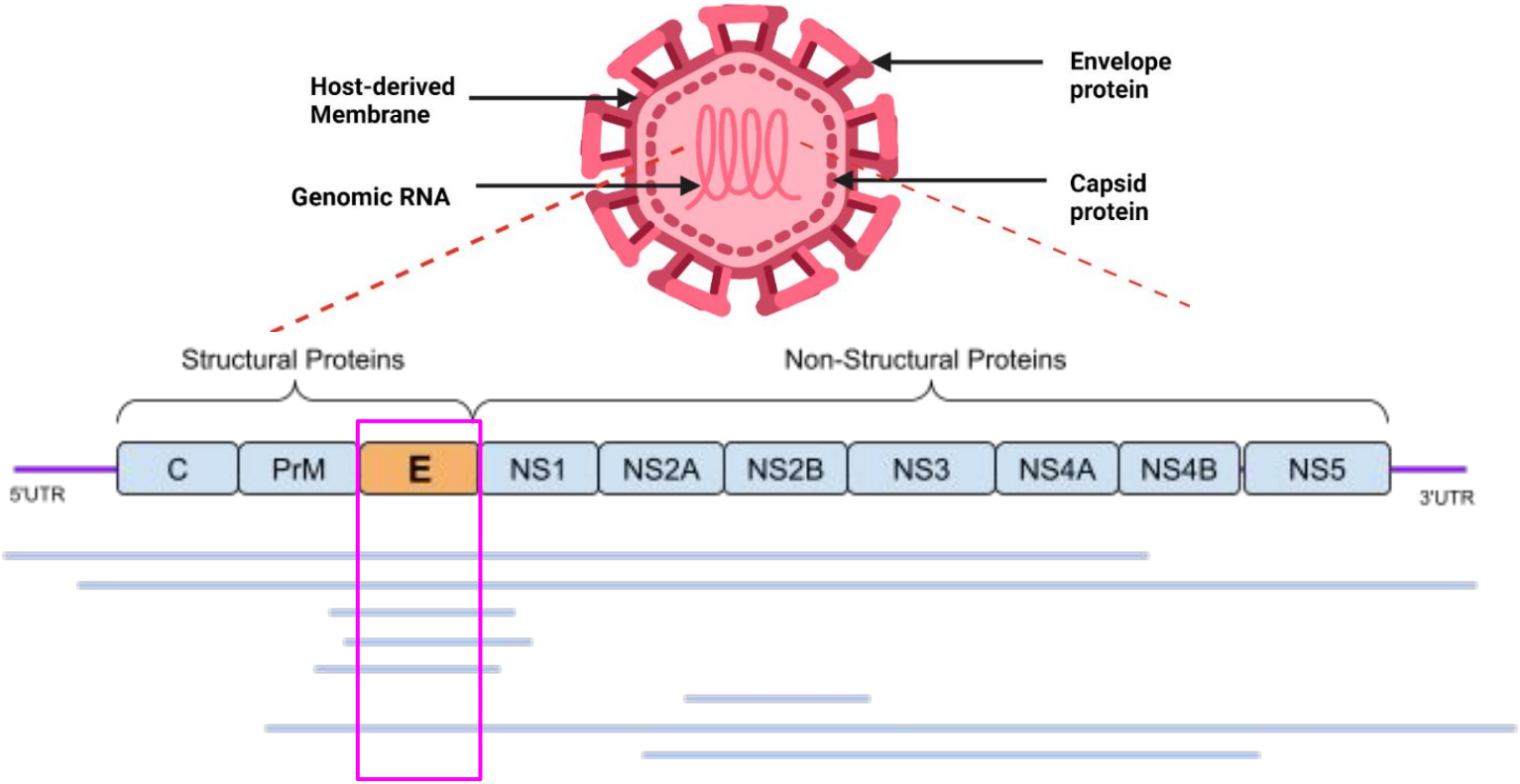
- The genome is about 11kb long encoding 10 genes

Dengue virus genome



- The genome is about 11kb long with 10 genes
- **The "E" gene is 1485nt** (our --min-length is 5000nt)

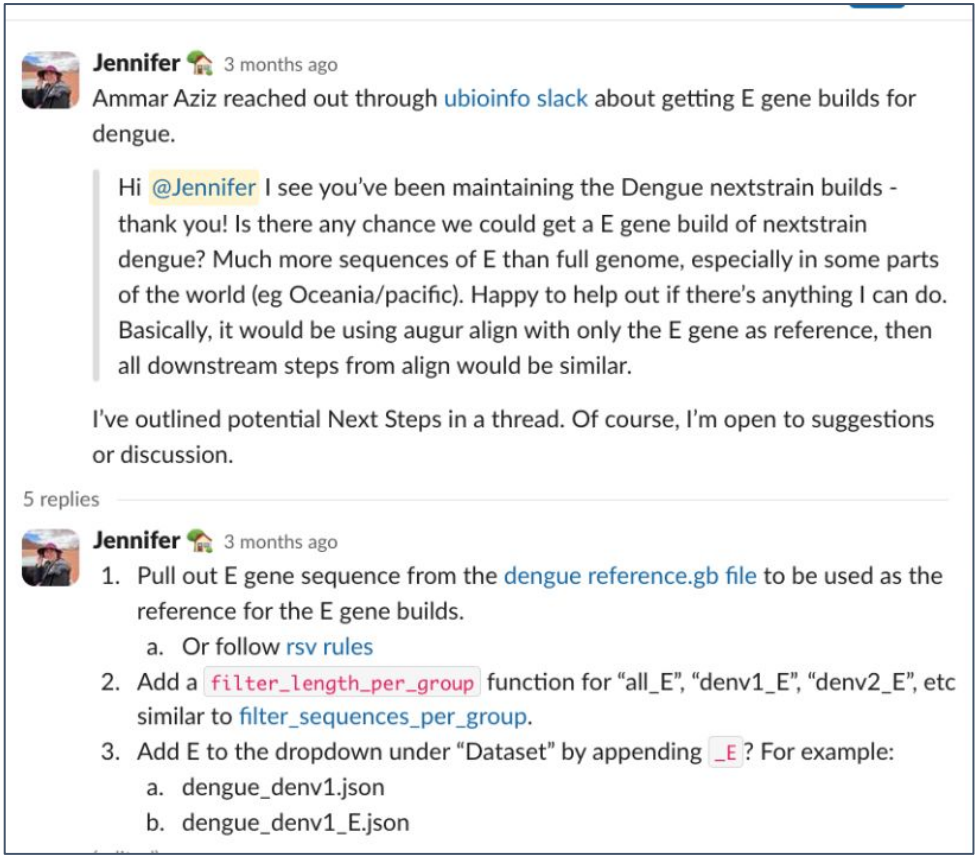
Dengue virus genome



- The genome is about 11kb long with 10 genes
- **The "E" gene is 1485nt** (our --min-length is 5000nt)
- For Dengue, we generate 5 Nextstrain trees (all, denv1 - 4)
- Each serotype has subclades (e.g. denv1/IV, denv2/AA)

Surface the problem to get feedback

Slack to #nextstrain-dev



Jennifer 3 months ago

Ammar Aziz reached out through ubioinfo slack about getting E gene builds for dengue.

> Hi @Jennifer I see you've been maintaining the Dengue nextstrain builds - thank you! Is there any chance we could get a E gene build of nextstrain dengue? Much more sequences of E than full genome, especially in some parts of the world (eg Oceania/pacific). Happy to help out if there's anything I can do. Basically, it would be using augur align with only the E gene as reference, then all downstream steps from align would be similar.

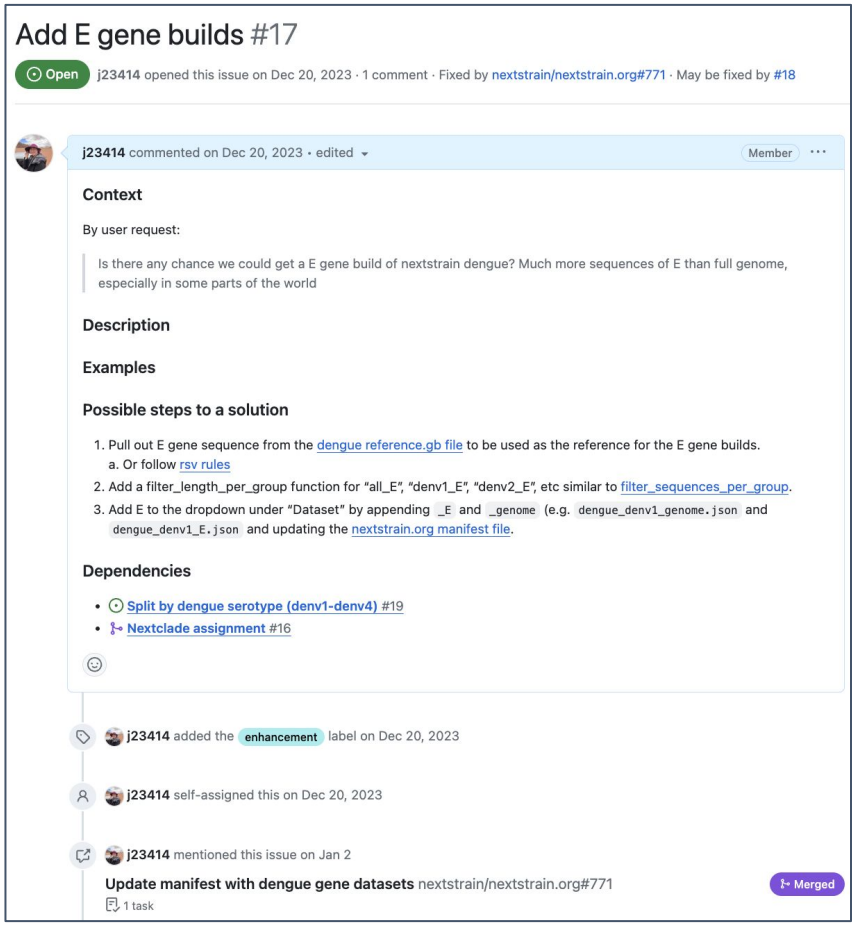
I've outlined potential Next Steps in a thread. Of course, I'm open to suggestions or discussion.

5 replies

Jennifer 3 months ago

1. Pull out E gene sequence from the dengue reference.gb file to be used as the reference for the E gene builds.
   a. Or follow rsv rules
2. Add a `filter_length_per_group` function for "all_E", "denv1_E", "denv2_E", etc similar to filter_sequences_per_group.
3. Add E to the dropdown under "Dataset" by appending `_E`? For example:
   a. dengue_denv1.json
   b. dengue_denv1_E.json

Create an issue on the dengue repo



Add E gene builds #17

Open · j23414 opened this issue on Dec 20, 2023 · 1 comment · Fixed by nextstrain/nextstrain.org#771 · May be fixed by #18

j23414 commented on Dec 20, 2023 • edited — Member

Context

By user request:

> Is there any chance we could get a E gene build of nextstrain dengue? Much more sequences of E than full genome, especially in some parts of the world

Description

Examples

Possible steps to a solution

1. Pull out E gene sequence from the dengue reference.gb file to be used as the reference for the E gene builds.
   a. Or follow rsv rules
2. Add a filter_length_per_group function for "all_E", "denv1_E", "denv2_E", etc similar to filter_sequences_per_group.
3. Add E to the dropdown under "Dataset" by appending `_E` and `_genome` (e.g. `dengue_denv1_genome.json` and `dengue_denv1_E.json` and updating the nextstrain.org manifest file.

Dependencies

- Split by dengue serotype (denv1-denv4) #19
- Nextclade assignment #16

j23414 added the enhancement label on Dec 20, 2023

j23414 self-assigned this on Dec 20, 2023

j23414 mentioned this issue on Jan 2

Update manifest with dengue gene datasets nextstrain/nextstrain.org#771 — Merged

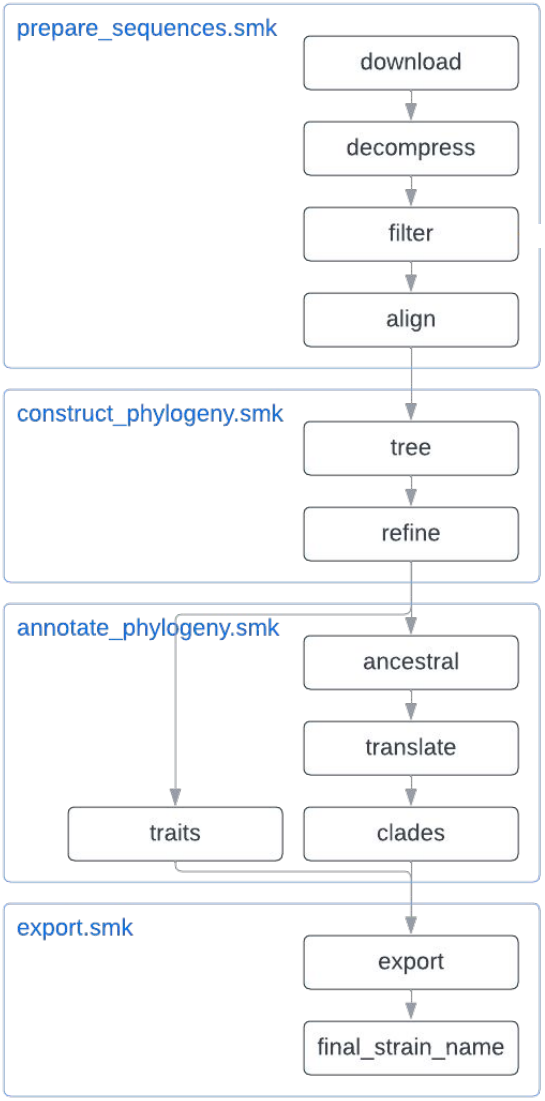
1 task

- Learned that a manifest needed updating
- Confirmed to follow the **RSV rules** for "F" and "G" gene trees

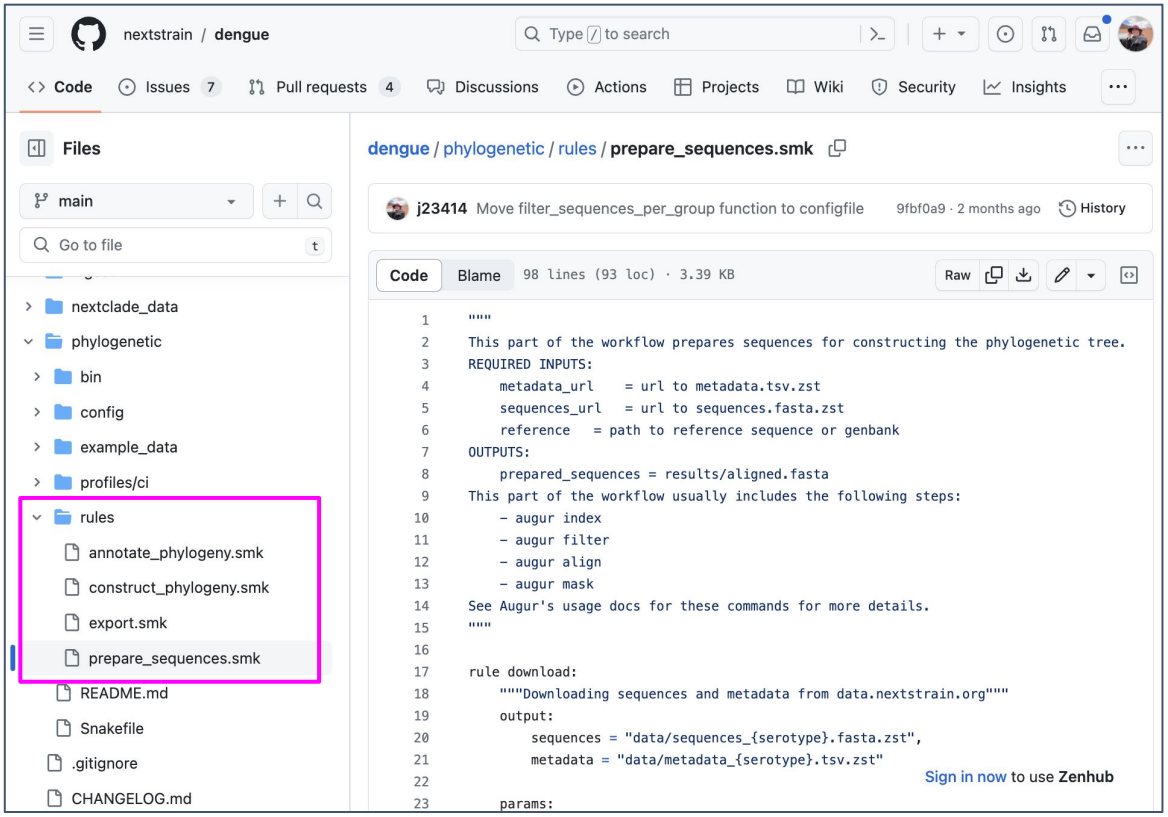
Outline

- Motivation
  - Dec 19, 2023 request for "E" gene trees
  - Surface the problem on slack and github to start the conversation
- **Overview of modifying the pipeline for "E" gene trees**
- Pushing to the live site and future directions

Dengue: phylogenetic pipeline

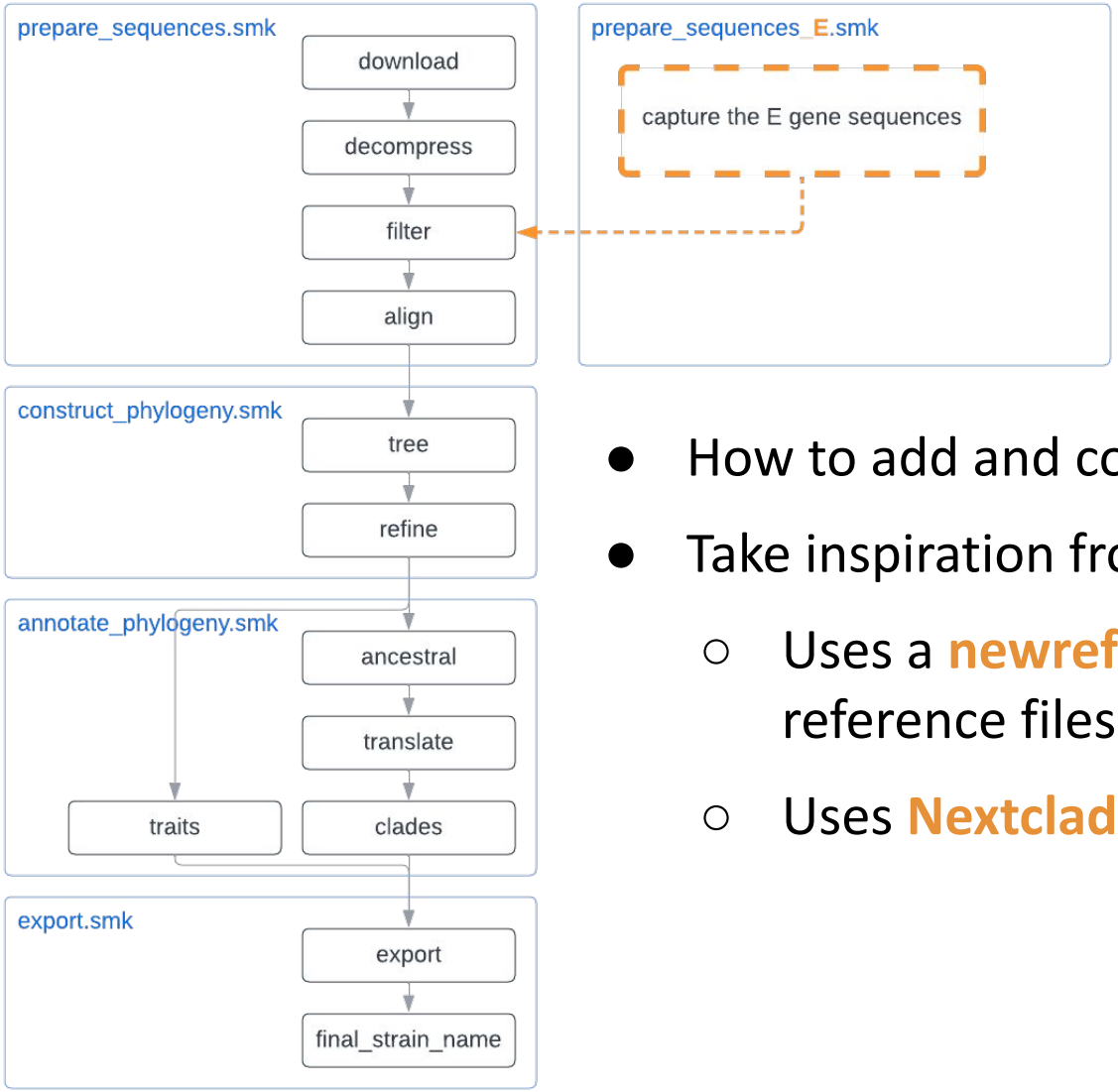


https://github.com/nextstrain/dengue



Pipeline is organized according to the
GitHub: nextstrain/pathogen-repo-guide

Dengue: Adding an "E" gene build



Consider moving this to “ingest” where indicator variables “E”, “genome” in the metadata.tsv

Consider renaming this to “prepare_gene_sequences.smk”

- How to add and connect the "E" gene builds?
- Take inspiration from [GitHub: nextstrain/rsv]
  - Uses a **newreference.py** to generate reference files
  - Uses **Nextclade** + **reference_gene.fasta**

(1/2) Create reference files for "E" gene

dengue / phylogenetic / config / reference_dengue_all.gb

j23414 Move phylogenetic workflow to a phylogenetic folder

Code Blame 271 lines (271 loc) · 17.1 KB

```
1   LOCUS       DENV4/NA/REFERENCE/2003                10649 bp    DNA     VRL 11-FEB-2016
2   DEFINITION  Dengue virus 4, complete genome.
3   ACCESSION   NC_002640
4   VERSION     NC_002640.1
5   DBLINK      BioProject:PRJNA15599
6   KEYWORDS    RefSeq.
7   SOURCE      Dengue virus 4
8     ORGANISM  Dengue virus 4
9               Viruses; ssRNA viruses; ssRNA positive-strand viruses, no DNA stage;
10              Flaviviridae; Flavivirus; Dengue virus group
...
50        CDS             441..713
51                        /gene="pr"
52                        /note="peptide pr"
53                        /product="protein pr"
54                        /protein_id="YP_009164957.1"
55        CDS             939..2423
56                        /gene="E"
57                        /product="envelope protein E"
58                        /protein_id="NP_740317.1"
59        CDS             2424..3479
60                        /gene="NS1"
61                        /product="nonstructural protein NS1"
62                        /protein_id="NP_740318.1"
63        CDS             3480..4133
64                        /gene="NS2A"
65                        /product="nonstructural protein NS2A"
```

Sign in now to use Zenhub

```
python newreference.py \
  --reference reference_dengue_all.gb \
  --output-genbank E.gb \
  --output-fasta E.fasta \
  --gene E
```

Consider adding “-start” and “-end” flags for sub regions of a gene (e.g. Measles locus in N)

- Captures the "CDS" coordinates for "E" gene
- Generates a FASTA and GenBank file

(1/2) Create reference files for "E" gene

E.gb

```
LOCUS       DENV4/NA/REFERENCE/2003 1485 bp    DNA             UNK 01-JAN-1980
DEFINITION  Dengue virus 4, complete genome.
ACCESSION   NC_002640
VERSION     NC_002640.1
KEYWORDS    .
SOURCE      .
  ORGANISM  .
            .
FEATURES             Location/Qualifiers
     CDS             1..1485
                     /gene="E"
                     /product="envelope protein E"
                     /protein_id="NP_740317.1"
     source          1..1485
                     /clone="rDEN4"
                     /db_xref="taxon:11070"
                     /mol_type="genomic RNA"
                     /organism="Dengue virus 4"
ORIGIN
        1 atgcgatgcg taggagtagg aaacagagac tttgtggaag gagtctcagg tggagcatgg
       61 gtcgacctgg tgctagaaca tggaggatgc gtcacaacca tggcccaggg aaaaccaacc
      121 ttggattttg aactgactaa gacaacagcc aaggaagtgg ctctgttaag aacctattgc
      181 attgaagcct caatatcaaa cataactacg gcaacaagat gtccaacgca aggagagcct
      241 tatctgaaag aggaacagga ccaacagtac atttgccgga gagatgtggt agacagaggg
      301 tggggcaatg gctgtggctt gtttggaaaa ggaggagttg tgacatgtgc gaagttttca
      361 tgttcgggga agataacagg caatttggtc caaattgaga accttgaata cacagtggtt
      421 gtaacagtcc acaatggaga cacccatgca gtaggaaatg acacatccaa tcatggagtt
      481 acagccatga taactcccag gtcaccatcg gtggaagtca aattgccgga ctatggagaa
      541 ctaacactcg attgtgaacc caggtctgga attgacttta atgagatgat tctgatgaaa
      601 atgaaaaaga aaacatggct cgtgcataag caatggtttt tggatctgcc tcttccatgg
      661 acagcaggag cagacacatc agaggttcac tggaattaca aagagagaat ggtgacattt
      721 aaggttcctc atgccaagag acaggatgtg acagtgctgg gatctcagga aggagccatg
      781 cattctgccc tcgctggagc cacagaagtg gactccggtg atggaaatca catgtttgca
      841 ggacatctta agtgcaaagt ccgtatggag aaattgagaa tcaagggaat gtcatacacg
      901 atgtgttcag gaaagttttc aattgacaaa gagatggcag aaacacagca tgggacaaca
      961 gtggtgaaag tcaagtatga aggtgctgga gctccgtgta aagtccccat agagataaga
```

E.fasta

```
>NC_002640.1 Dengue virus 4, complete genome
ATGCGATGCGTAGGAGTAGGAAACAGAGACTTTGTGGAAGGAGTCTCAGGTGGAGCATGG
GTCGACCTGGTGCTAGAACATGGAGGATGCGTCACAACCATGGCCCAGGGAAAACCAACC
TTGGATTTTGAACTGACTAAGACAACAGCCAAGGAAGTGGCTCTGTTAAGAACCTATTGC
ATTGAAGCCTCAATATCAAACATAACTACGGCAACAAGATGTCCAACGCAAGGAGAGCCT
TATCTGAAAGAGGAACAGGACCAACAGTACATTTGCCGGAGAGATGTGGTAGACAGAGGG
TGGGGCAATGGCTGTGGCTTGTTTGGAAAAGGAGGAGTTGTGACATGTGCGAAGTTTTCA
TGTTCGGGGAAGATAACAGGCAATTTGGTCCAAATTGAGAACCTTGAATACACAGTGGTT
GTAACAGTCCACAATGGAGACACCCATGCAGTAGGAAATGACACATCCAATCATGGAGTT
ACAGCCATGATAACTCCCAGGTCACCATCGGTGGAAGTCAAATTGCCGGACTATGGAGAA
CTAACACTCGATTGTGAACCCAGGTCTGGAATTGACTTTAATGAGATGATTCTGATGAAA
ATGAAAAAGAAAACATGGCTCGTGCATAAGCAATGGTTTTTGGATCTGCCTCTTCCATGG
ACAGCAGGAGCAGACACATCAGAGGTTCACTGGAATTACAAAGAGAGAATGGTGACATTT
AAGGTTCCTCATGCCAAGAGACAGGATGTGACAGTGCTGGGATCTCAGGAAGGAGCCATG
CATTCTGCCCTCGCTGGAGCCACAGAAGTGGACTCCGGTGATGGAAATCACATGTTTGCA
GGACATCTTAAGTGCAAAGTCCGTATGGAGAAATTGAGAATCAAGGGAATGTCATACACG
ATGTGTTCAGGAAAGTTTTCAATTGACAAAGAGATGGCAGAAACACAGCATGGGACAACA
GTGGTGAAAGTCAAGTATGAAGGTGCTGGAGCTCCGTGTAAAGTCCCCATAGAGATAAGA
GATGTAAACAAGGAAAAAGTGGTTGGGCGTATCATCTCATCCACCCCTTTGGCTGAGAAT
ACCAACAGTGTAACCAACATAGAATTAGAACCCCCCTTTGGGGACAGCTACATAGTGATA
GGTGTTGGAAACAGCGCATTAACACTCCATTGGTTCAGGAAAGGGAGTTCCATTGGCAAG
ATGTTTGAGTCCACATACAGAGGTGCAAAACGAATGGCCATTCTAGGTGAAACAGCTTGG
GATTTTGGTTCCGTTGGTGGACTGTTCACATCATTGGGAAAGGCTGTGCACCAGGTTTTT
GGAAGTGTGTATACAACCATGTTTGGAGGAGTCTCATGGATGATTAGAATCCTAATTGGG
TTCTTAGTGTTGTGGATTGGCACGAACTCGAGGAACACTTCAATGGCTATGACGTGCATA
GCTGTTGGAGGAATCACTCTGTTTCTGGGCTTCACAGTTCAAGCA
```

- The gene coordinates are updated
- The fasta begins with the start codon "ATG"

(2/2) Use Nextclade to pull out "E" gene

```
nextclade run \
   --input-ref E.fasta \
   --output-fasta E_sequences.fasta \
   --min-seed-cover 0.01 \
   --min-length 1000 \
   --silent \
   sequences_all.fasta
```

```
real 0m17.322s
user 2m8.163s
sys  0m0.500s
```

phylogenetic — less E_sequences.fasta — 98×40

```
>X65240
ATGCGTTGCATAGGAATATCAAATAGAGACTTTGTAGAAGGGGTTTCAGGAGGAAGCTGGGTTGACATAGTCTTAGAACATGGAAGTTGTGTGACGAC
GATGGCAAAAAACAAACCAACATTGGATTTTGAACTGATAAAAACAGAAGCCACACAACTGCCACTCTAAGGAAG-TACTGTATAGAAGCCTGA----
CCAATACAACAACAGAATCTCGTTGCCCAACACAAGGGGAACCCAGTCTAAATGAAGAGCAGGACAAAAGGTTCGTCTGC---AAACACTTGGTAGAC
AGAGGATGGGGAAATGGATGTGGATTATTTGGAAAAGGAGGTATTGTGACCTGTGCTATGTTTACATGCAAAAAGAACATGGAAGGAAACATCGTGCA
ACCAGAAAATTTGGAATACACCATCGTGATAACACCTCACTCAGGAGAAGAGCACGCTGTAGGTAACGACACAGGAAAACATGGCAAGGAAATCAAAA
TAACACCCAGAGTTCCATCACAG-AAGCAGAACTAACAGGCTATGGCACCGTCACGATGGAGTGCTCTCCGAGAACGGGCCTGGACTTCAATGAGATA
GTACTGCTGCAGATGGAAGACAAAGCTTGGCTAGTGCACAGGCAATGGTTCCTAGACCTGCCGTTACCATGGCTACCCGGAGCGGACACA-CA-AGGA
TCA----AATTACAAAGAGACATTGGTCACTTTCAAAAATCCCCATGCGAAGAAACAGGATGTCGTTGTTTTAGGATCTCAAGAAGGGGCCATGCACA
CGGCACTCACAGGGGCCACAGAAATCCA-------GATGGAAATTAC-TATTCACAGGACATCTCAAGTGCAGACTGAGAATGGACAAACTACAGCTC
AAAGGAATGTCATACTCTATGTGTACAGGAAAGTTTCAAATTGTGAAGGAAATAGCAGAAACACAACATGGAACAATAGTTATCAGAGTACAATATGA
AGGAGACGGCTCTCCATGTAAGATCCCTCTTGAGATAATGGATGGAAAAAAG--ACATGTCTTAGGTCGCCTGATTACATTAACCCGATCGTAACAGA
AAACC--CAGT-----CAACATAGAAGCAGAACCTCCATTCGGAGACAGCTACATCATCATAGGAGTAGGGGACA----ATTGAAACTCCACTGGTTT
AAGAAAGGAAGTTCCATCGGCCAAATGTTTGAGACAACAATGAGAGGAGCGAAGAGAATGGCCATTTTAGGTGACACAGCCTGGGATTTTGGATCCCT
GGGAGGAGTGTTTACATCTATAGGAAAGGCTCTCCACCAAGTTTTCGGAGCTATCTATGGGGCCGCTTTTAGTGGGGTCTCATGGACTATGAAAATCC
TCATCGGAGTCATCATCACATGGATAGGAATGAATTCACGTAGCACCTCACTGTCTGTGACTAGTATTGGTG--GGAGTCATAACACTGTACTTGGGA
GCCATGGTGCAGGCT
>X76219
ATGCGATGCGTGGGAATAGGCAACAGAGACTTCGTGGAAGGACTGTCAGGAGGAACGTGGGTGGATGTGGTACTGGAGCATGGAAGTTGCGTCACCAC
CATGGCAAAAGATAAACCAACATTGGACATTGAACTCTTGAAGAC---------GGAGGTCACTGCGTAAA----CTGT-GCATTGAAGCTAAAATAT
CAAACACCACCACCGATTCAAGATGTCCAACACAAGGGGAGCCACA-CTGGTGGAAGAACAAGACGCGAACTTCGTGTGTCGACGAACGTTTGTGGAC
AGAGGCTGGGGCAATGGCTGTGGGCTTTTCGGAAAAGGTAGCCTAATAACGTGTGCTAAGTT--CAAGTGTGTG--ACAAAACAGGAAAGATTGTTCA
ATATGAGAACTTGAAATATTCAGTGATAGTCACCGTCCACACTGGTGACCAGCACCAGGTGGGAAATGAGACCACAGAACATGGAATTGCAACCATAA
CACCTCCTACGTCA---------GAAATACAGCTGACCGACTACGGAGCTCTTACATTGGATTGCTCACCCAGAACAGGGCTAGACTTTAATGAGATG
GTGTTGTTGACAATGAAAGAAAAATCATGGCTTGTCCACAAACAATGGTTTCTAGACTTACCACTGCCCTGGACCTCGGGAGCTTCAACACCAGAGAC
T---TGGAAC--------AGAGAATGGTTACATTTAAGACAGCTCATGCAAAGAAGCAGGAAGTAGTCGTACTAGGATCACAAGAAGGAGCAATGCACA
CTGCGTTGACCGGAGCGACAGAAATCCAAACGTCTGACGACAA--AAATTTTTGCAGGACACTTGAAATGTAGACTAAAAATGGACAAACTGACCTTA
AAAGGGATGTCATATGTGATGTGCACAGGA-----TTCAAGTGAGAAAGAAGTGGCTGAGACCCAGCATGGAACTGTTCTAGTGCAGGTTAAATACGA
AGGAACAGATGCACCATGCAAGATCCCCTTTTAGATGAGAAAGGTAACCCAGAA----------TGGGAGATTGATAACAGCCAACCCCATAGCTGAGA
AAAAC--CAGT-----CAACATTGAGGCAGAACCACCTTTTGGTGAGAATTACATCGTGGTAGGAGCAGGTGAAAAAGCTTTGAAACTAAGCTGGTTC
AAGAAAGGAAGCAGCATTGGGAAAATGCTTGAAGCAACTGCCCGAGGAGCACGAAGGACGGCCATCCTAGGAGACACCGCATGGGACTTCGGTTCTAT
AGGAGGAGTGTTCACGTCTGTGGGAAAACTGGTACACCAGATCTTTGGAACTGCATATGGAGTTTTGTTCAGCGGTGTTTCCTGGACTATGAAAATAG
GAATAGGGATTCTGCTGACATGGCTAGGATTAAATTCAAGGAGCACGTCCCTTTCGACGACGTGCATTGCAGTTGGCATGGTTACACTGTACCTAGGA
GTCATGGTTCAGGCG
>X15433
ATGCGTTGCATAGGAATATCAAATAGAGACTTTGTAGAGGGGGTTTCAGGAGGAAGCTGGGTTGACATAGTCTTAGAACATGGAAGCTGTGTGACGAC
GATGGCAAAAAACAAACCAACATTGGATTTTGAACTGATAAAAACAGAAGCCAAACAGCCTGCCACCCTAAGGAAGTACTGTATAGAAGCCTTA----
CCAACACAACAACAGAATCTCGTTGCCCAACACAAGGGGAACCCAGCCTAAATGAAGAGCAGGACAAACACTCCAT---------------GGTAGAC
AGAGGATGGGGAAATGGATGTGGACTATTTGGAAAGGGAGGCATTGTGACCTGTGCTATGTTCACATGCCAAAAGAACATGGAAGGAAAAATCGTGCA
:
```

- Starts with "ATG"
- Seems to be aligned
- Lower threshold for diverse viruses ([Nextclade 3.2 changelog](#))

Compare "Nextclade: and "Augur align"

```
nextclade run \
   --input-ref E.fasta \
   --output-fasta E_sequences.fasta \
   --min-seed-cover 0.01 \
   --min-length 1000 \
   --silent \
   sequences_all.fasta
```

```
augur align \
   --reference-sequence E.fasta \
   --output augur_E_sequences.fasta \
   --fill-gaps \
   --sequences sequences_denv4.fasta
```

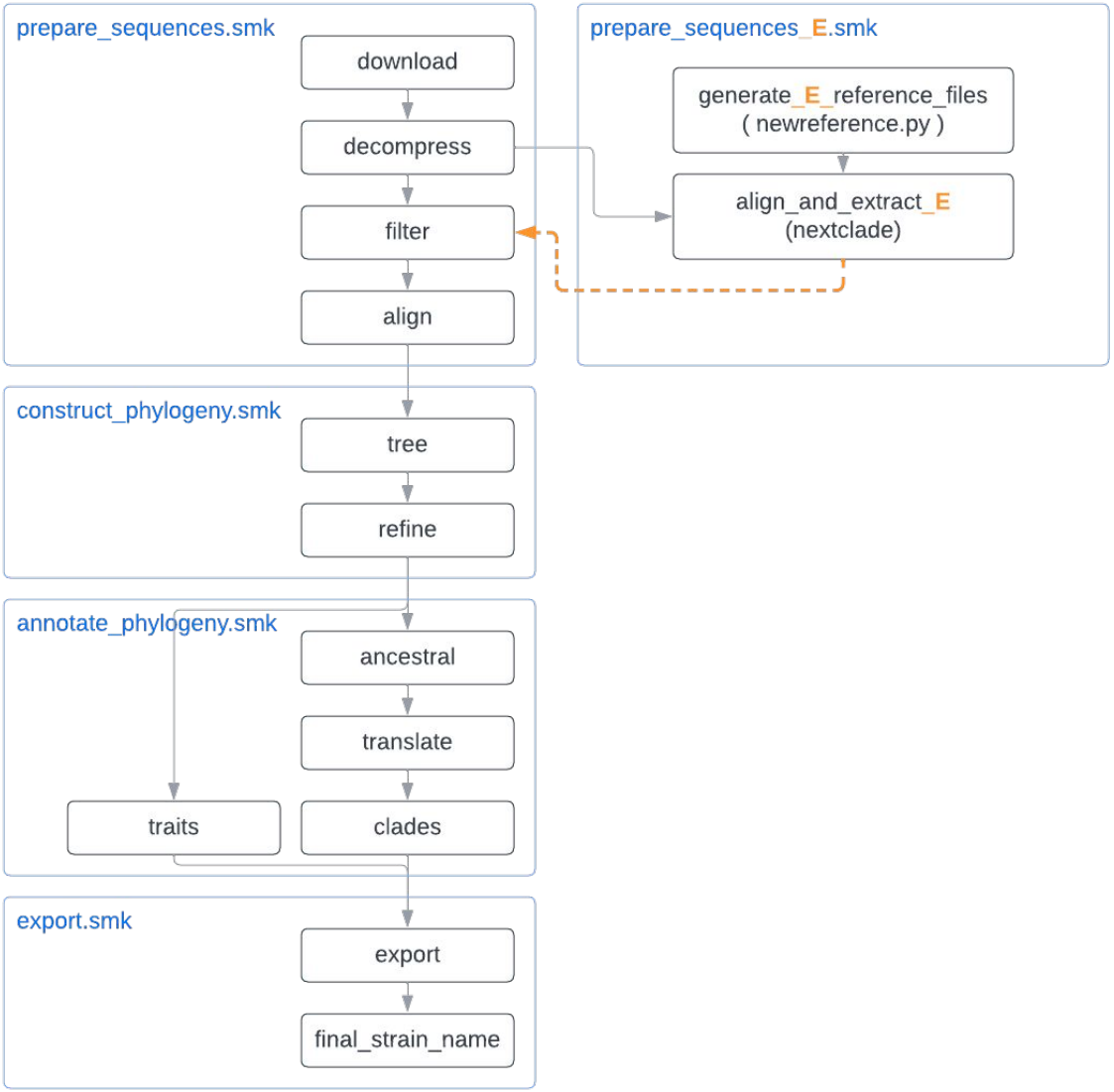
```
real 0m17.322s
user 2m8.163s
sys  0m0.500s
```

```
real 85m37.558s
user 85m3.708s
sys  0m8.019s
```

sequences_all.fasta (46,342)
time: less than a minute

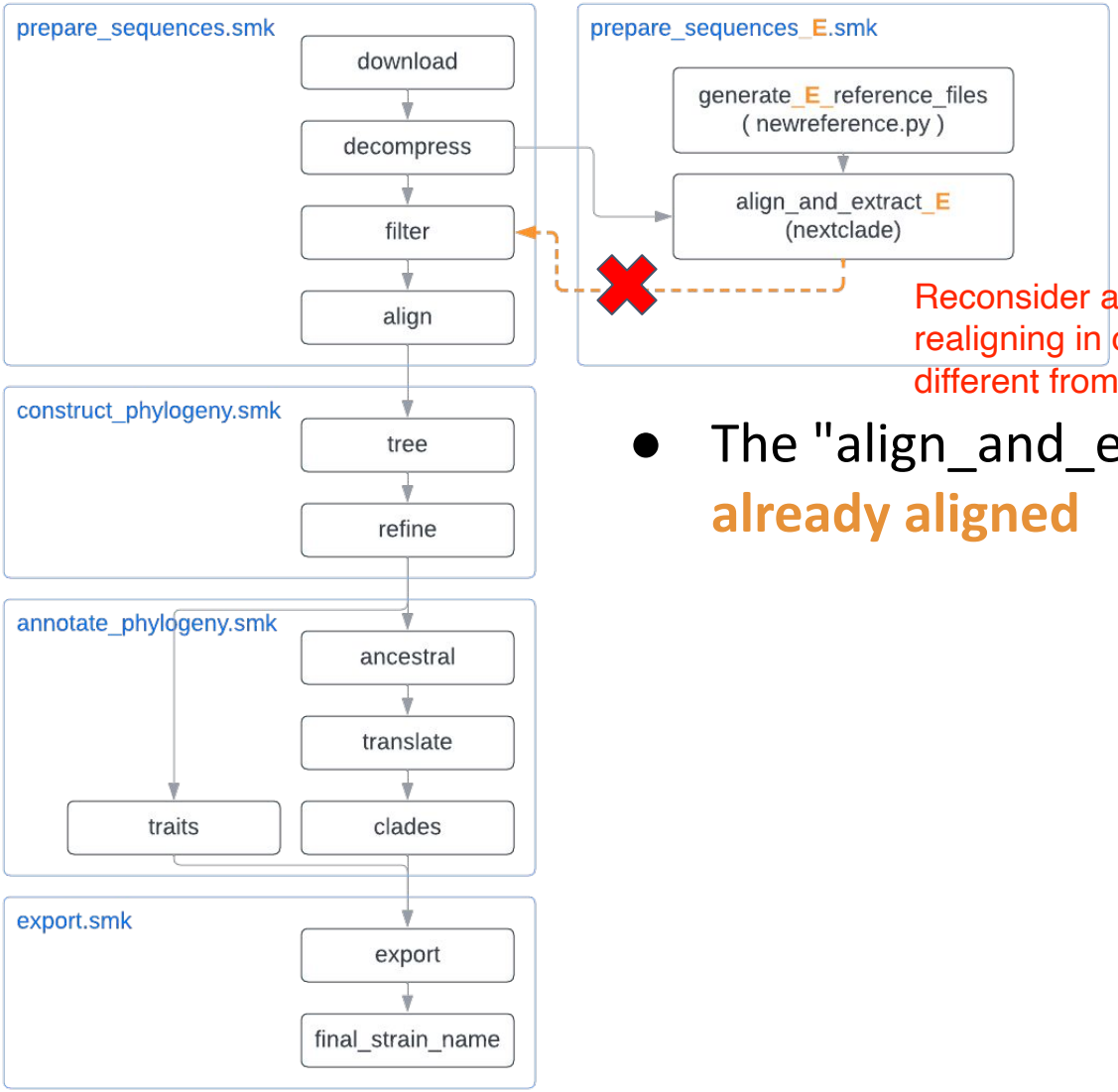
sequences_denv4.fasta (3,874)
time: around 1.5 hours

Prepare "E" gene sequences



How to connect the **aligned E sequences** to the pipeline?

Prepare "E" gene sequences

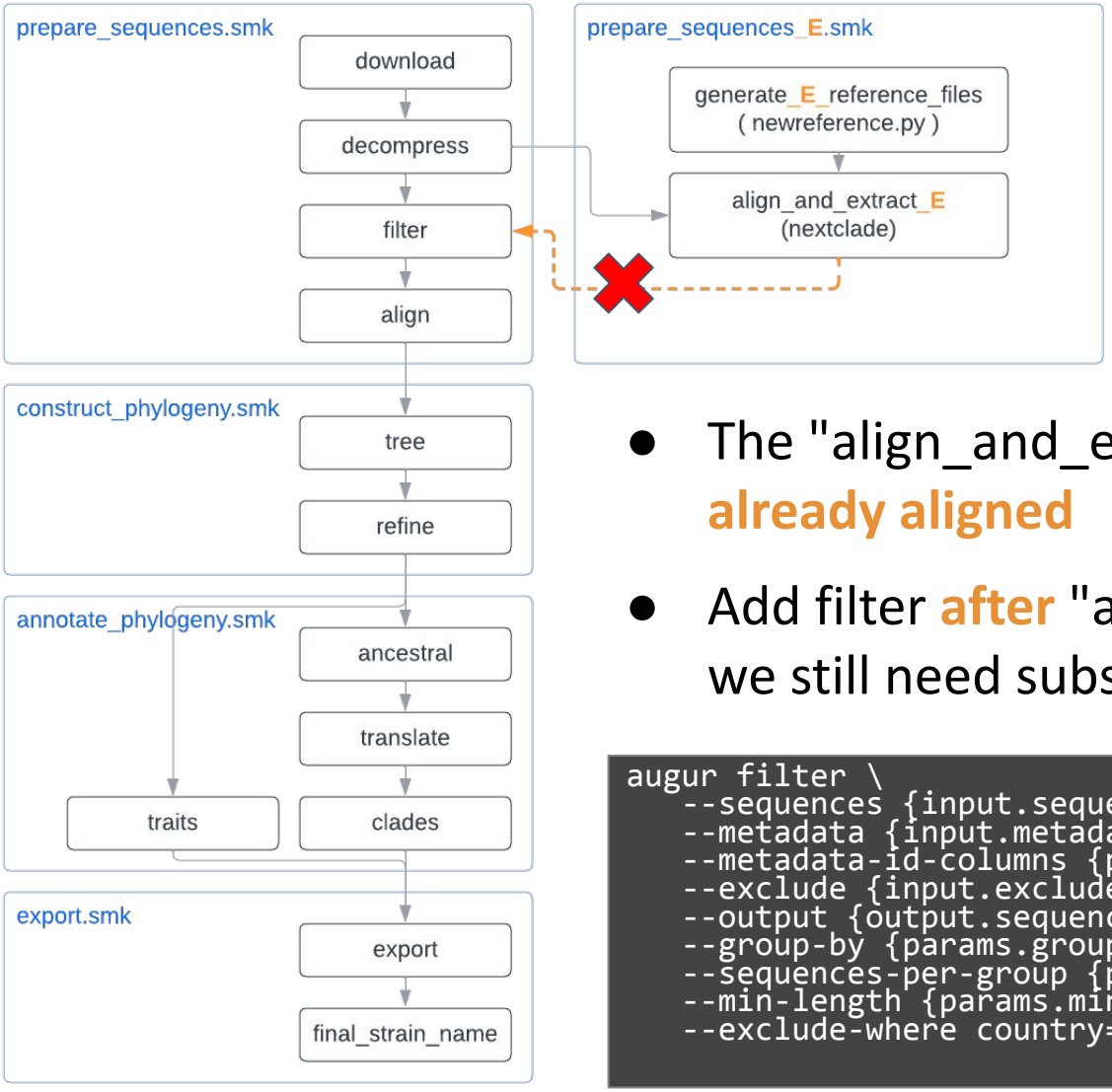


How to connect the **aligned E sequences** to the pipeline?

Reconsider attaching at the filter step, and realigning in case the Nextclade alignment is different from MAFFT (may simplify wildcards)

- The "align_and_extract_E" output is **already aligned**

Prepare "E" gene sequences

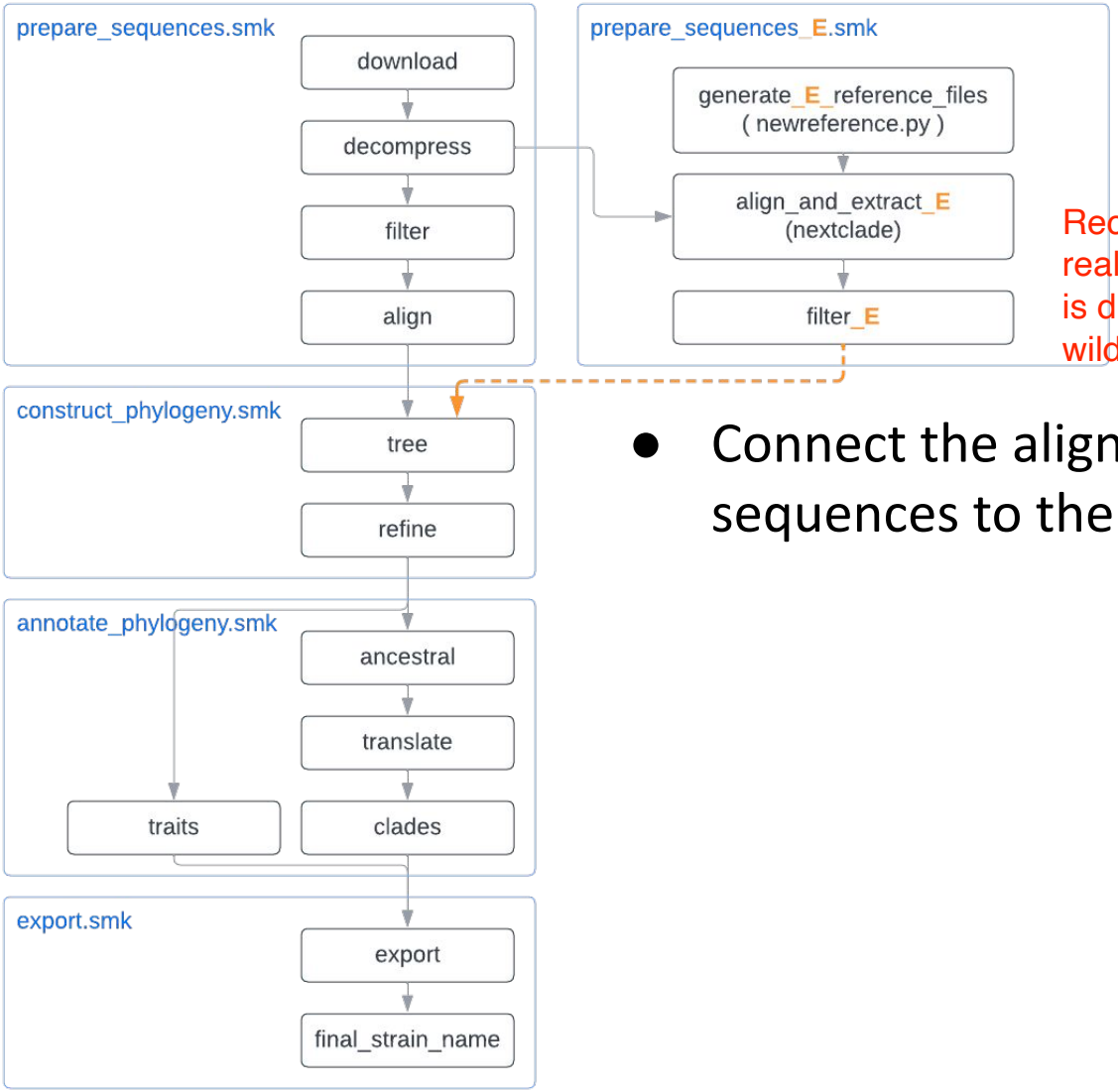


How to connect the **aligned E sequences** to the pipeline?

- The "align_and_extract_E" output is **already aligned**
- Add filter **after** "align_and_extract_E" since we still need subsampling

```
augur filter \
   --sequences {input.sequences} \
   --metadata {input.metadata} \
   --metadata-id-columns {params.strain_id} \
   --exclude {input.exclude} \
   --output {output.sequences} \
   --group-by {params.group_by} \
   --sequences-per-group {params.sequences_per_group} \
   --min-length {params.min_length} \
   --exclude-where country=? region=? date=? \
```

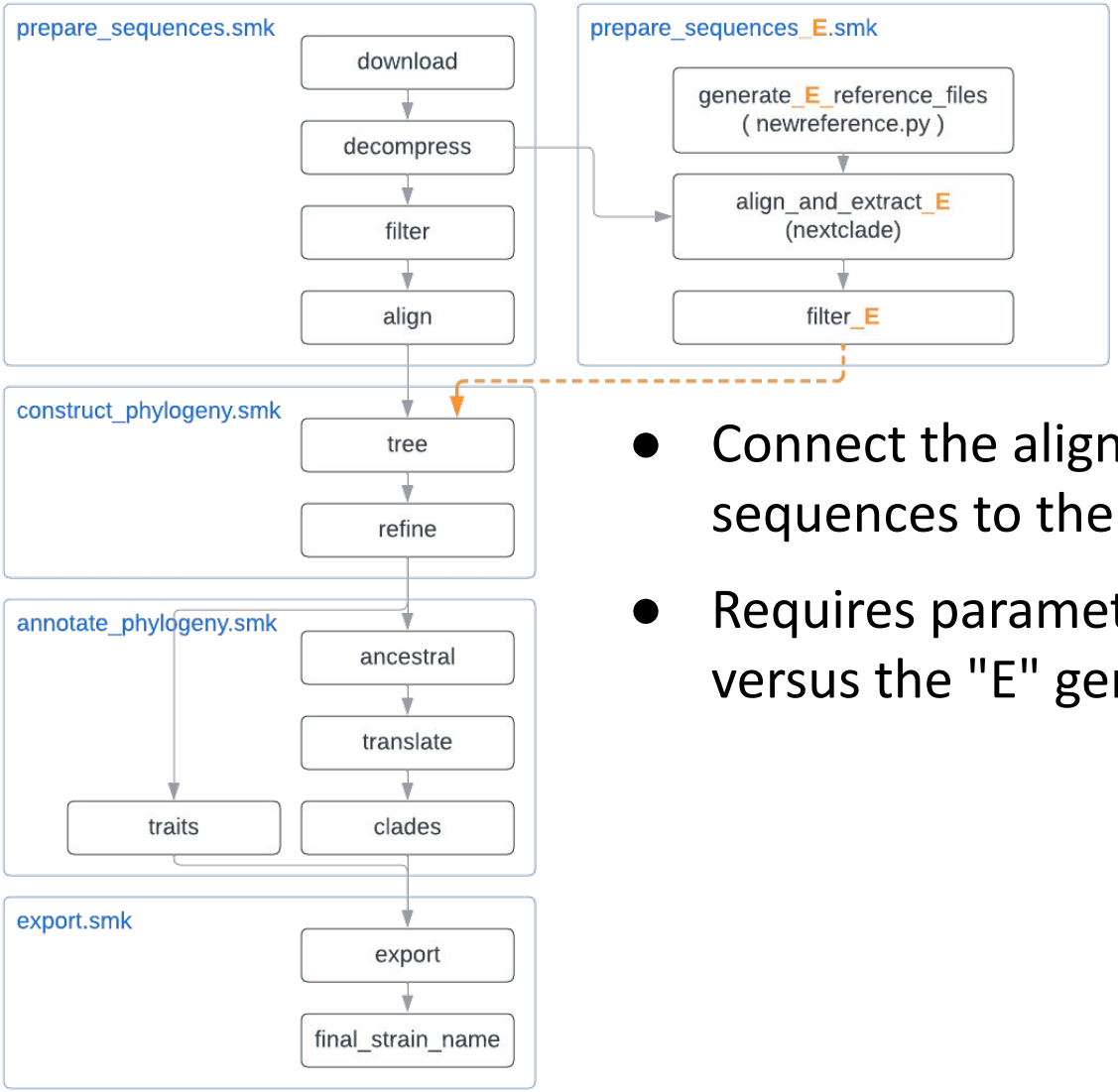
Push "E" sequences through pipeline



Reconsider attaching at the filter step, and realigning in case the Nextclade alignment is different from MAFFT (may simplify wildcards)

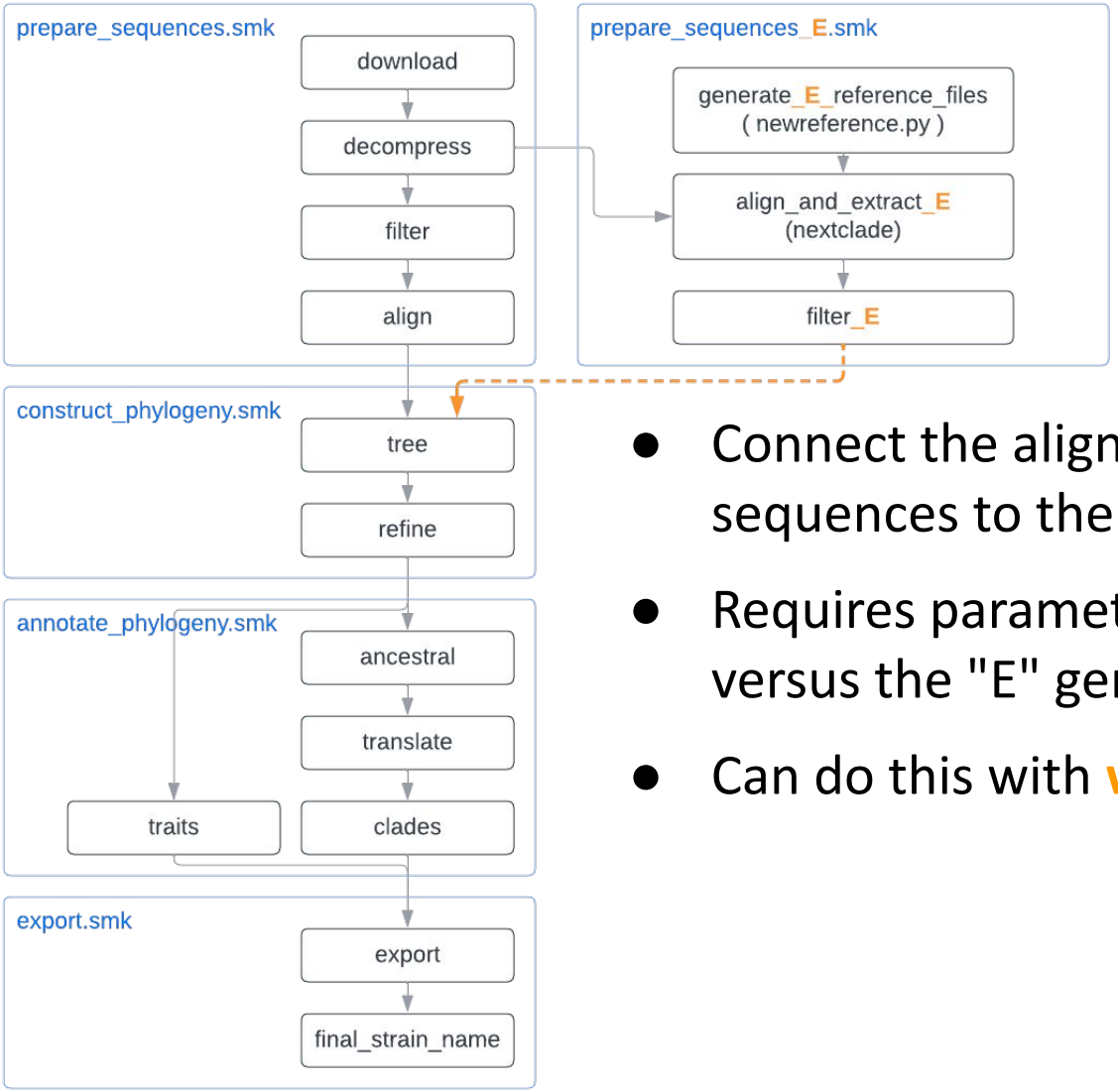
- Connect the aligned and subsampled E gene sequences to the rest of the pipeline

Push "E" sequences through pipeline



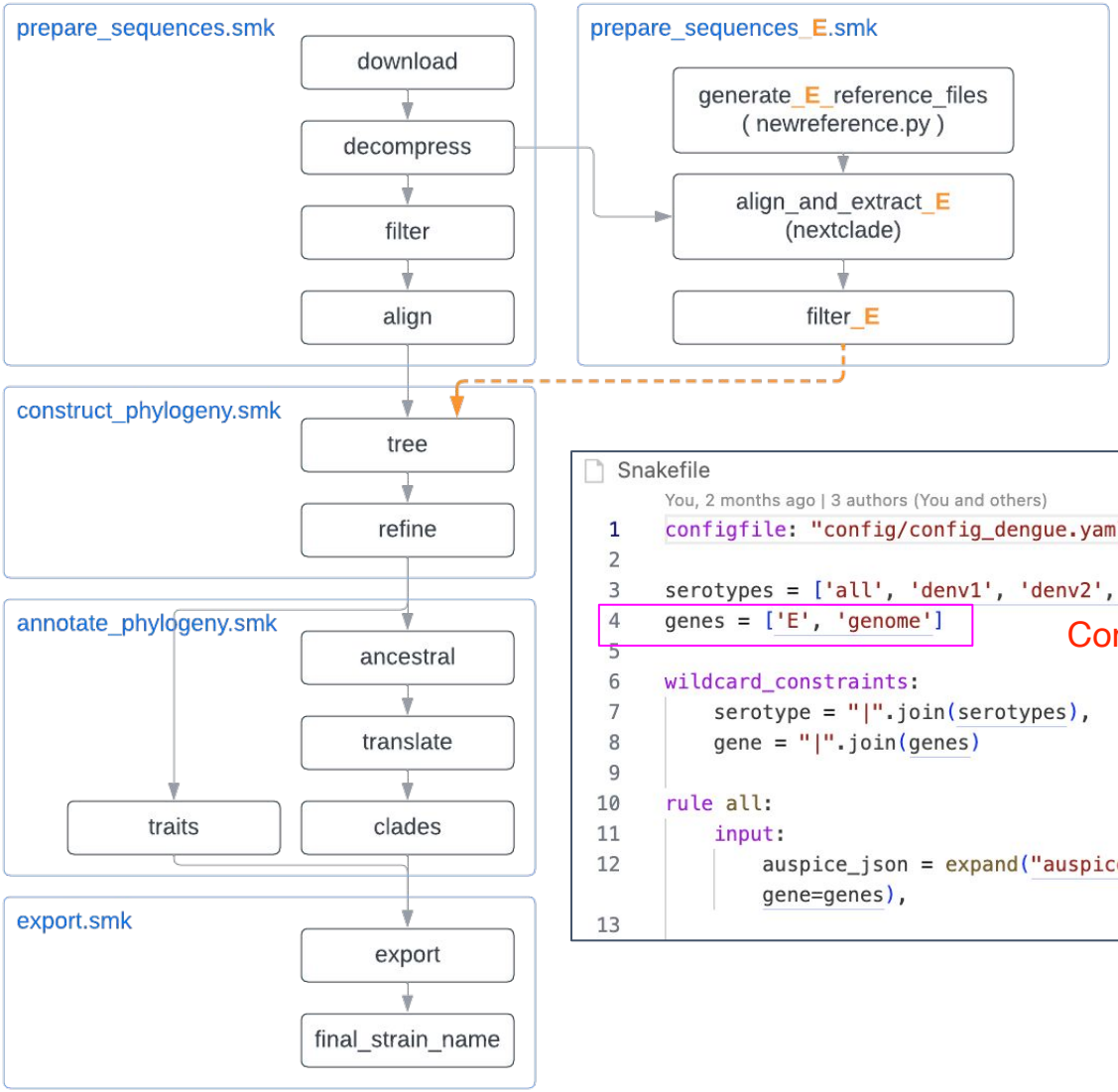
- Connect the aligned and subsampled E gene sequences to the rest of the pipeline
- Requires parameterizing whole "Genome" versus the "E" gene files

Push "E" sequences through pipeline



- Connect the aligned and subsampled E gene sequences to the rest of the pipeline
- Requires parameterizing whole "Genome" versus the "E" gene files
- Can do this with **wildcards**

Sidenote on {wildcards}



Add wildcards to the phylogenetic/Snakefile

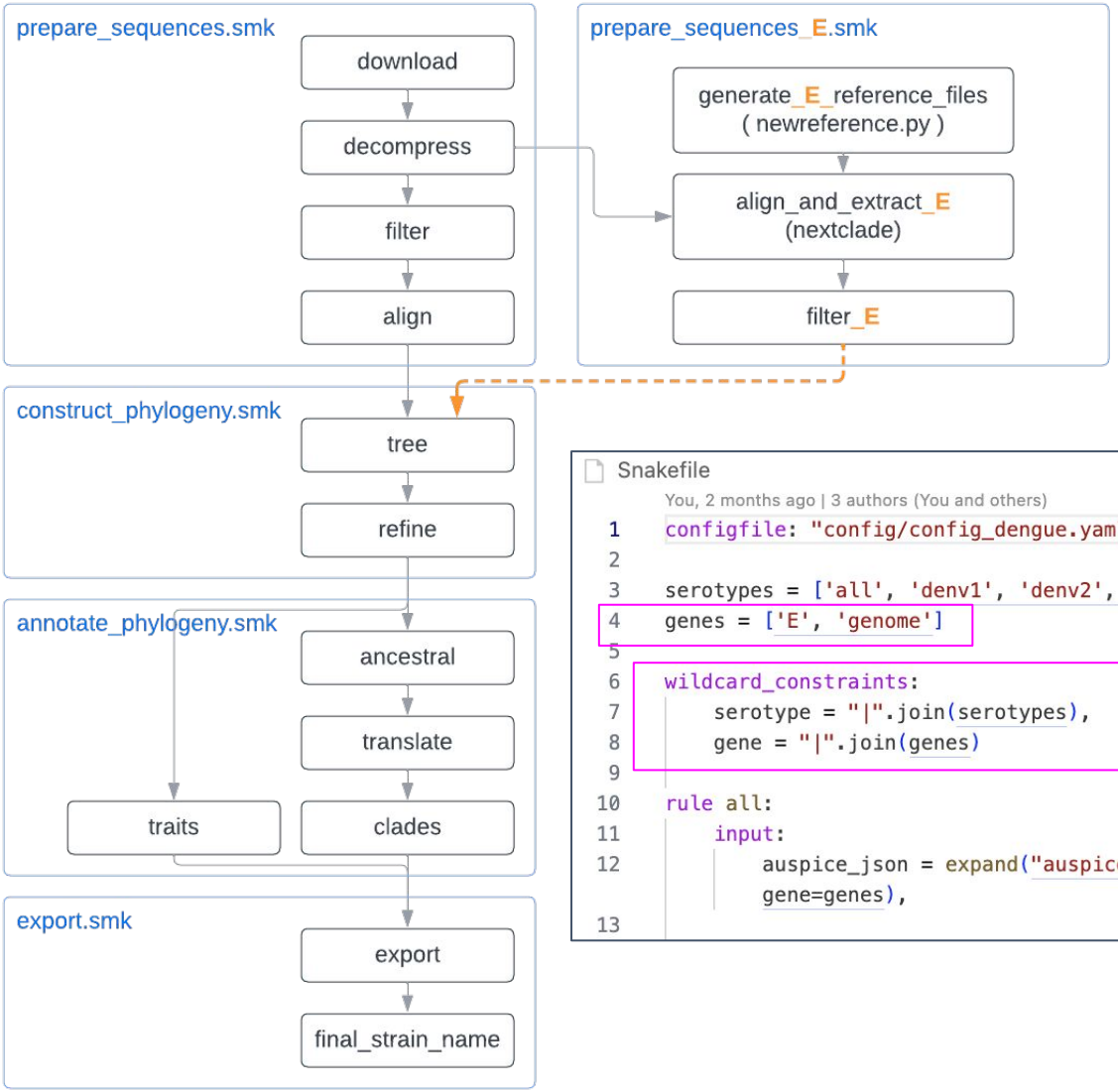
Snakefile

You, 2 months ago | 3 authors (You and others)

```
1  configfile: "config/config_dengue.yaml"      You, 2 months ago • Update the CI workflow …
2
3  serotypes = ['all', 'denv1', 'denv2', 'denv3', 'denv4']
4  genes = ['E', 'genome']
5
6  wildcard_constraints:
7      serotype = "|".join(serotypes),
8      gene = "|".join(genes)
9
10 rule all:
11     input:
12         auspice_json = expand("auspice/dengue_{serotype}_{gene}.json", serotype=serotypes,
           gene=genes),
13
```

Consider moving this to config.yaml

Sidenote on {wildcards}



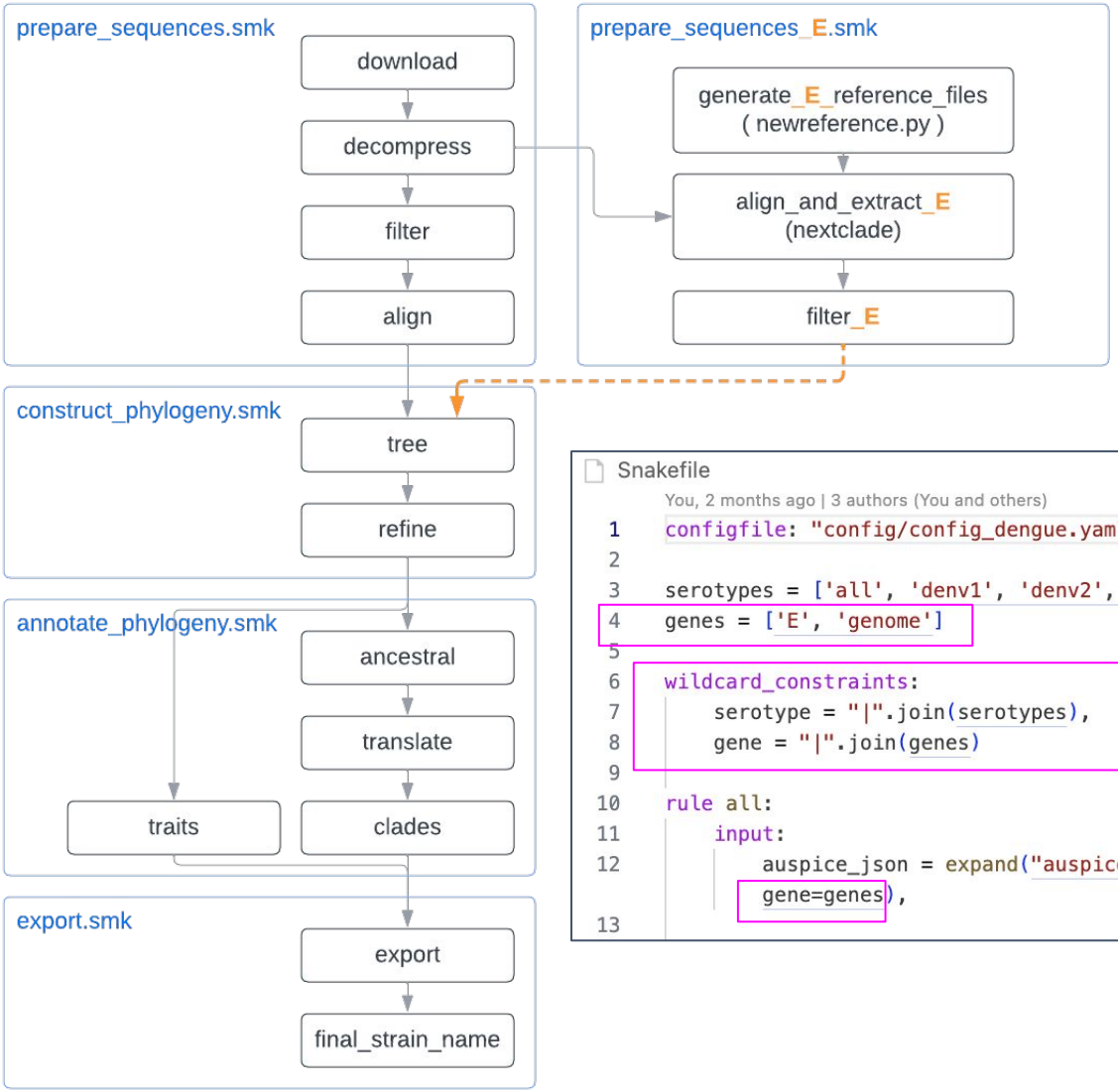
Add wildcards to the phylogenetic/Snakefile

Snakefile

You, 2 months ago | 3 authors (You and others)

```
1  configfile: "config/config_dengue.yaml"      You, 2 months ago • Update the CI workflow …
2
3  serotypes = ['all', 'denv1', 'denv2', 'denv3', 'denv4']
4  genes = ['E', 'genome']
5
6  wildcard_constraints:
7      serotype = "|".join(serotypes),
8      gene = "|".join(genes)
9
10 rule all:
11     input:
12         auspice_json = expand("auspice/dengue_{serotype}_{gene}.json", serotype=serotypes,
           gene=genes),
13
```

Sidenote on {wildcards}



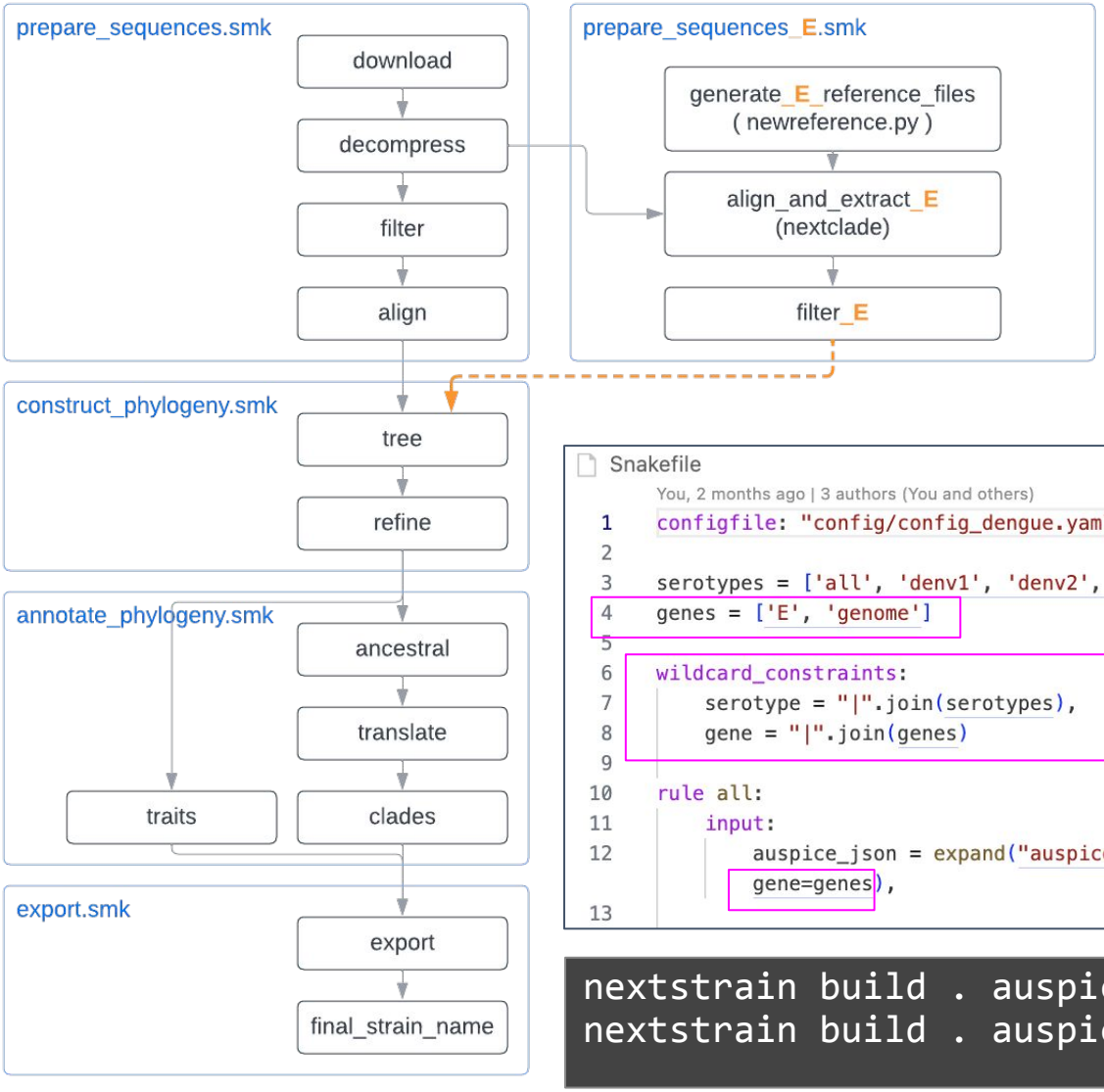
Add wildcards to the phylogenetic/Snakefile

Snakefile

You, 2 months ago | 3 authors (You and others)

```
1  configfile: "config/config_dengue.yaml"      You, 2 months ago • Update the CI workflow …
2
3  serotypes = ['all', 'denv1', 'denv2', 'denv3', 'denv4']
4  genes = ['E', 'genome']
5
6  wildcard_constraints:
7      serotype = "|".join(serotypes),
8      gene = "|".join(genes)
9
10 rule all:
11     input:
12         auspice_json = expand("auspice/dengue_{serotype}_{gene}.json", serotype=serotypes,
           gene=genes),
13
```

Sidenote on {wildcards}



Add wildcards to the phylogenetic/Snakefile

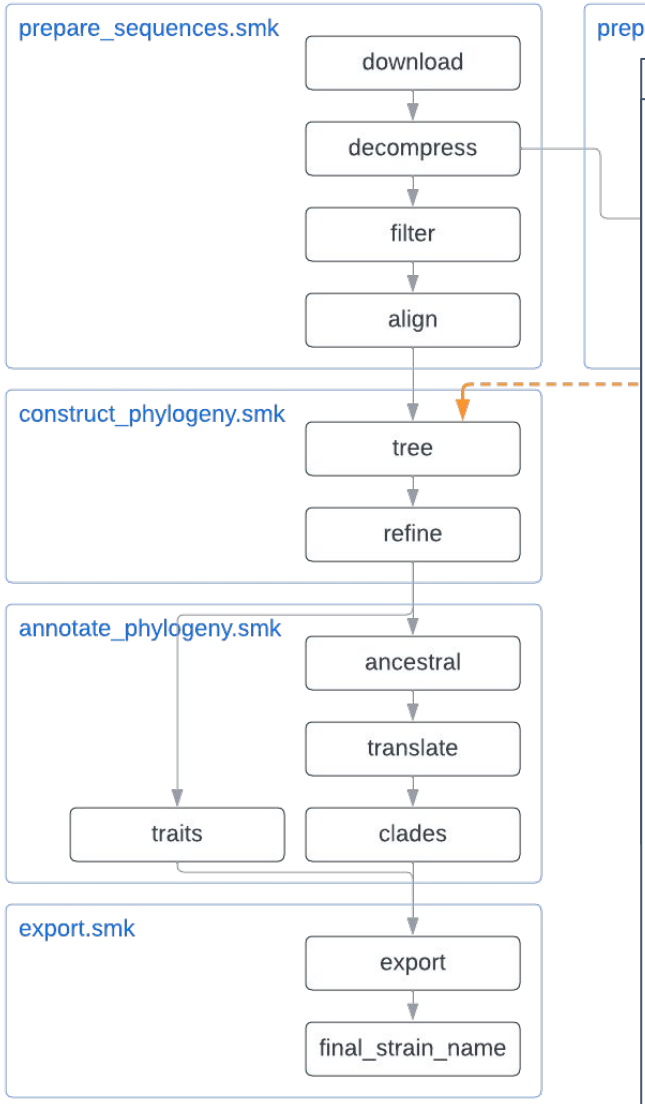
Snakefile

You, 2 months ago | 3 authors (You and others)

```
1  configfile: "config/config_dengue.yaml"      You, 2 months ago • Update the CI workflow …
2
3  serotypes = ['all', 'denv1', 'denv2', 'denv3', 'denv4']
4  genes = ['E', 'genome']
5
6  wildcard_constraints:
7      serotype = "|".join(serotypes),
8      gene = "|".join(genes)
9
10 rule all:
11     input:
12         auspice_json = expand("auspice/dengue_{serotype}_{gene}.json", serotype=serotypes,
           gene=genes),
13
```

```
nextstrain build . auspice/dengue_denv4_genome.json
nextstrain build . auspice/dengue_denv4_E.json
```

Sidenote: use _{gene} wildcards



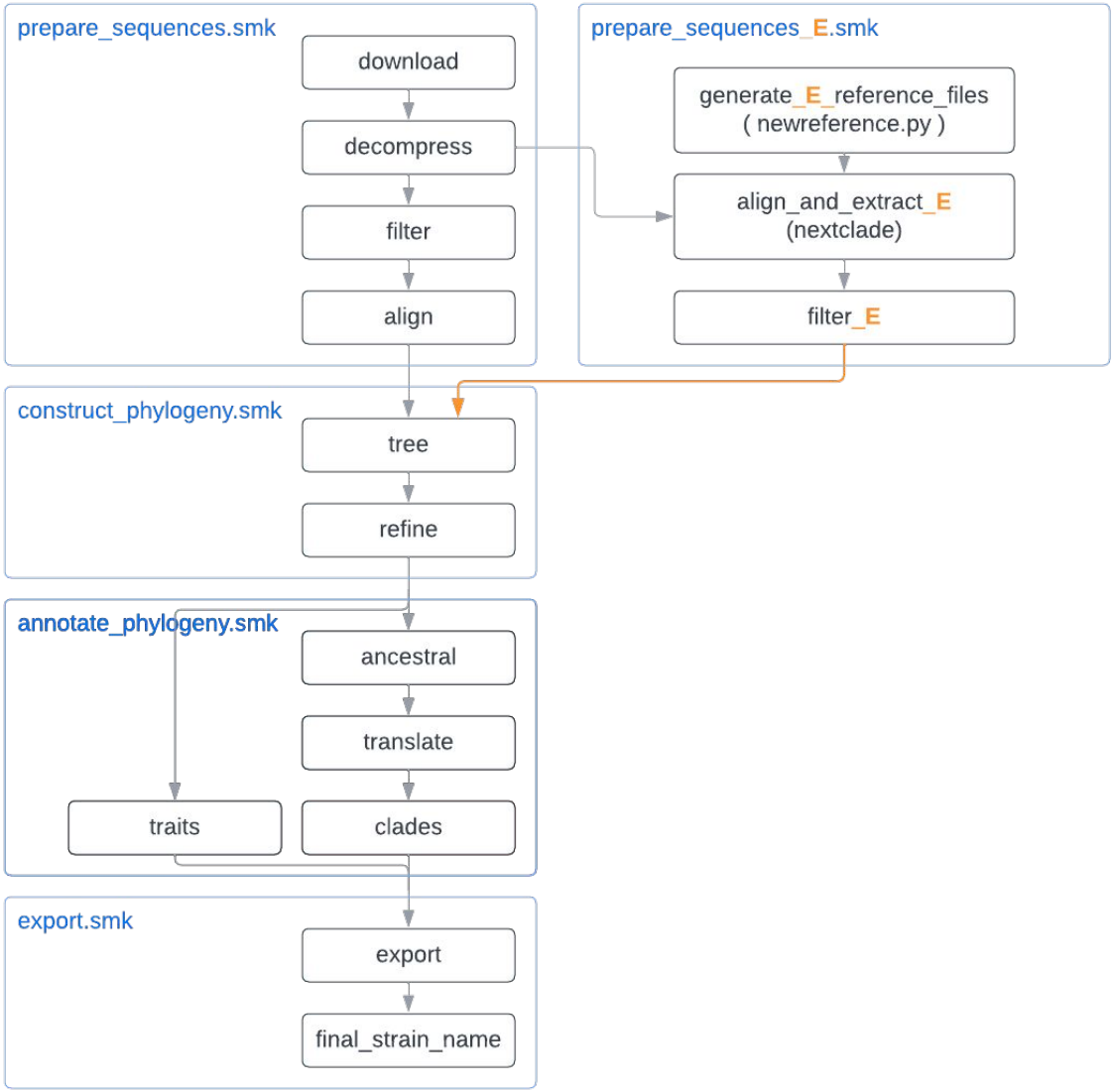
prepare_sequences_E.smk

```
3   serotypes = ['all', 'denv1', 'denv2', 'denv3', 'denv4']
4   genes = ['E', 'genome']
5
6   wildcard_constraints:
7       serotype = "|".join(serotypes),
8       gene = "|".join(genes)
```

rules > construct_phylogeny.smk

```
15  rule tree:
16      """Building tree"""
17      input:
18          alignment = "results/aligned_{serotype}_{gene}.fasta"
19      output:
20          tree = "results/tree-raw_{serotype}_{gene}.nwk"
21      shell:
22          """
23          augur tree \
24              --alignment {input.alignment} \
25              --output {output.tree} \
26              --nthreads 1
27          """
28
29  rule refine:
30      """
31      Refining tree
32        - estimate timetree
33        - use {params.coalescent} coalescent timescale
34        - estimate {params.date_inference} node dates
35        - filter tips more than {params.clock_filter_iqd} IQDs from clock expectation
36      """
37      input:
38          tree = "results/tree-raw_{serotype}_{gene}.nwk",
39          alignment = "results/aligned_{serotype}_{gene}.fasta",
40          metadata = "data/metadata_{serotype}.tsv"
41      output:
42          tree = "results/tree_{serotype}_{gene}.nwk",
43          node_data = "results/branch-lengths_{serotype}_{gene}.json",
44      params:
45          coalescent = "const",
46          date_inference = "marginal",
47          clock_filter_iqd = 4,
```

Connected! Try running the pipeline



Complications with "clades"

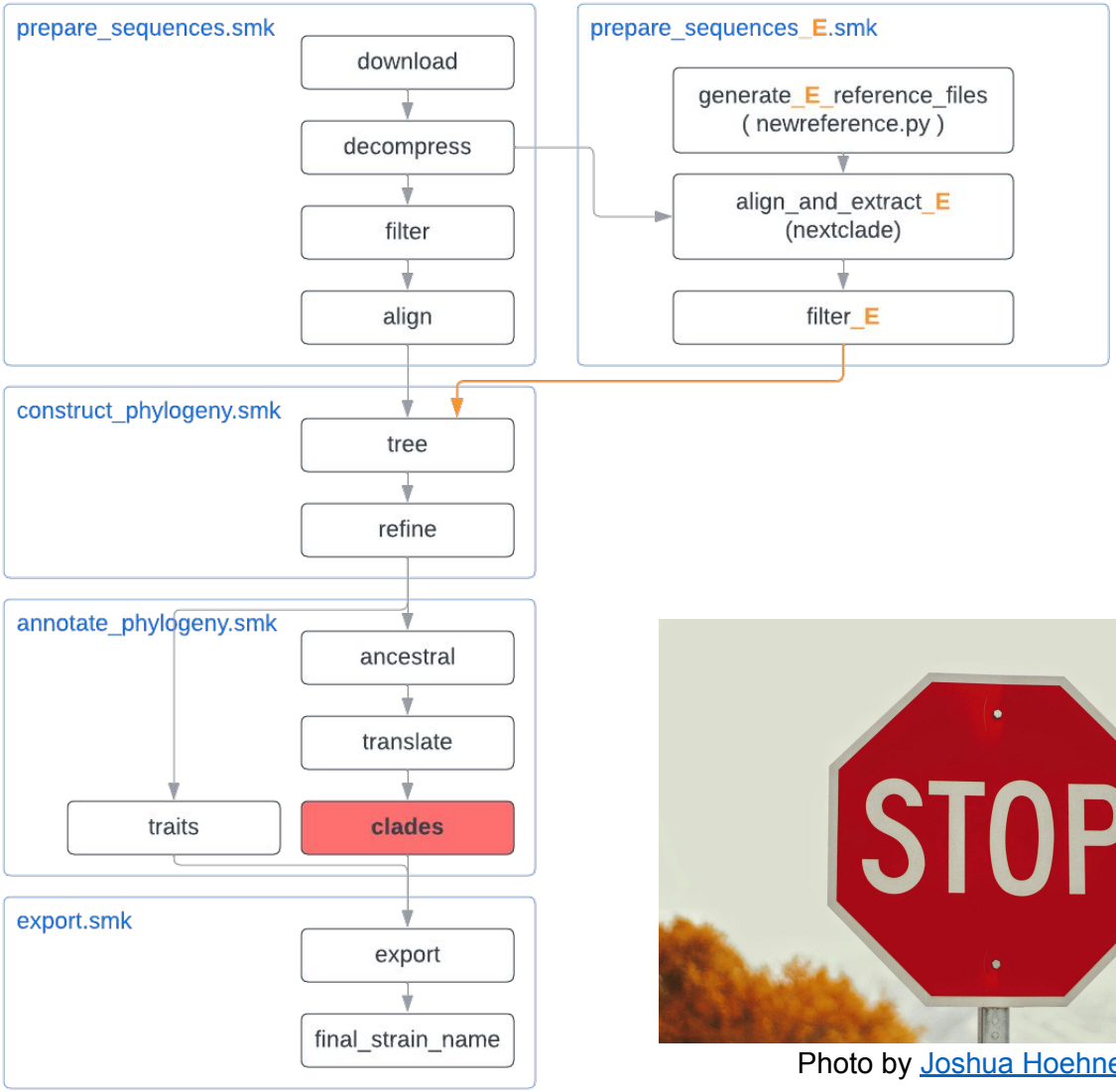
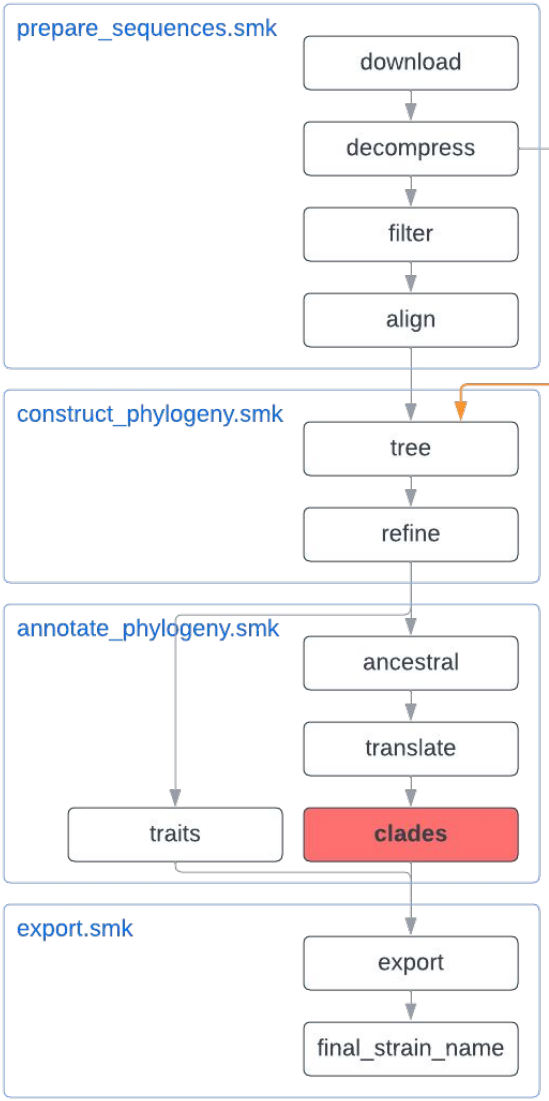


Photo by Joshua Hoehne on Unsplash

Complications with "clades"

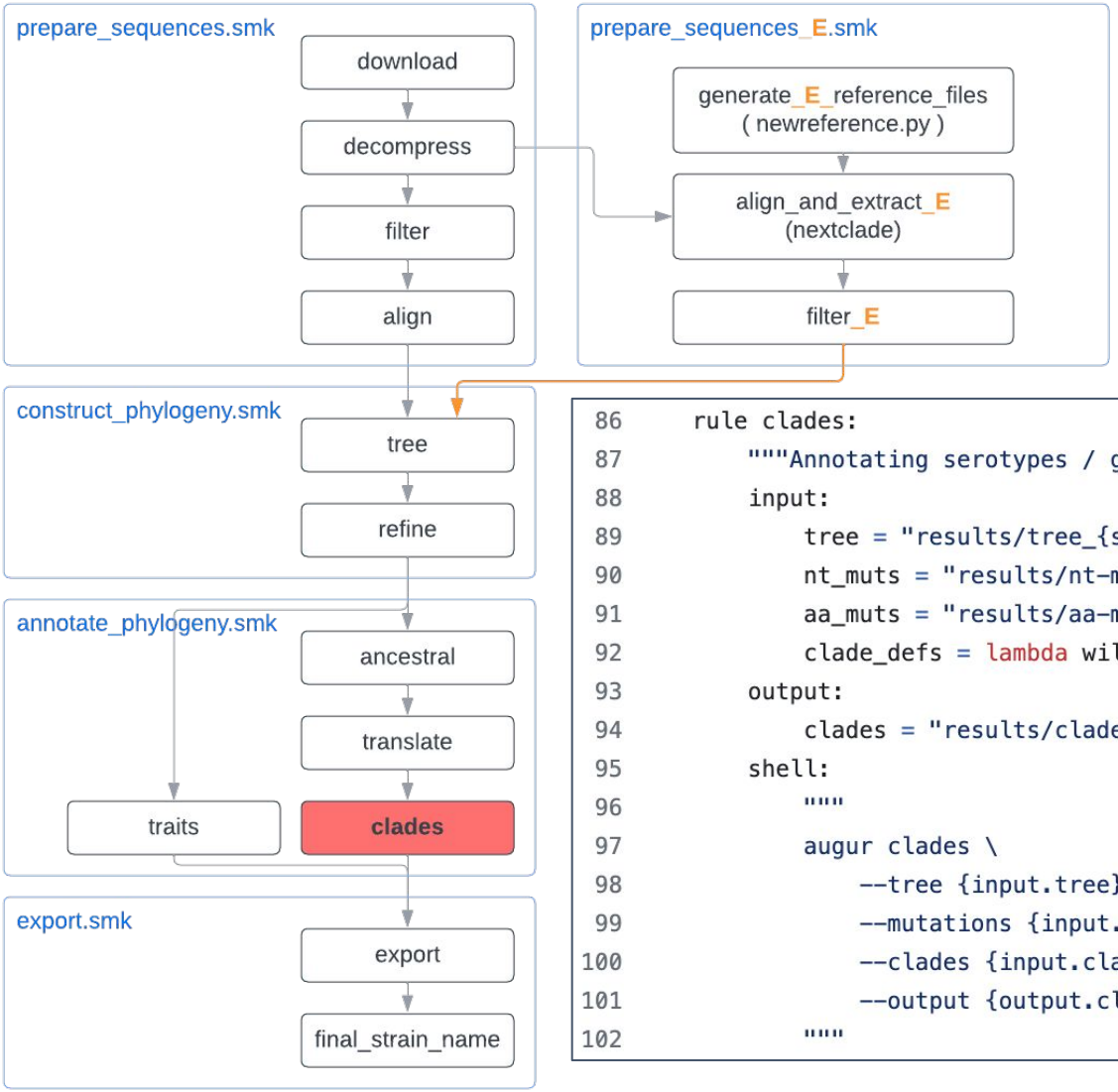



```
augur clades \
   --tree {input.tree} \
   --mutations {input.nt_muts} {input.aa_muts} \
   --clades clades_genotypes.tsv \
   --output {output.clades}
```

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	clade	gene	site	alt		clade	gene	site	alt		clade	gene	site	alt
2	DENV1/I	E	461	V		DENV2/AM	E	71	D		DENV3/I	M	128	F
3	DENV1/I	E	484	L		DENV2/AM	E	81	T		DENV3/I	E	68	V
4	DENV1/I	M	122	R		DENV2/AM	E	129	I		DENV3/II	M	57	A
5	DENV1/II	E	345	A		DENV2/AM	NS1	21	V		DENV3/II	NS5	749	K
6	DENV1/II	E	432	M		DENV2/AM	NS1	73	S		DENV3/III	E	132	Y
7	DENV1/III	E	297	V		DENV2/AM	NS1	99	V		DENV3/III	E	301	T
8	DENV1/III	M	118	R		DENV2/AM	NS1	170	R		DENV3/IV	NS1	139	S
9	DENV1/IV	E	339	S		DENV2/AA	E	491	A		DENV3/IV	NS5	638	P
10	DENV1/IV	M	72	E		DENV2/AA	M	15	G					
11	DENV1/IV	E	88	T		DENV2/AA	M	39	I		clade	gene	site	alt
12	DENV1/V	NS1	324	R		DENV2/AI	E	484	I		DENV4/I	E	429	L
13	DENV1/V	NS2A	142	P		DENV2/AI	NS1	222	N		DENV4/I	NS1	98	S
14	DENV1/V	NS3	185	K		DENV2/AI	NS5	687	I		DENV4/II	E	265	A
15	DENV1/V	NS5	834	E		DENV2/AII	NS1	51	Q		DENV4/II	E	46	T
16						DENV2/AII	NS2A	142	R		DENV4/II	NS1	246	S
17						DENV2/AII	NS3	160	S		DENV4/S	E	132	V
18						DENV2/C	E	71	A		DENV4/S	E	154	S
19						DENV2/C	E	149	N		DENV4/S	E	162	T
20						DENV2/C	E	462	V					
21						DENV2/S	E	59	F					
22						DENV2/S	E	236	M					
23						DENV2/S	E	432	V					
24														

Check what happens when aligning “E” gene to “WGS” Nextclade dataset
“DENV2/AII” may be classified up the branch as “DENV2/AI”

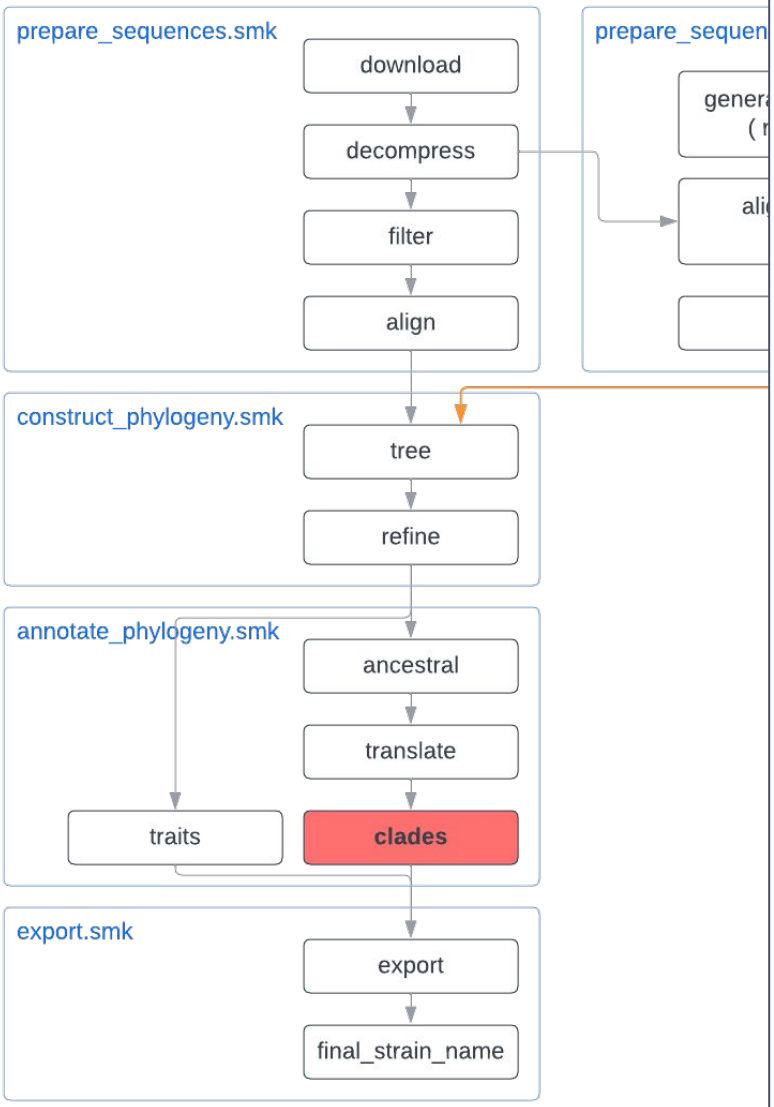
fix: (1/3) Only run clades for "_genome"



Consider dropping the augur clades call in the phylogenetic workflow, while "augur clades" is still needed in the Nextclade workflow.

```
86    rule clades:
87        """Annotating serotypes / genotypes"""
88        input:
89            tree = "results/tree_{serotype}_genome.nwk",
90            nt_muts = "results/nt-muts_{serotype}_genome.json",
91            aa_muts = "results/aa-muts_{serotype}_genome.json",
92            clade_defs = lambda wildcards: config['clades']['clade_definitions'][wildc
93        output:
94            clades = "results/clades_{serotype}_genome.json"
95        shell:
96            """
97            augur clades \
98                --tree {input.tree} \
99                --mutations {input.nt_muts} {input.aa_muts} \
100               --clades {input.clade_defs} \
101               --output {output.clades}
102           """
```

fix: (2/3) Rely on "Nextclade" metadata



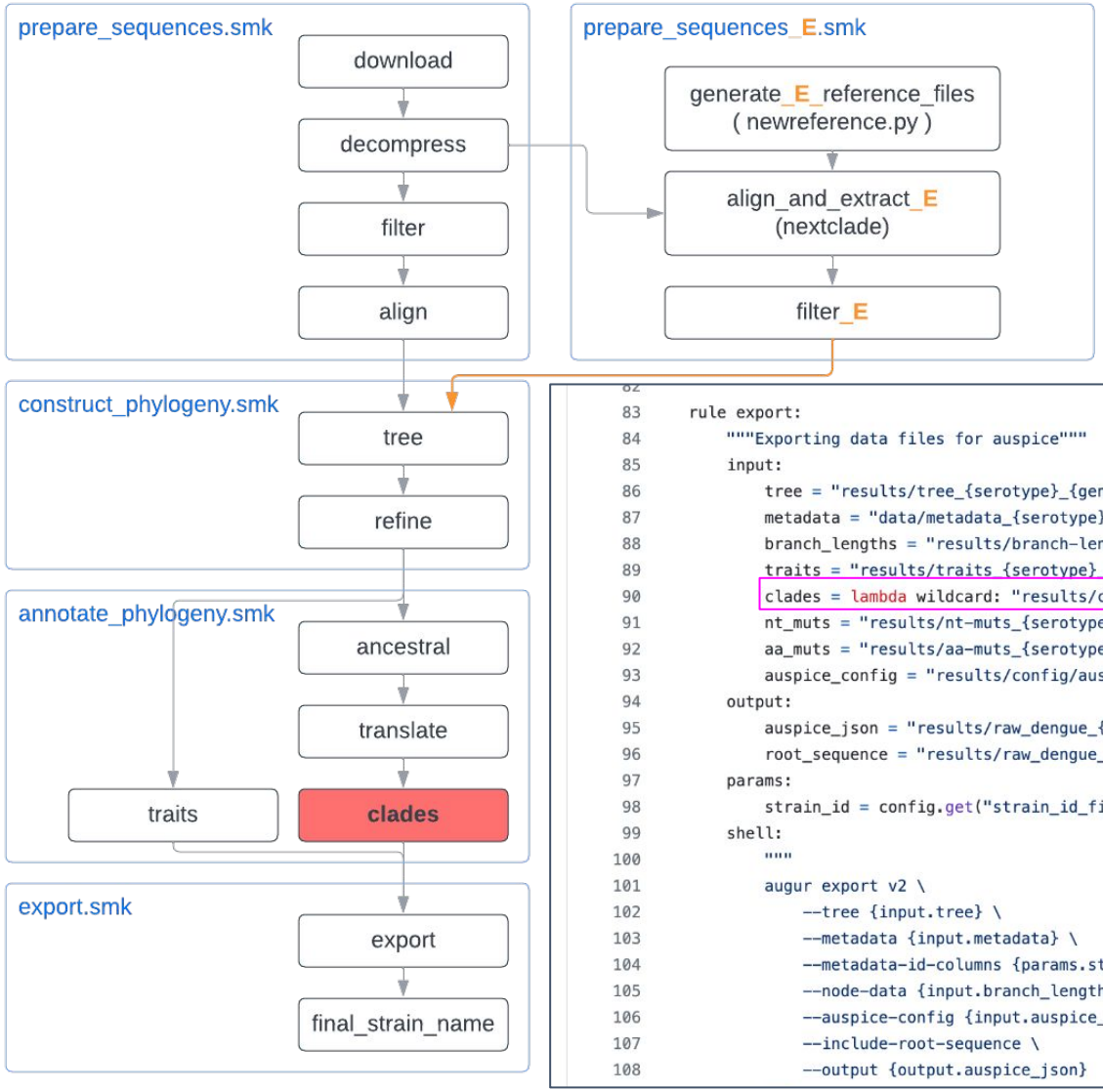
dengue / phylogenetic / config / **config_dengue.yaml**

Code | Blame | 43 lines (40 loc) · 1.11 KB

```
10      sequences_per_group:
11        all: '10'
12        denv1: '36'
13        denv2: '36'
14        denv3: '36'
15        denv4: '36'
16      E_root_sequence:
17        all: 'NC_002640'
18        denv1: 'NC_001477'
19        denv2: 'NC_001474'
20        denv3: 'NC_001475'
21        denv4: 'NC_002640'
22
23    traits:
24      sampling_bias_correction: '3'
25      traits_columns:
26        all_genome: 'region'
27        denv1_genome: 'country region'
28        denv2_genome: 'country region'
29        denv3_genome: 'country region'
30        denv4_genome: 'country region'
31        all_E: 'region nextclade_type'
32        denv1_E: 'country region nextclade_subtype'
33        denv2_E: 'country region nextclade_subtype'
34        denv3_E: 'country region nextclade_subtype'
35        denv4_E: 'country region nextclade_subtype'
```

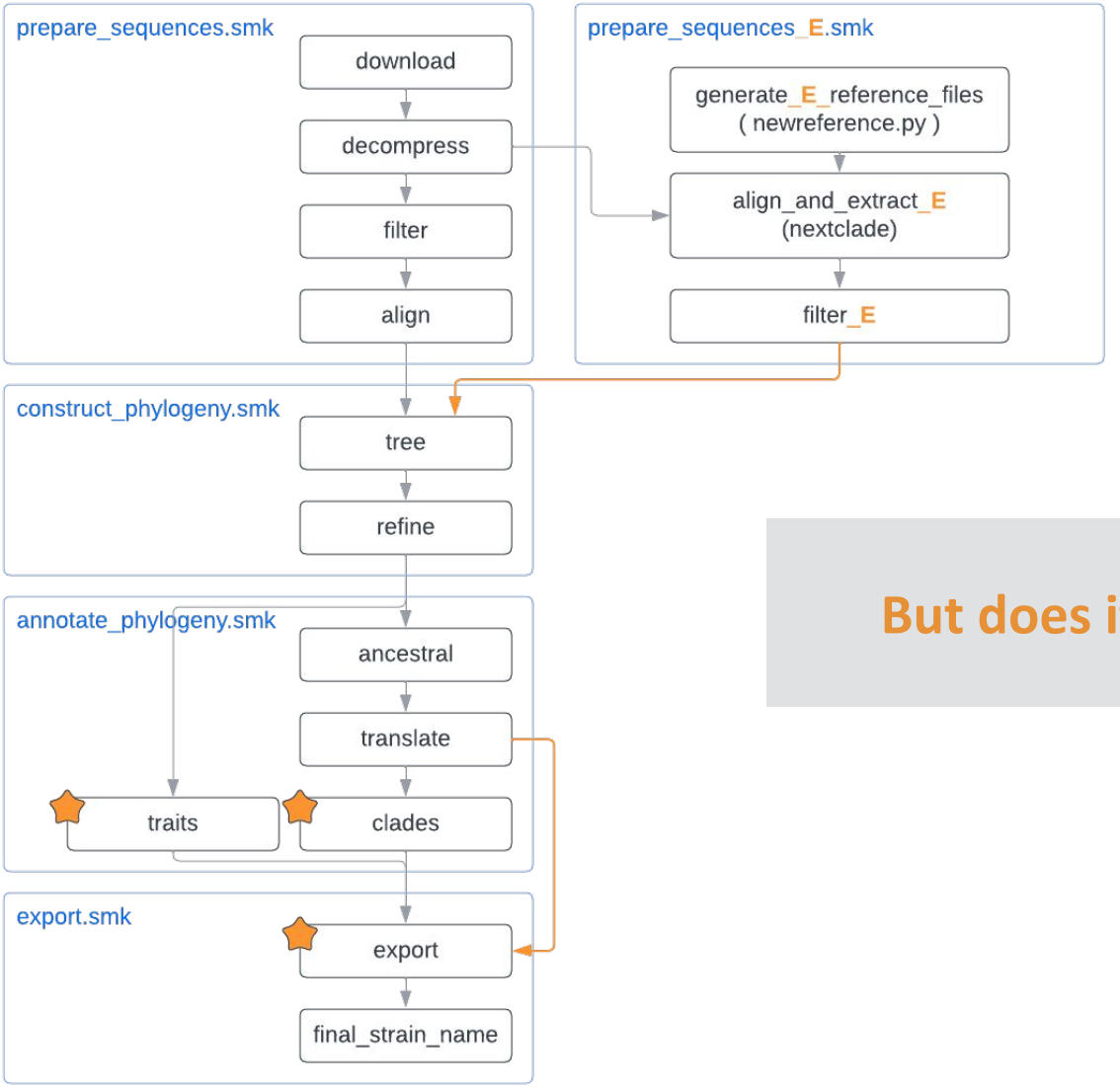
Consider using "nextclade_subtype" for both genome and E gene trees

fix: (3/3) Add if/else in "export" rule




```
83  rule export:
84      """Exporting data files for auspice"""
85      input:
86          tree = "results/tree_{serotype}_{gene}.nwk",
87          metadata = "data/metadata_{serotype}.tsv",
88          branch_lengths = "results/branch-lengths_{serotype}_{gene}.json",
89          traits = "results/traits_{serotype}_{gene}.json",
90          clades = lambda wildcard: "results/clades_{serotype}_{gene}.json" if wildcard.gene in ['genome'] else [],
91          nt_muts = "results/nt-muts_{serotype}_{gene}.json",
92          aa_muts = "results/aa-muts_{serotype}_{gene}.json",
93          auspice_config = "results/config/auspice_config_{serotype}_{gene}.json",
94      output:
95          auspice_json = "results/raw_dengue_{serotype}_{gene}.json",
96          root_sequence = "results/raw_dengue_{serotype}_{gene}_root-sequence.json",
97      params:
98          strain_id = config.get("strain_id_field", "strain"),
99      shell:
100         """
101         augur export v2 \
102             --tree {input.tree} \
103             --metadata {input.metadata} \
104             --metadata-id-columns {params.strain_id} \
105             --node-data {input.branch_lengths} {input.traits} {input.clades} {input.nt_muts} {input.aa_muts} \
106             --auspice-config {input.auspice_config} \
107             --include-root-sequence \
108             --output {output.auspice_json}
```

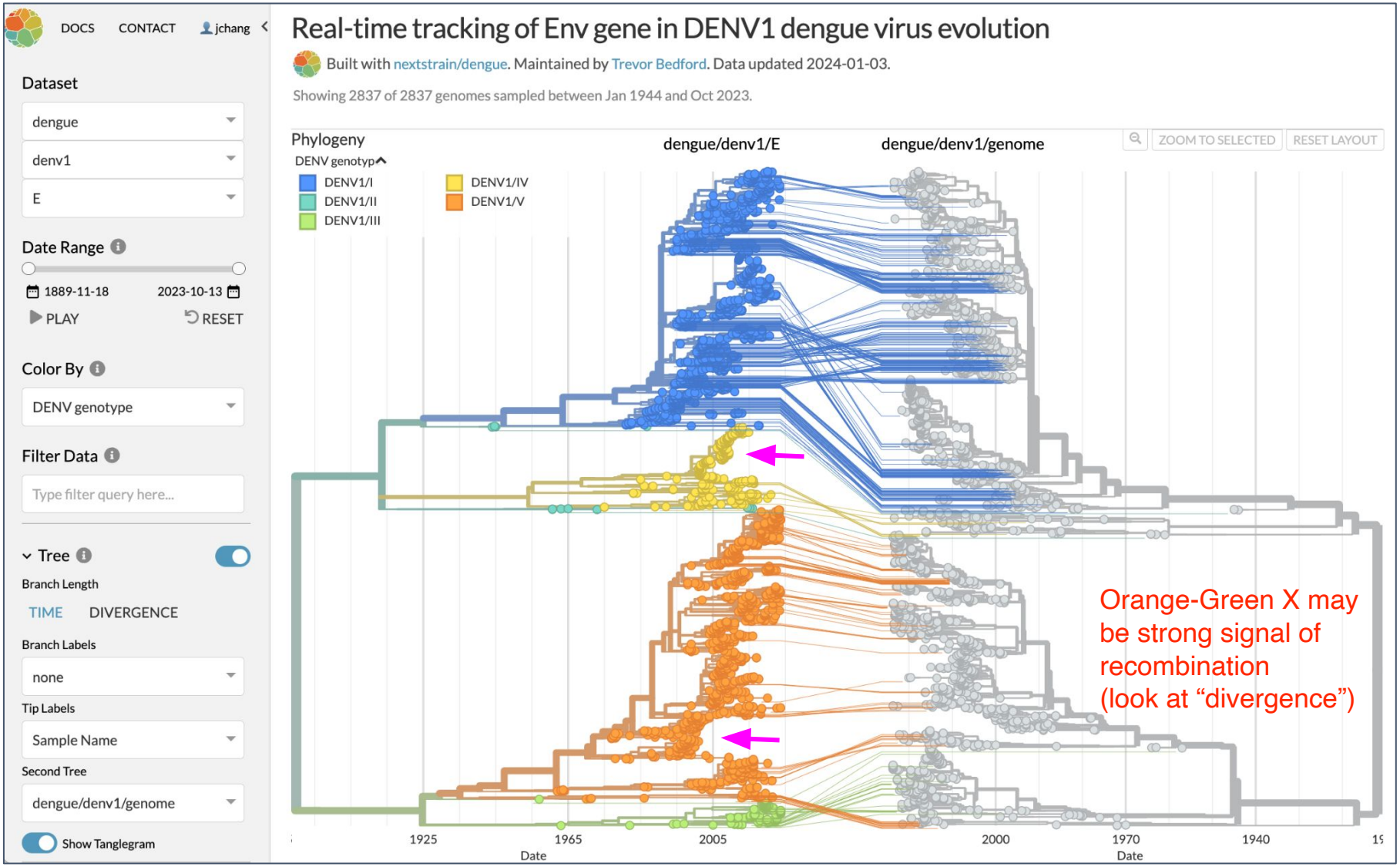
Dengue pipeline (with "E" gene)



But does it work?

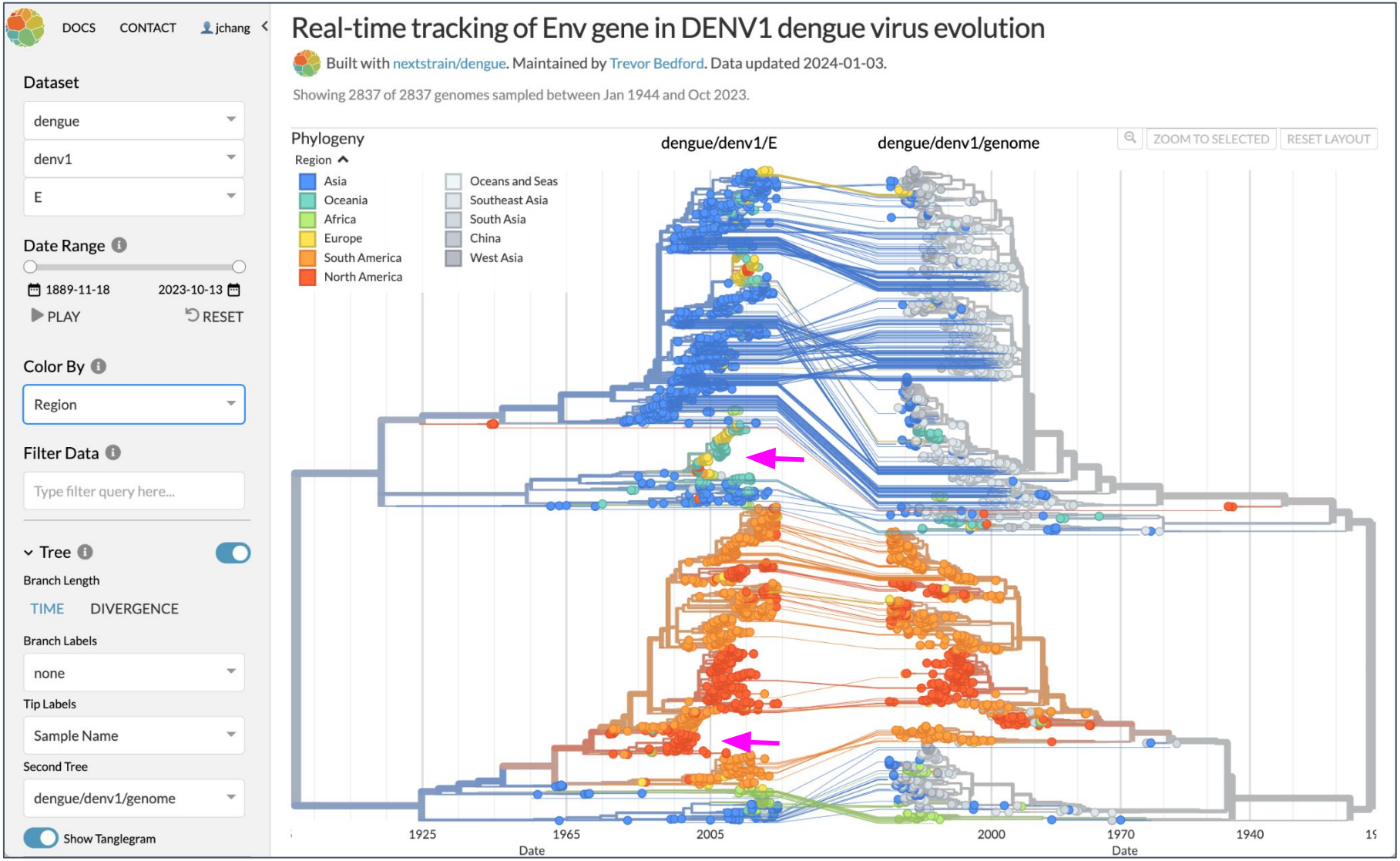
"E" gene trees

https://next.nextstrain.org/dengue/denv1/E:dengue/denv1/genome



"E" gene trees

https://next.nextstrain.org/dengue/denv1/E:dengue/denv1/genome



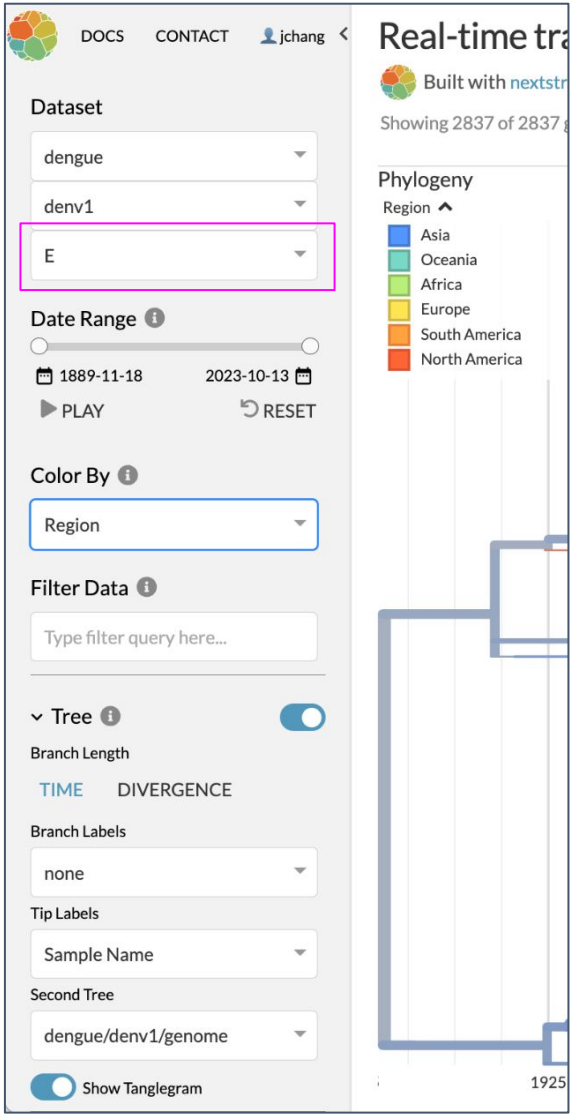
Outline

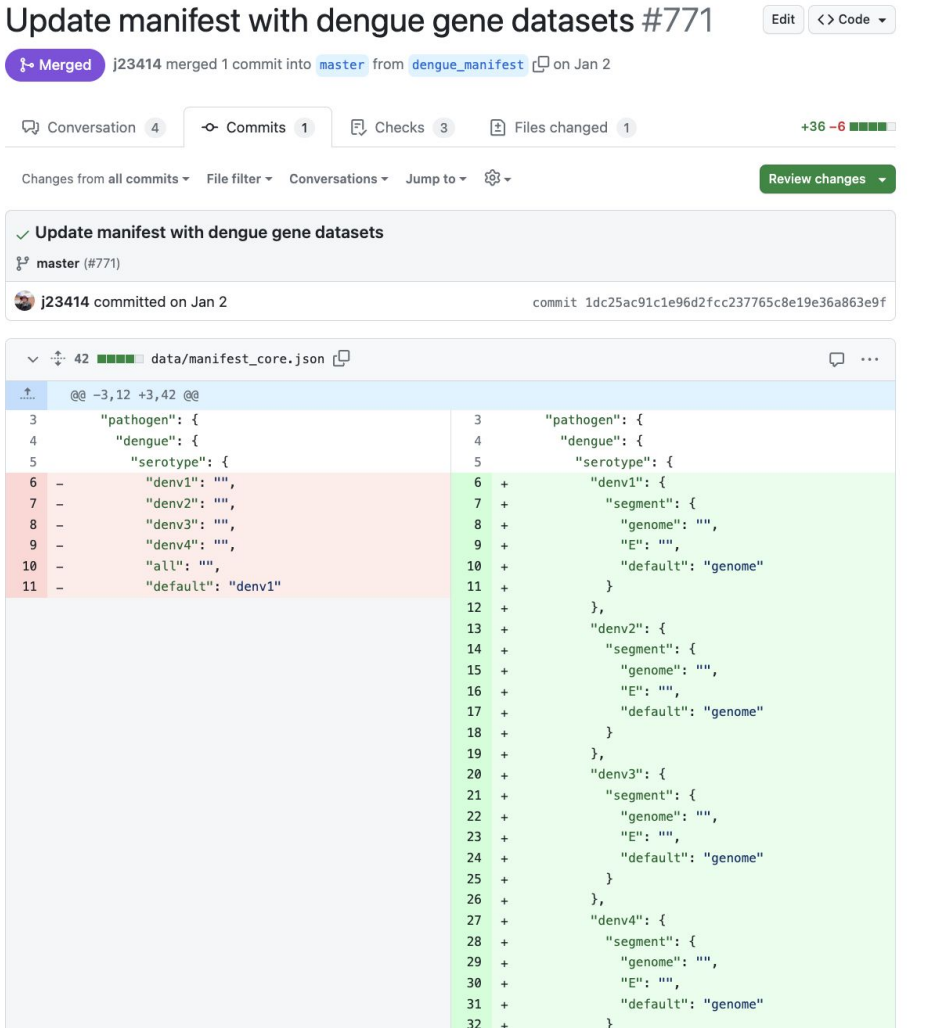
- Motivation
  - Dec 19, 2023 request for "E" gene trees
  - Surface the problem on slack and github to start the conversation
- Overview of modifying the pipeline for "E" gene trees
  - Use newreference.py
  - Use Nextclade or Augur align
  - Use wildcards to parameterize "gene" vs "genome"
  - A work-around for "augur clades" if gene not in the "clades.tsv"
- **Pushing to the live site and future directions**

Outline

- Motivation
  - Dec 19, 2023 request for "E" gene trees
  - Surface the problem on slack and github to start the conversation
- Overview of modifying the pipeline for "E" gene trees
  - Use newreference.py
  - Use Nextclade or Augur align
  - Use wildcards to parameterize "gene" vs "genome"
  - A work-around for "augur clades" if gene not in the "clades.tsv"
- **Pushing to the live site and future directions**
  - Change the Manifest

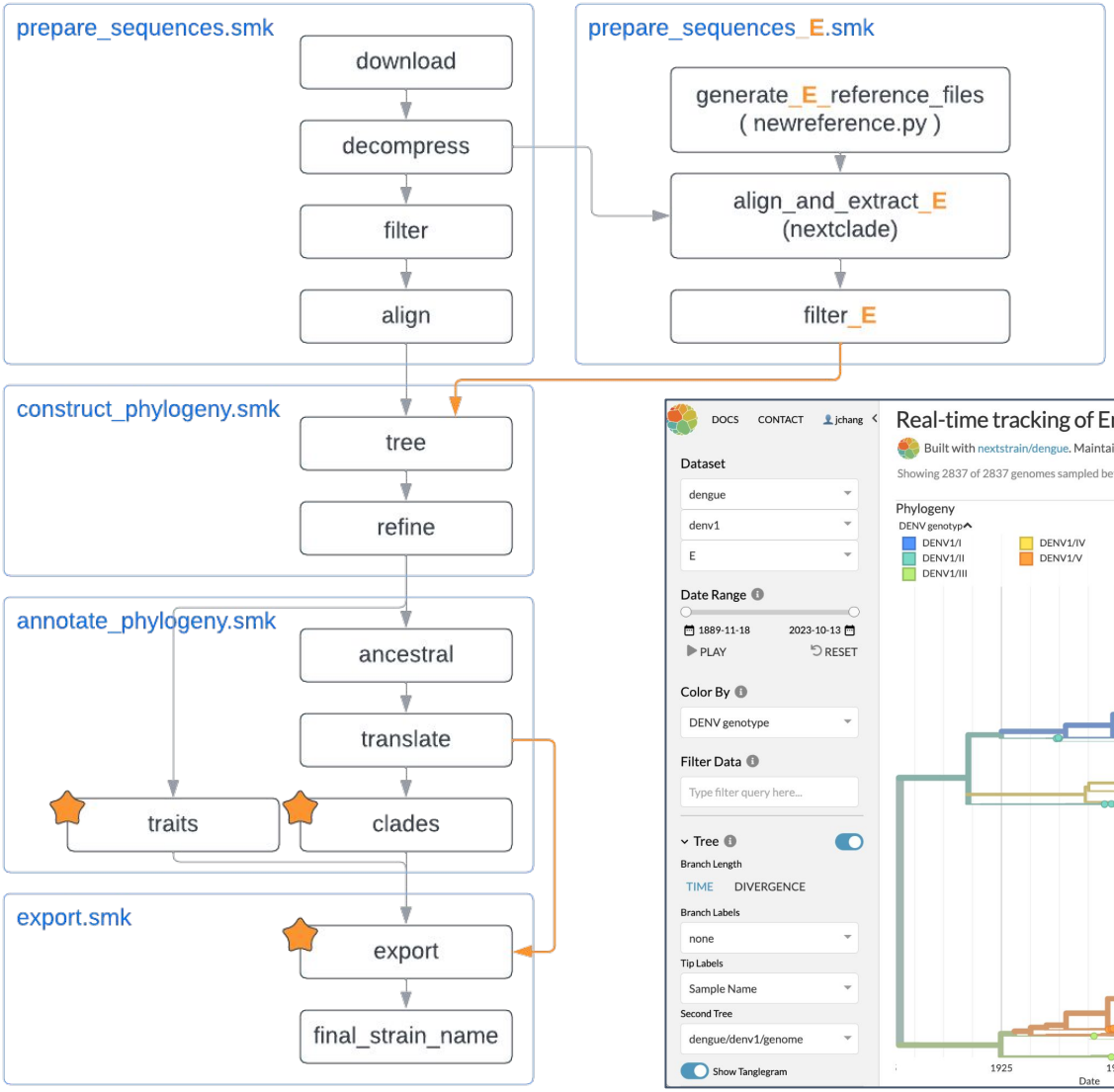
Update "manifest" on nextstrain.org

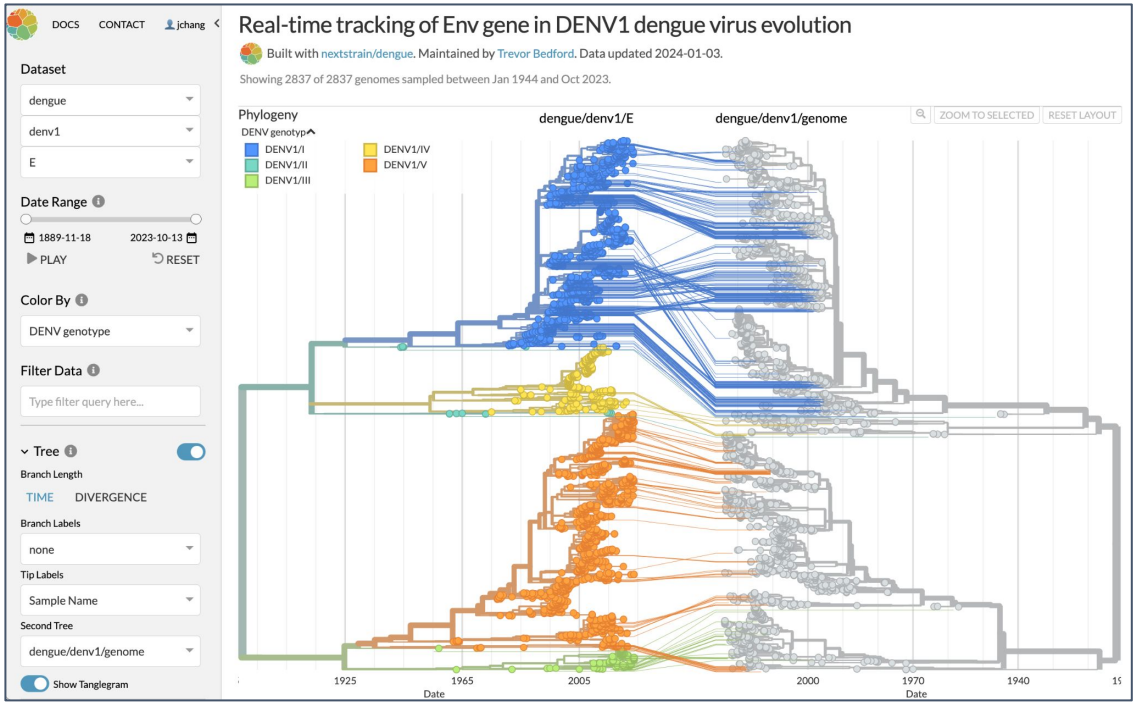




https://github.com/nextstrain/nextstrain.org/pull/771

In summary - the gene pipeline





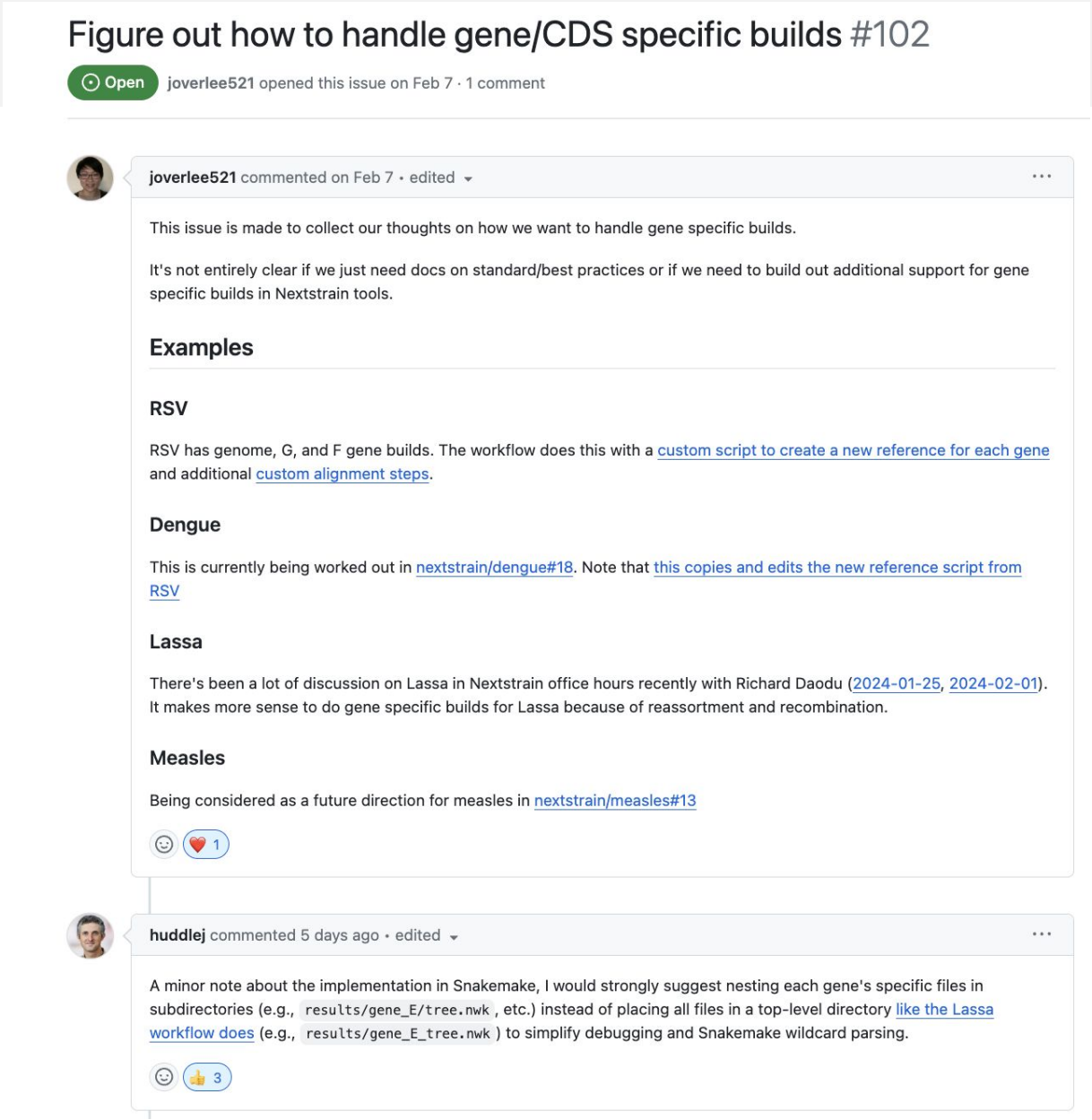


Figure out how to handle gene/CDS specific builds #102

Open joverlee521 opened this issue on Feb 7 · 1 comment

joverlee521 commented on Feb 7 • edited

This issue is made to collect our thoughts on how we want to handle gene specific builds.

It's not entirely clear if we just need docs on standard/best practices or if we need to build out additional support for gene specific builds in Nextstrain tools.

Examples

RSV

RSV has genome, G, and F gene builds. The workflow does this with a custom script to create a new reference for each gene and additional custom alignment steps.

Dengue

This is currently being worked out in nextstrain/dengue#18. Note that this copies and edits the new reference script from RSV

Lassa

There's been a lot of discussion on Lassa in Nextstrain office hours recently with Richard Daodu (2024-01-25, 2024-02-01). It makes more sense to do gene specific builds for Lassa because of reassortment and recombination.

Measles

Being considered as a future direction for measles in nextstrain/measles#13

❤️ 1

huddlej commented 5 days ago • edited

A minor note about the implementation in Snakemake, I would strongly suggest nesting each gene's specific files in subdirectories (e.g., `results/gene_E/tree.nwk`, etc.) instead of placing all files in a top-level directory like the Lassa workflow does (e.g., `results/gene_E_tree.nwk`) to simplify debugging and Snakemake wildcard parsing.

👍 3

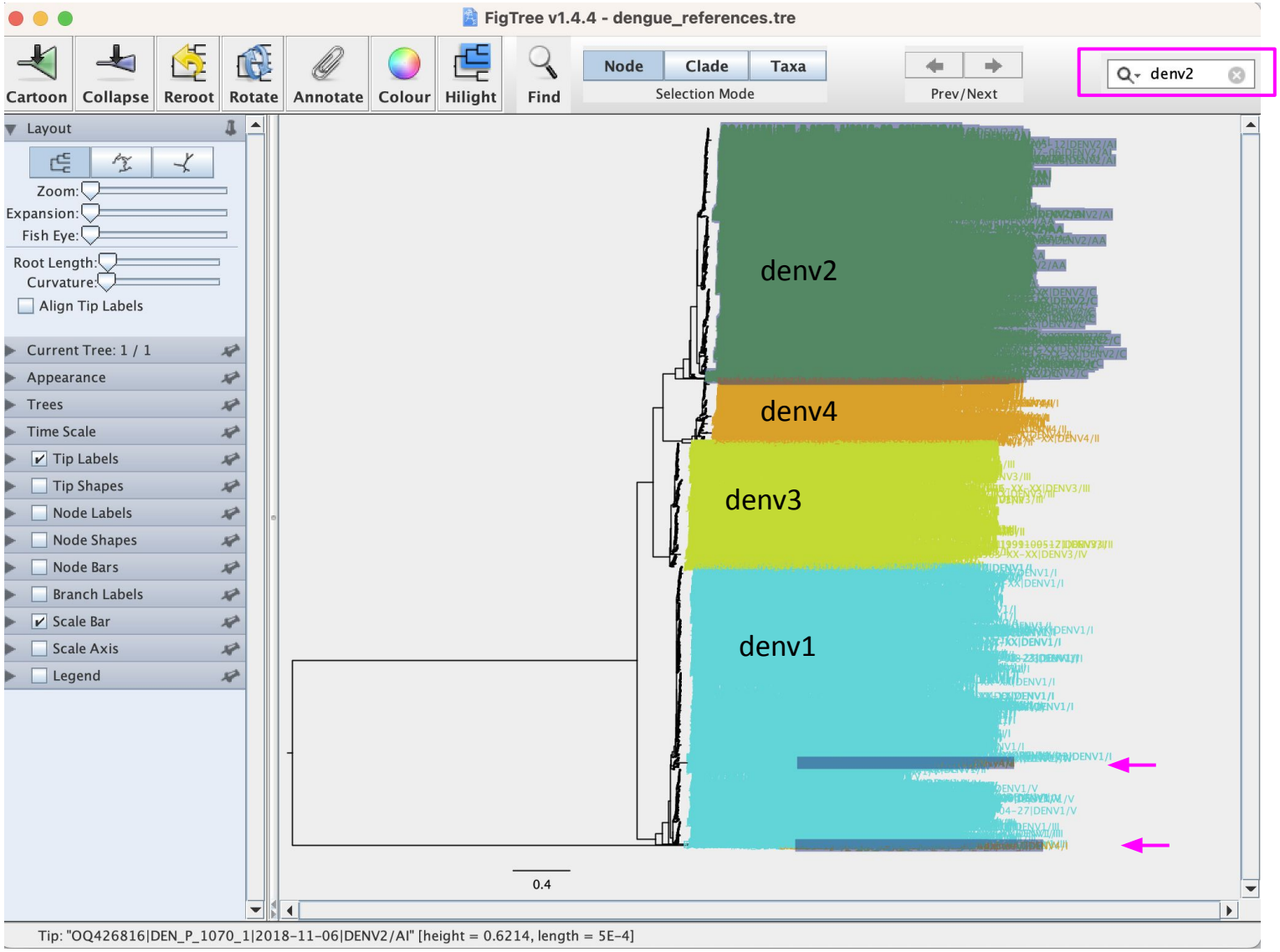
Can continue the discussion on a GitHub issue:

https://github.com/nextstrain/private/issues/102

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.

Still working on Nextclade assignment



denv2?