

# 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

# viruses      Dec 19, 2023

Ammar Aziz and you

 **Ammar Aziz** 3 months ago  
Hi @Jennifer Chang 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.

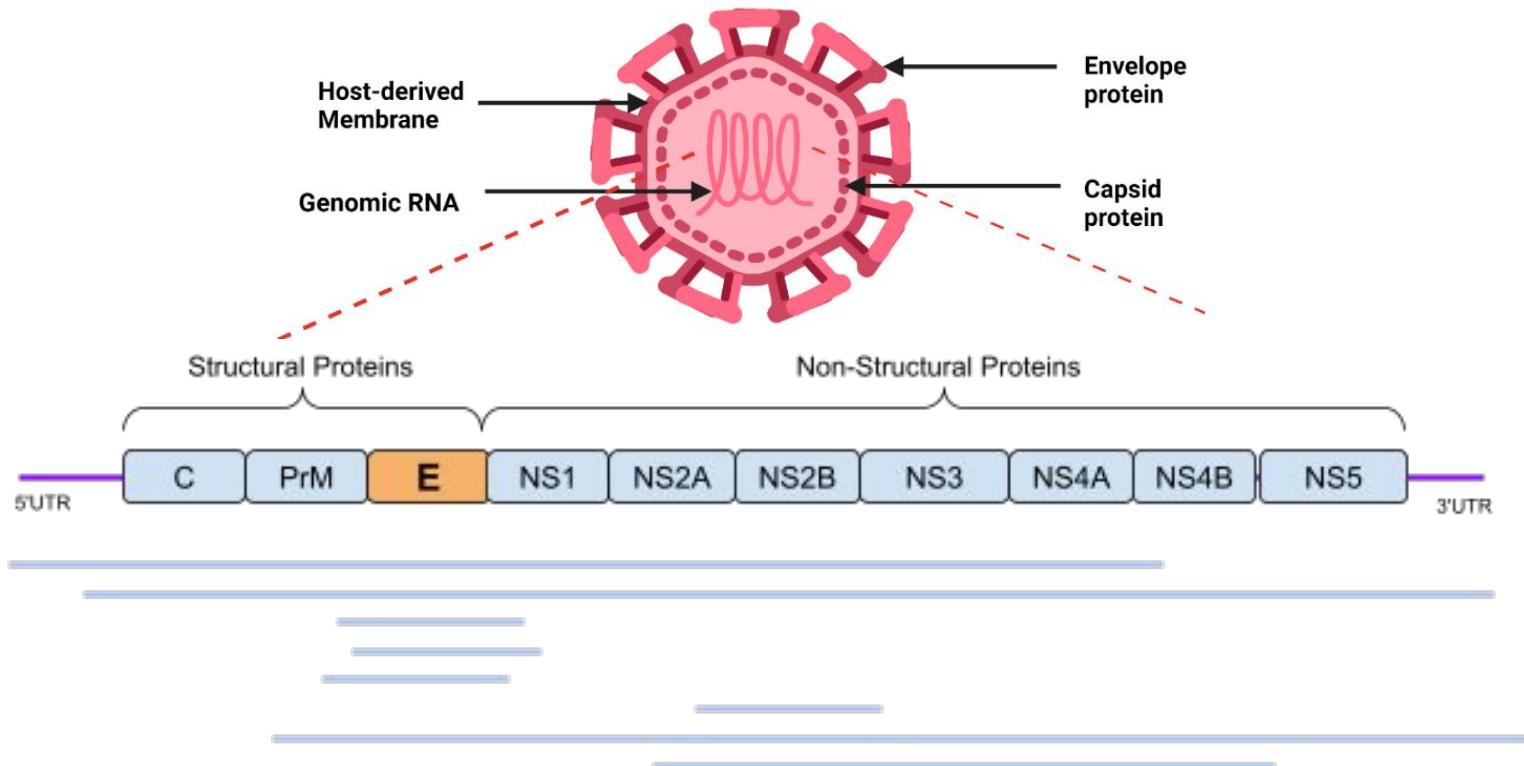
 **Jennifer Chang** 3 months ago  
Thanks for bringing this up! I'll look into getting E gene builds and perhaps get in touch. The current site is split by serotype (all, denv1-denv4), I assume we'd aim at each one having a E gene build.

 **Ammar Aziz** 3 months ago  
Yeah ideally for each serotype have an Egene build.

 **Ammar Aziz** 3 months ago  
In our part of the world the majority of countries (pacific/oceania) sequence the E gene. So while it looks like there's nothing in that region, there's some surveillance happening.

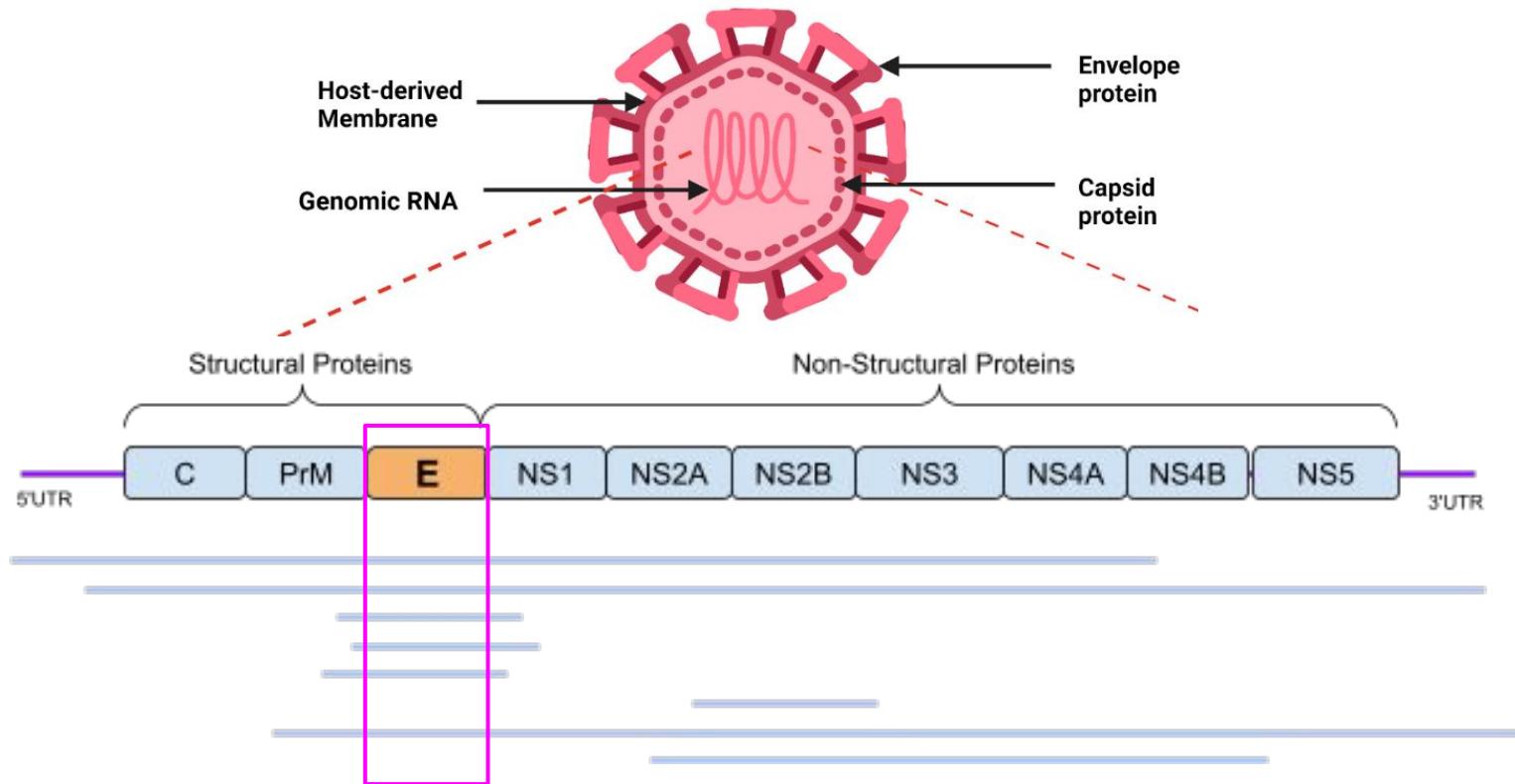
 **Ammar Aziz** 3 months ago  
I have to confess selfishness here, I want to point collaborators to nextstrain when they ask questions about transmission between neighboring countries. With an E gene build I can do just that!

# Dengue virus genome



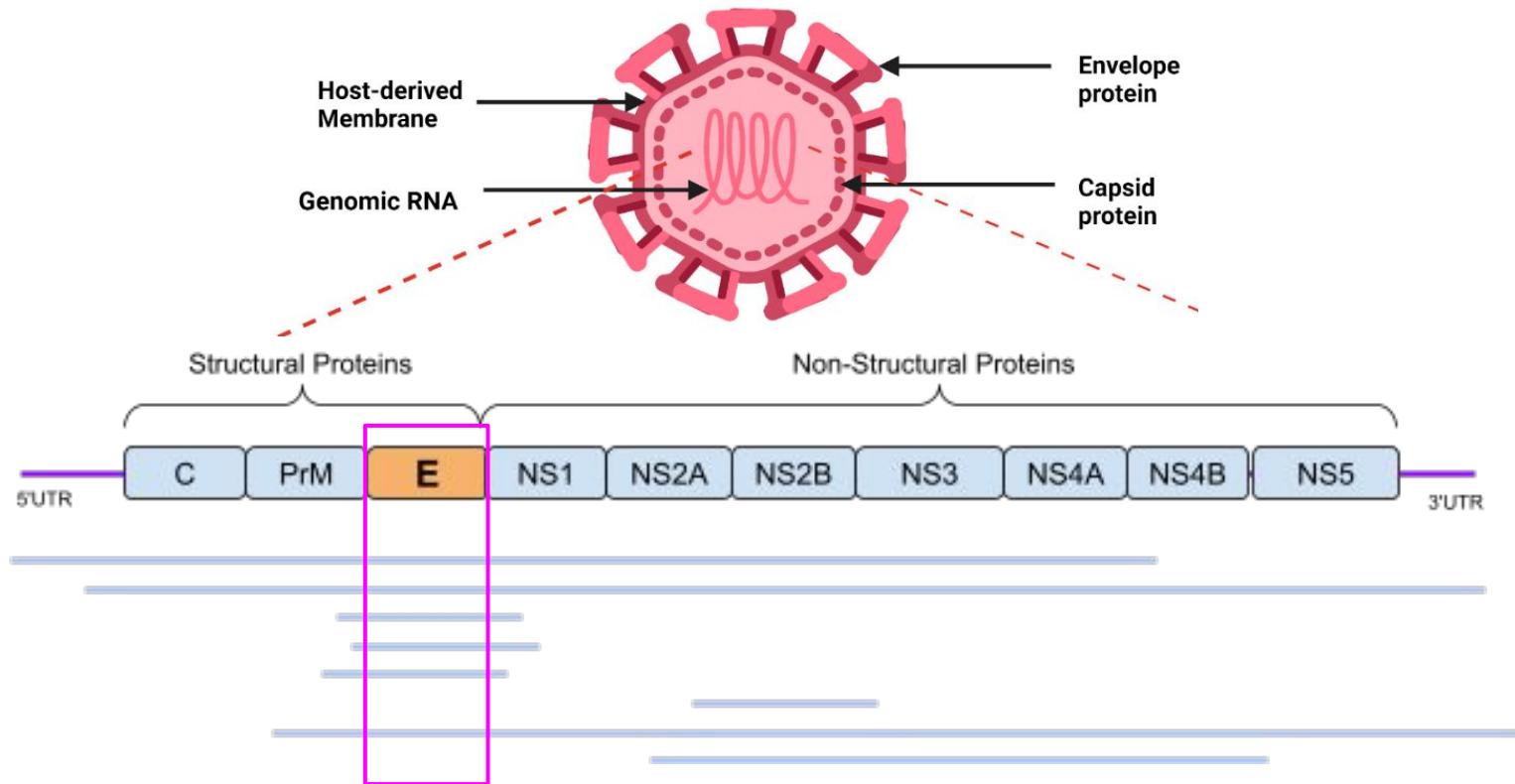
- The genome is about 11kb long encoding 10 genes

# 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

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

## 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

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`? For example:
  - a. dengue\_denv1.json
  - b. dengue\_denv1\_E.json

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

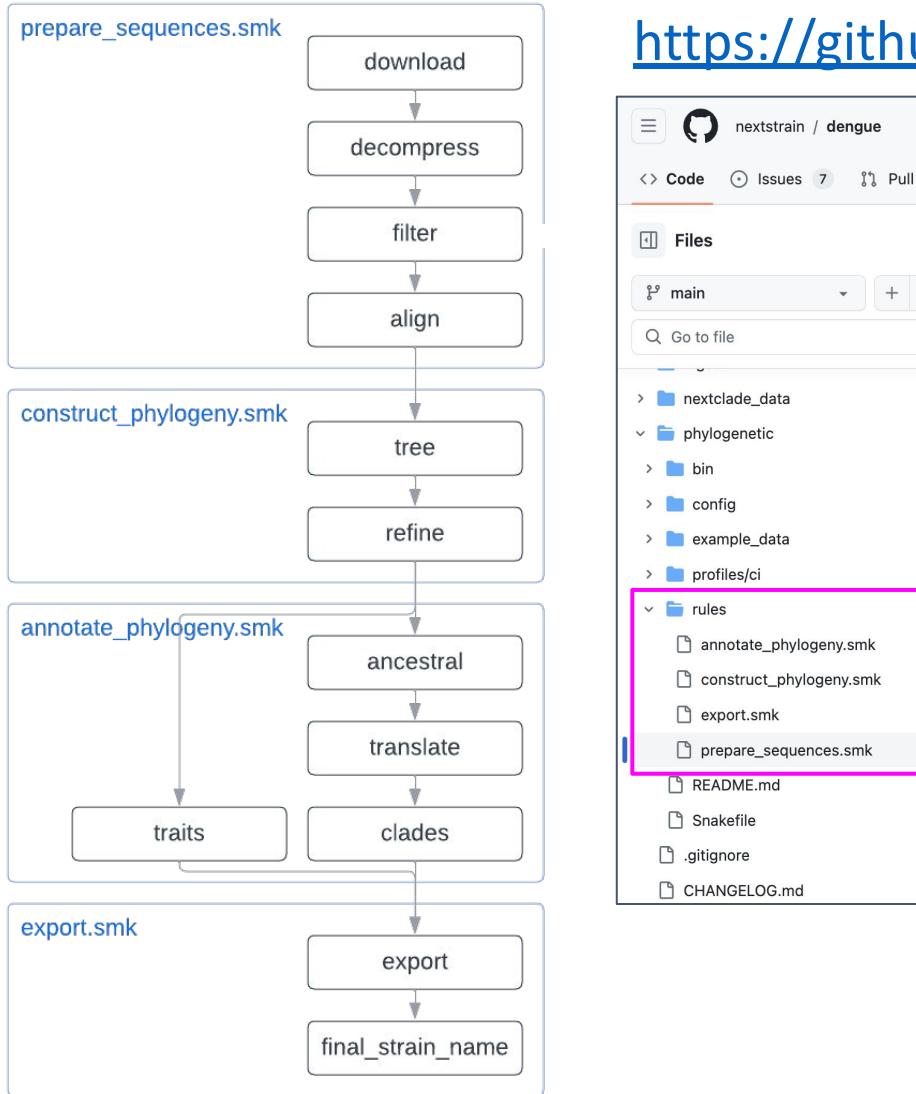
j23414 mentioned this issue on Jan 2  
Update manifest with dengue gene datasets nextstrain/nextstrain.org#771  
1 task

Merged

# 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

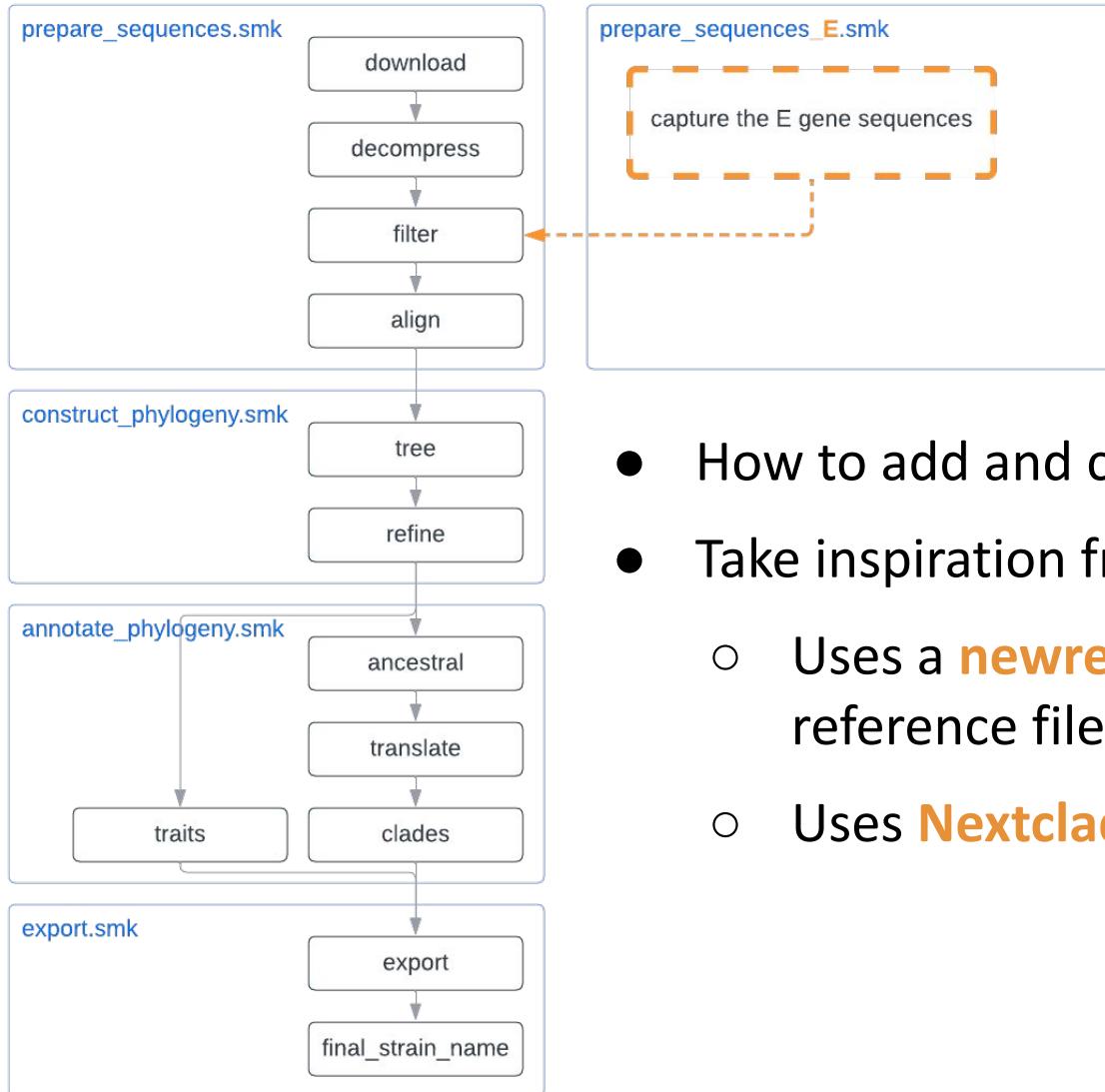


<https://github.com/nextstrain/dengue>

```
1 """
2 This part of the workflow prepares sequences for constructing the phylogenetic tree.
3 REQUIRED INPUTS:
4     metadata_url      = url to metadata.tsv.zst
5     sequences_url     = url to sequences.fasta.zst
6     reference         = path to reference sequence or genbank
7 OUTPUTS:
8     prepared_sequences = results/aligned.fasta
9 This part of the workflow usually includes the following steps:
10    - augur index
11    - augur filter
12    - augur align
13    - augur mask
14 See Augur's usage docs for these commands for more details.
15 """
16
17 rule download:
18     """Downloading sequences and metadata from data.nextstrain.org"""
19     output:
20         sequences = "data/sequences_{serotype}.fasta.zst",
21         metadata = "data/metadata_{serotype}.tsv.zst"
22
23 params:
```

Pipeline is organized according to the  
[GitHub: nextstrain/pathogen-repo-guide](https://github.com/nextstrain/pathogen-repo-guide)

# Dengue: Adding an "E" gene build



- How to add and connect the "E" gene builds?
- Take inspiration from [GitHub: nextstrain/rsv](https://github.com/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

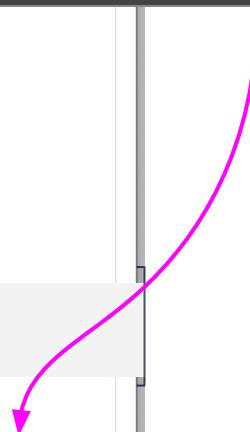
Code Blame 271 lines (271 loc) · 17.1 KB

```
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
```

...

```
49           /protein_id="NP_740317.1"
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"
66           /protein_id="NP_740319.1"
```

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



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

[Sign in now](#) to use Zenhub

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

E.gb

```
1 LOCUS      DENV4/NA/REFERENCE/2003 1485 bp    DNA          UNK 01-JAN-1980
2 DEFINITION  Dengue virus 4, complete genome.
3 ACCESSION   NC_002640
4 VERSION     NC_002640.1
5 KEYWORDS    .
6 SOURCE      .
7 ORGANISM    .
8 .
9 FEATURES    Location/Qualifiers
10 CDS        1..1485
11             /gene="E"
12             /product="envelope protein E"
13             /protein_id="NP_740317.1"
14 source      1..1485
15             /clone="rDEN4"
16             /db_xref="taxon:11070"
17             /mol_type="genomic RNA"
18             /organism="Dengue virus 4"
19 ORIGIN
20             1 atgcgatcg taggatgtt aaacagagac tttgttggaaag gagtctcagg tggagcatgg
21             61 gtcgacctgg tgcttagaaca tggaggatgc gtcacaacca tggcccaggg aaaaccaacc
22             121 ttggattttg aactgactaa gacaacagcc aaggaagtgg ctctgttaag aacctattgc
23             181 attgaaggct caatatcaaataactacg gcaacaagat gtccaaacgca aggagagcct
24             241 tatctgaaag aggaacaggcca acatggccgtt gagatgttgtt agacagagggg
25             301 tggggcaatgtgtggcgtt gtttggaaaa ggaggatgttg tgacatgtgc gaatgtttca
26             361 tggccgggaa agataacagg caatgttgc caaatgttgc acctgttgc cacatgggtt
27             421 gtaacagtcc acaatggaga cacccatgtca gtggaaatgtt acacatccaa tcatgttgc
28             481 acagccatgtca taactccatgtca gtcacccatgtca aattgttgcgtt ctatgttgc
29             541 ctaacactcg attgttgcgtt cttgttgcgtt attgactttt atggatgtt tctgttgcgtt
30             601 atggaaaatgtt aacatgttgcgtt cttgttgcgtt attgactttt atggatgtt tctgttgcgtt
31             661 acatgttgcgtt cttgttgcgtt attgactttt atggatgtt tctgttgcgtt
32             721 aatgttgcgtt atgttgcgtt acatgttgcgtt atggatgtt tctgttgcgtt
33             781 cttgttgcgtt atgttgcgtt acatgttgcgtt atggatgtt tctgttgcgtt
34             841 ggttgcgtt atgttgcgtt acatgttgcgtt atggatgtt tctgttgcgtt
35             901 atgttgcgtt atgttgcgtt acatgttgcgtt atggatgtt tctgttgcgtt
36             961 atgttgcgtt atgttgcgtt acatgttgcgtt atggatgtt tctgttgcgtt
```

E.fasta

```
1 >NC_002640.1 Dengue virus 4, complete genome
2 ATGCGATGCCTAGGAGTAGGAACAGAGACTTTGTGAAAGGAGTCTCAGGGTGGAGCATGG
3 GTGACCTGGTGTCTAGAACATGGAGGATGCGTCACAACCATGGCCAGGGAAACCAAC
4 TTGGATTTGAACCTGACTAAGACAACAGCAAGGAAGTGGCTGTAAAGAACCTATTGC
5 ATTGAAGCCTCAATATCAAACATAACTACGGCAACAAGATGTCCAACGCAGAGGAGCCT
6 TATCTGAAAGAGGAACAGGCCAACAGTACATTGCCGGAGAGATGTGGTAGACAGAGGG
7 TGGGGCAATGGCTGTGCTTGGAAAAGGAGGAGTGTGACATGTGCGAAGTGGTTCA
8 TGTTGGGGAAAGATAACAGGCAATTGGTCAAATTGAGAACCTGAATACAGTGGTT
9 GTAACAGTCACATGGAGACACCCATCGAGTAGGAAATGACACATCCAATCATGGAGTT
10 ACAGCCATGATAACTCCAGGTACCATCGGTGAAAGTCAATTGCCGACTATGGAGAA
11 CTAACACTCGATTGTGAAACCCAGGTCTGAAATTGACTTAAATGAGATGATTGATGAA
12 ATGAAAAGAAAACATGGCTGTGCTAAAGCAATGGTTTGGATCTGCCTCTCCATGG
13 ACAGCAGGAGCAGACACATCAGAGTTCACTGGAAATTACAAGAGAGAATGGTACATTG
14 AAGGTTCTCATGCCAAGAGACAGGATGTGACAGTGCTGGATCTCGAGGAGGCCATG
15 CATTCTGCCCTCGCTGGAGCCACAGAAGTGGACTCCGGTGTGGAAATCACATGTTGCA
16 GGACATCTTAAGTGCAGGAGTGGCTATGGAGAAATTGAGAATCAAGGGATGTCTACAG
17 ATGTTTCAGGAAAGTTCAATTGACAAGAGATGGCAGAAACACAGCATGGGACAACA
18 GTGGTGAAGTCAAGTATGAAGGTGCTGGAGCTCCGTGAAAGTCCCCTAGAGATAAGA
19 GATGTAACAAAGGAAAAAGTGGTGGCGTATCATCTCATCCACCCCCCTTGGCTGAGAAT
20 ACCAACAGTGTAAACCAACATAGAATTAGAACCCCCCTTGGGACAGCTACATAGTGTAA
21 GGTGTTGGAAACAGGCCATTAACACTCCATTGGTGTGGAAAGGGAGTCCATTGGCAAG
22 ATGTTTGAGTCCACATACAGAGGTGCAAACAGAATGGCATTCTAGGTGAAACAGCTTGG
23 GATTTGGTTCGGTGGACTGTTCACATCATTGGAAAGGCTGTGCAACAGGTTTTT
24 GGAAGTGTGTATACAACCATGTTGGAGGAGTCTCATGGATGATTAGAATCTAATTGGG
25 TTCTTAGTGTGTGGATTGGCACGAACTCGAGGAACACTTCATGGCTATGACGTGCATA
26 GCTGTTGGAGGAATCACTGTGTTCTGGGCTTCACAGTTCAGCA
27
```

- 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
```

- Starts with "ATG"
  - Seems to be aligned
  - Lower threshold for diverse viruses

(Nextclade 3.2 changelog )

## phylogenetic — less E\_sequences.fasta — 98x40

```
>X65240
ATGCGTTGCATAGGAATATCAAATAGAGACTTGTAGAAGGGGTTTCAGGAGGAAGCTGGGTGACATAGCTTAGAACATGGAAGTTGTGACGAC
GATGCCAAAAAAACAAACATTGGATTGAACTGATAAAAACAGAACCCACACAACCTGCACTCTAAGGAAG-TACTGTATAGAACGCTGA---
CCAATCACACACAGAACCTCGTTGCCAACACAAGGGGAAACCCAGCTAAATGAAGAGCAGGACAAAGGTTCGCTCGC---AAACACTTGTAGAC
AGAGGATGGGAAATGGATGTGGATTATGGAAAAGGAGTATTGTGACCTGTGCTATGTTACATGCAAAAGAACATGGAAGGAAACATGTCGCA
ACCAGAAAATTGGAAATACCCATCGTATAACACCTCACTCAGGAGAACGAGCACCGCTGAGGTAACGGACACAGGAAACATGCAAGGAAATCAAA
TAACACCCAGATTTCCATCACAG-AAGCAGAACTAACAGGTTATGGCACCGTCAGCAGGATGGAGTCTCCGAGAACGGGCTGGACTCAATGAGATA
GACTGCTGCAGATGGAAAGAACAGCTTGCCTAGTCACAGGCAATGGTTCTAGACCTCGGTTACCATGGCTACCCGGAGCGGACACA-CA-AGGA
TCA----ATTACAAAGAGACATTGTCATCTTCAAAATCCCCATGCGAACAAACAGGATGTCGTTTTAGGATCTCAAGAAGGGGCCATGCACA
CGGCACTCACAGGGGCCAGAAATCCA-----GATGGAATTAC-TATTCACAGGACATCTCAAGTGCGACTGAGAACATGGACAAACTACAGCTC
AAAGGAATGTCATACTATGTCAGGAAAGTTCAAATTGTAAGGAAATACCGAACACAAACATGGAAACAATAGTTATCAGAGTACAATATGA
AGGAGACGGCTCTCATGTAAGATCCCTTGGAGATAATGGATGGAAAAAAAG--ACATGTCAGGCTGCGCTGATTACATTAACCCGATTCAGAACAGA
AAAC---CAGT----CAACATAGAACGAGAACCTCCATTGGAGACAGCTACATCATGAGTGGGAGA---ATTGAAACTCCACTGGTT
AAGAAAGGAAGTTCATCGGCAAATGTTGAGACAACATGAGGGAGCGAACAGAATGCCATTAGGTCAGACAGGCTGGATTGGATCCCT
GGGAGGAGTGTACATCATAGAAAGGCTCTCCACAAAGTTTGGAGCTATCTATGGGGCGCTTTAGTGGGCTCATGGACTATGAAAATCC
TCATCGGAGTCATCATCACATGGATAGGAATTCACTGAGCACCTCACTGTCTGTGACTAGTATGGTGG--GGAGTCATAACACTGTACTTGGG
GCCATGGTCAGGCT
>X76219
ATGCGATGCGTGGGAATAGGCAACAGAGACTTCGTGAAAGGACTGTCAAGGAGAACGTGGGTGGATGTGGTACTGGAGCATGGAAGTTGCGTACCCAC
CATGCCAAAAGATAAAACCAACATTGGACATTGAACTCTGAAAGAC-----GGAGGTCACTGCGTAA---CTGT-GCATTGAAGCTAAAATAT
CAAACACCACCCACGATTCAAGATGTCCAACACAAGGGGAGCCACA-CTGGTGGAAAGAACAGACGCGAACTCTGTGTGCGACGAACGTTTGTGGAC
AGAGGCTGGGAATGGCTGGCTCTTGGAAAAGGATGCGCTAAACCTGTCATAAGT---CAAGTGTGTG-CACAAACAGGAAAGGATTTCTCA
ATATGAGAACTTGAATATTCACTGATGACTCGCTCCACACTGGTCAAGGACACCGAACAGGTTGGAAATGAGAACACAGAACATGGAAATTGCAACCATTA
CACCTCTCATGCTA-----GAAATACAGCTGACCGACTACCGAGCTCTTCAATTGGATTGTCACCCAGAACAGGGCTAGACTTTAATGAGATG
GTGTTGTCATGAAATGAAAGAAAATCATGGTTGTCACAAACATGGTTCTAGACTTACACTGCCCTGGACCTGGAGCTTCACACCCAGAGAC
T---TGAAC-----AGAGAATGGTTACATTAAAGACAGCTATGCAAGAACGAGCAGGAAGTGTGACTAGGATCACAAGAAGGAGCAATGCA
CTGCGTGTGACCGGAGCGACAGAACATCCACAGCTGCGACCAA--AAATTGGTGCAGGACACTGAATGAGACTAAAAATGGCAACACTGACCTTA
AAAGGAGTGTCAATGTCATGTCACAGG-----TTCAAGTGAGAACAGGACTGGCTGAGACCCAGCATGGAACTGTCTAGTCAGGTTAACATGAG
AGGAACAGATGCACTGCAAGATCCCCTTGGATGAGAAAAGGATGCAAGGAGCTGAGACCCAGCATGGAACTGTCTAGTCAGGTTAACATGAG
AAAAC---CAGT----CAACATTGAGGAGAACACCCATTGGTGAAGAACATGCCCAGGAGAGCACGAAGGACGCCATCTAGGAGACACCGCATGGACTTC
AAGAAAGGAAGCAGCATTGGAAAATGCTTGAAGCAACTGCCCAGGAGAGCACGAAGGACGCCATCTAGGAGACACCGCATGGACTTC
AGGAGGAGTGTTCACGCTGTGGAAAATGGTACCCAGATCTTGTGAACTGATATGGAGTTTGTTCAGCGGTGTTCTGGACTATGAAAATAG
GAATAGGGATTCTGTCATGGCTAGGATTAAATTCAAGGAGCACGTCCTTGGACGAGCTGCACTGGTGGCATGGTACACTGTACCTAGGA
GTCATGGTCAGGCG
>X15433
ATGCGTTGCATAGGAATATCAAATAGAGACTTGTAGAGGGGTTTCAGGAGGAAGCTGGGTGACATAGCTTAGAACATGGAAGCTGTGACGAC
GATGCCAAAAAAACAAACATTGGATTGAACTGATAAAAACAGAACCCACACGCCCTGCCACCTAAGGAAGTACTGTATAGAACGCTTA---
CCAACACACACAGAACATCTGTTGCCAACACAAGGGAAACCCAGCTAAATGAAGAGCAGGACAAACACTCCAT-----GGTAGAC
AGGAGTGGGAAATGGATGTGGACTATTGGAAAGGGAGGAGCTTGTGACCTGTGCTATGTTCACATGCCAAAAGAACATGGAAGGAAAATCGTGCAC
:
```

# 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
```

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

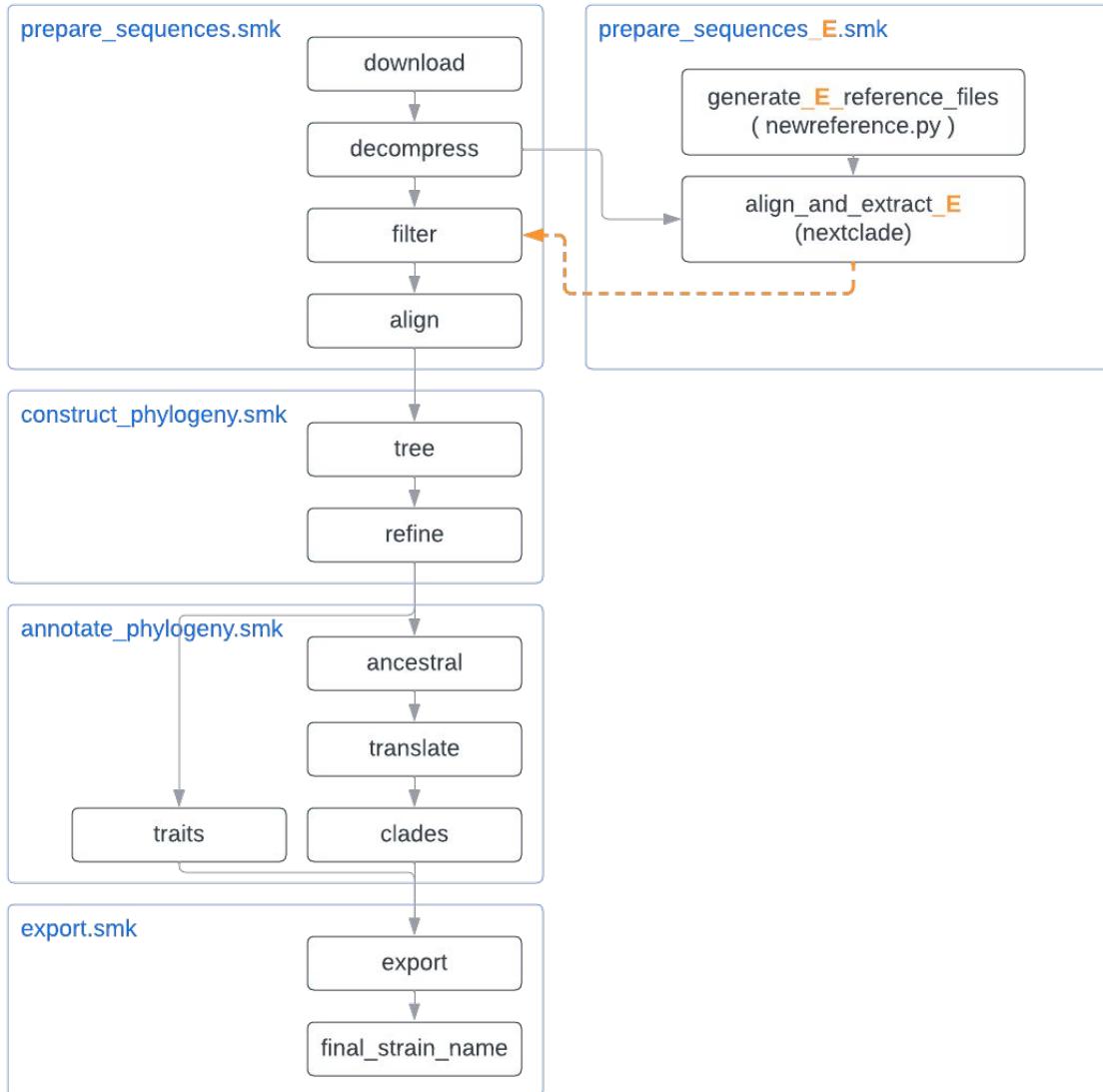
sequences\_all.fasta (46,342)  
time: less than a minute

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

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

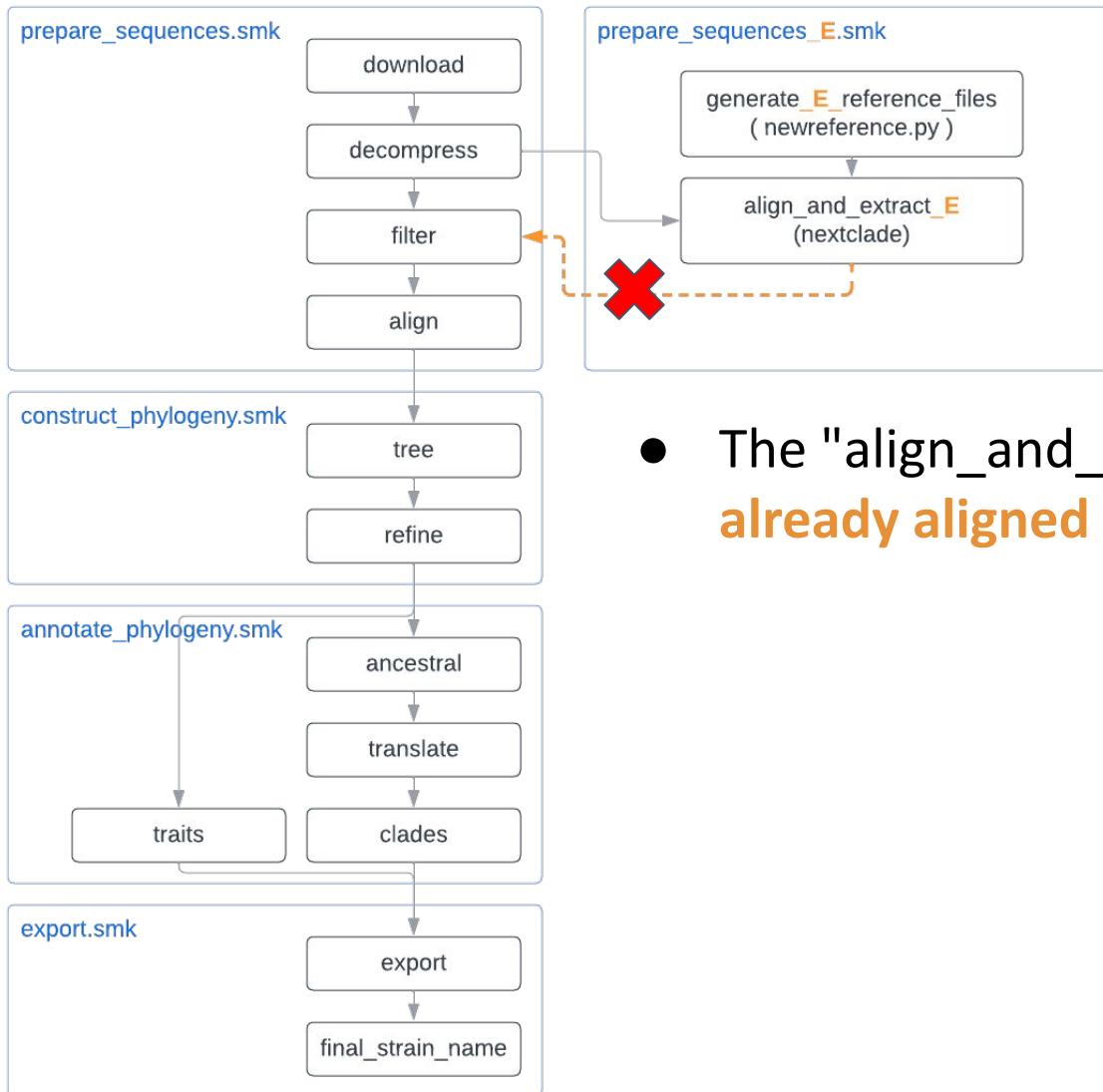
sequences\_denv4.fasta (3,874)  
time: around 1.5 hours

# Prepare "E" gene sequences



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

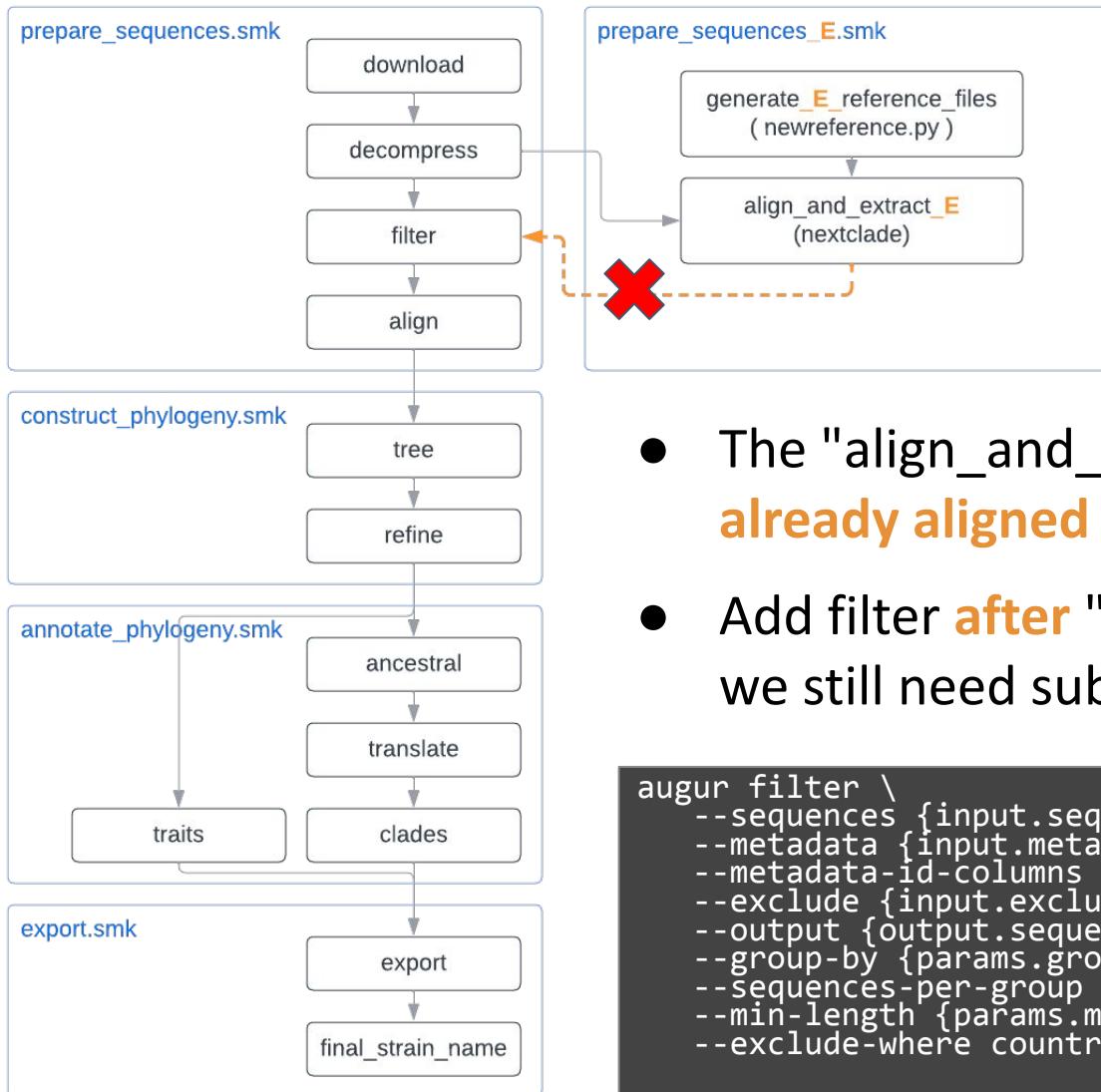
# Prepare "E" gene sequences



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

- The "align\_and\_extract\_E" output is **already aligned**

# Prepare "E" gene sequences

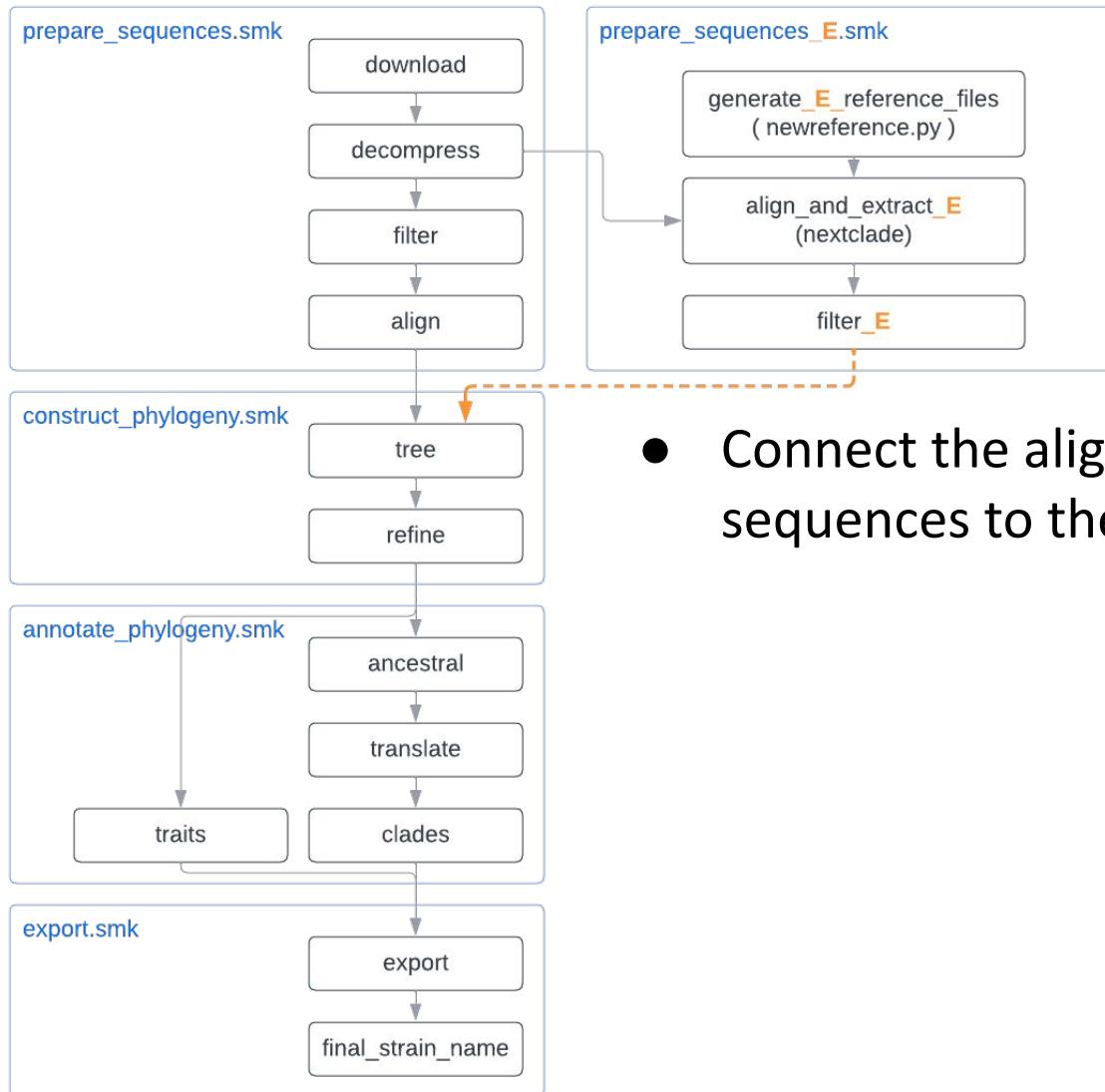


# How to connect the aligned E sequences to the pipeline?

- The "align\_and\_extract\_E" output is **already aligned**
  - Add filter **after** "align\_and\_extract\_E" since we still need subsampling

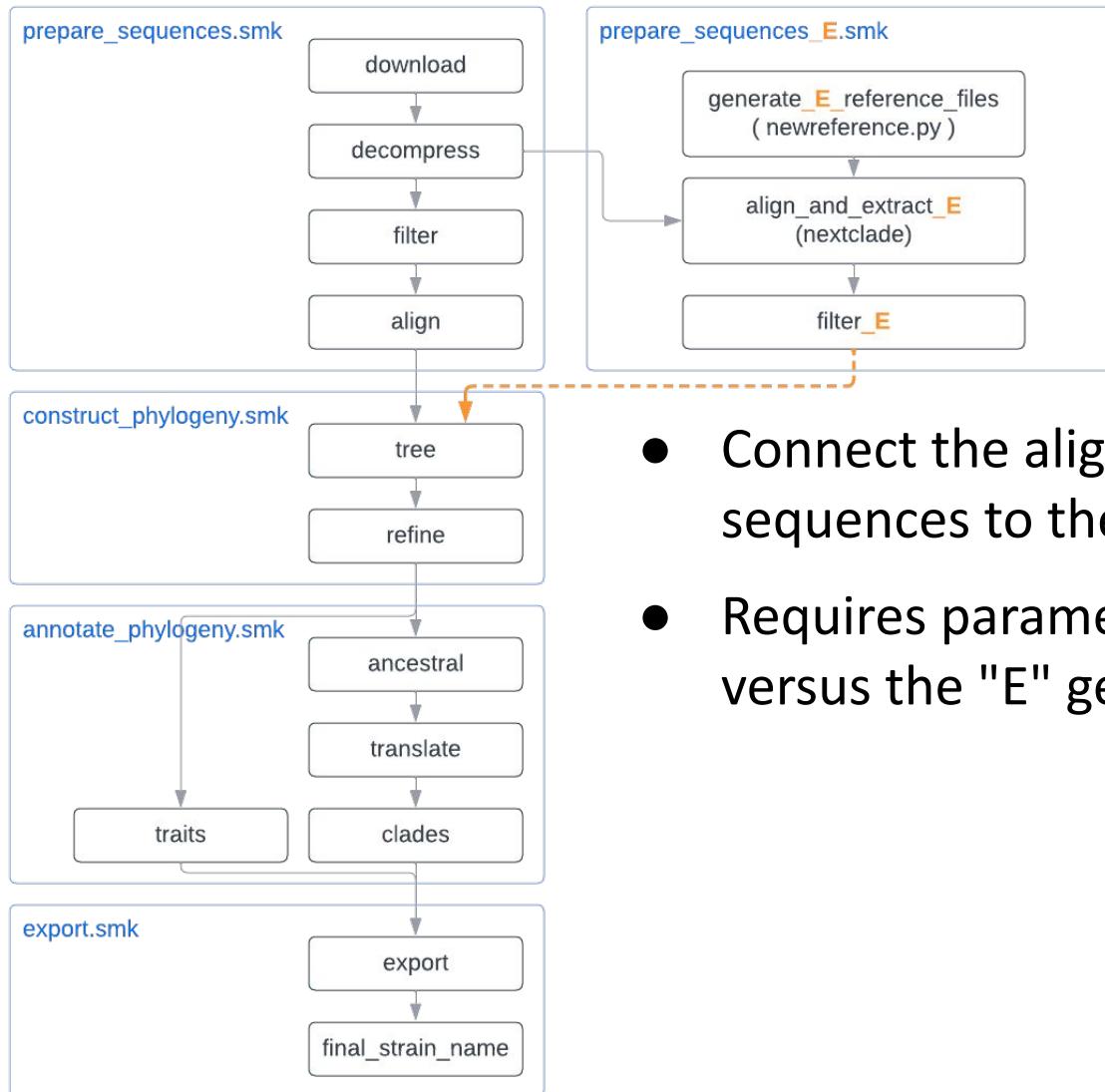
```
augur filter \
--sequences {input.sequences} \
--metadata {input.metadata} \
--metadata-id-columns {params.strain_id} \
--exclude {input.exclude} \
--output {output.sequences} \
--group-by {params.group_by} \
--sequences-per-group {params.sequences_per_group} \
--min-length {params.min_length} \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

# Push "E" sequences through pipeline



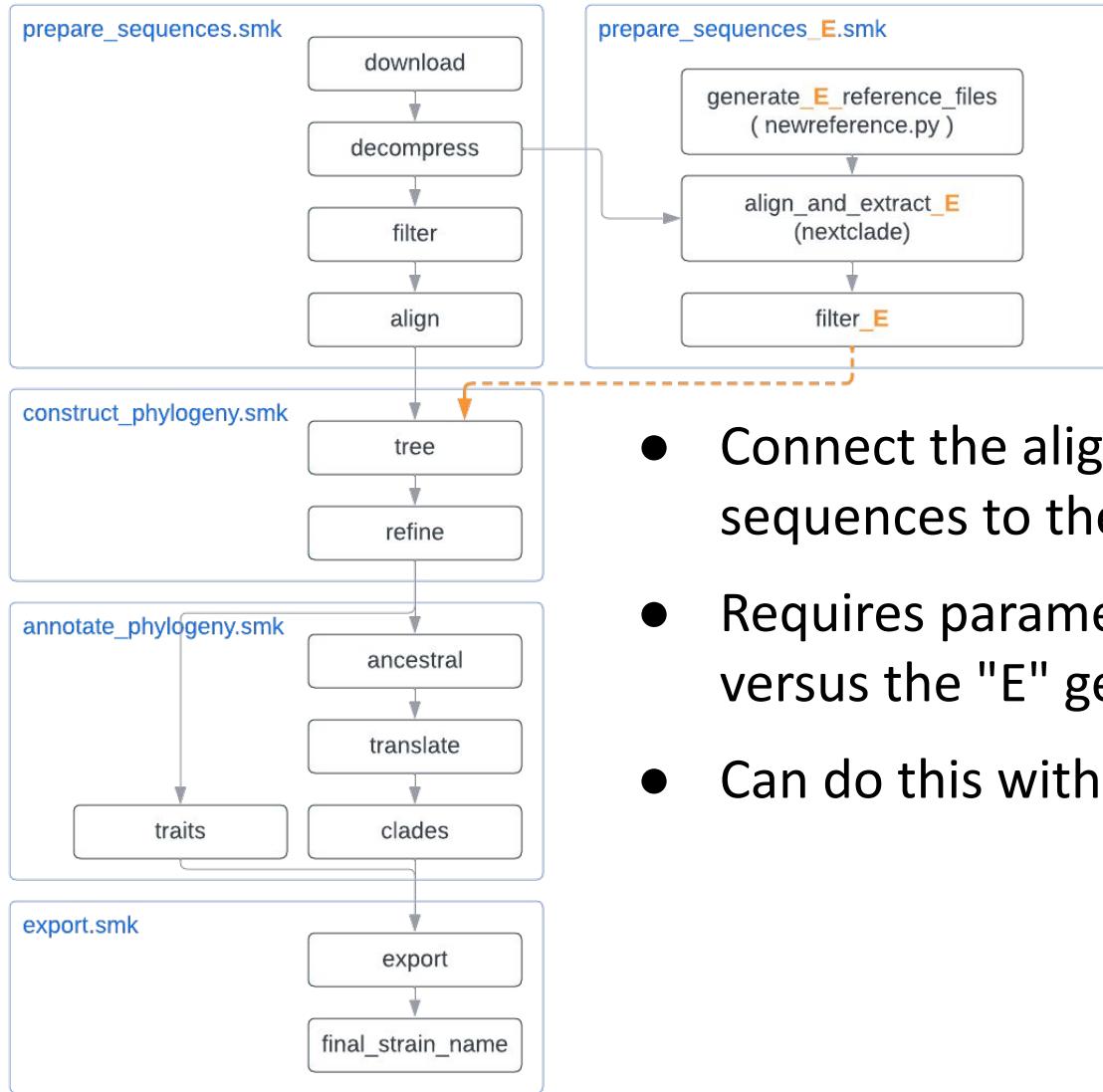
- Connect the aligned and subsampled E gene sequences to the rest of the pipeline

# Push "E" sequences through pipeline



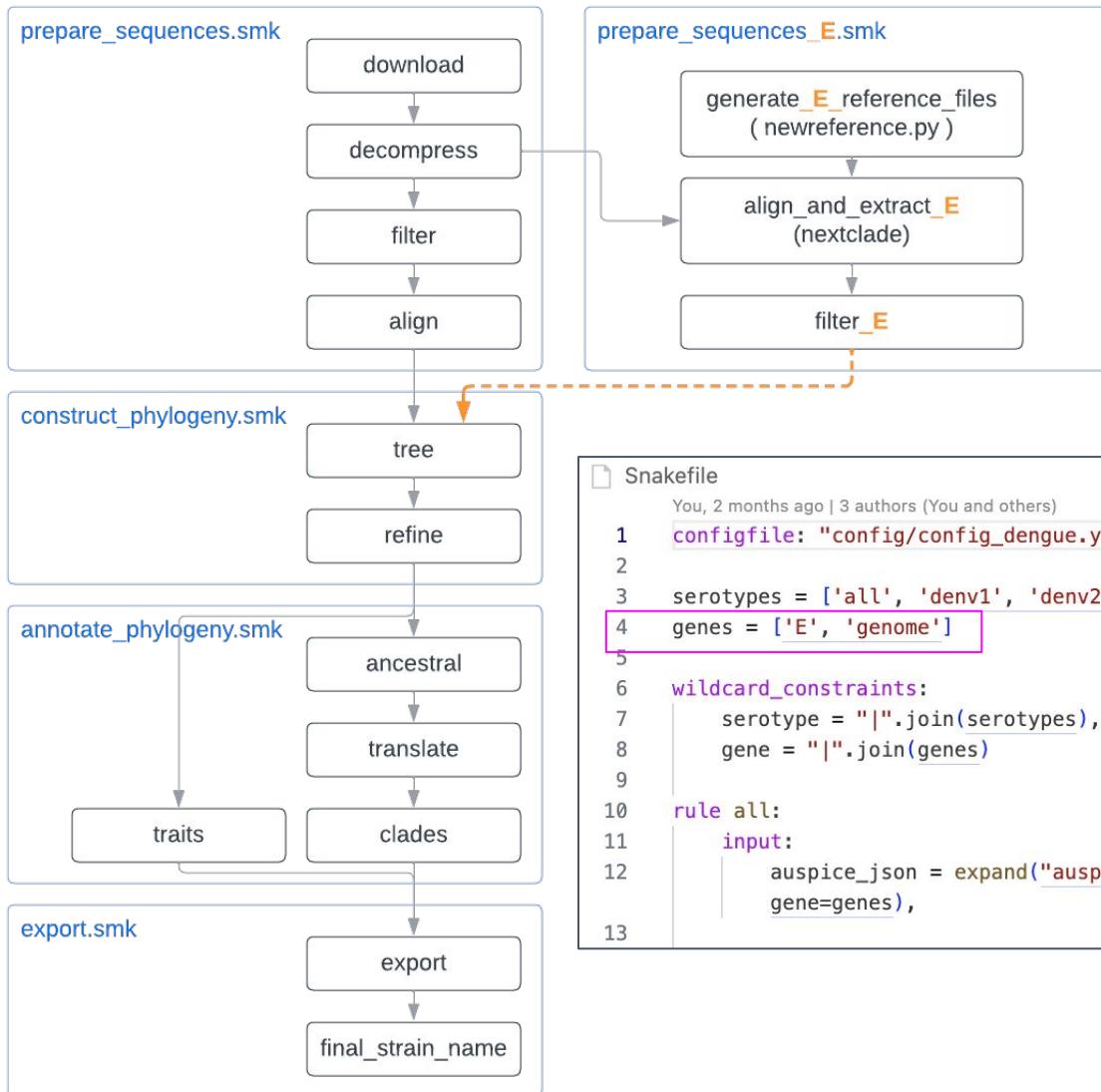
- Connect the aligned and subsampled E gene sequences to the rest of the pipeline
- Requires parameterizing whole "Genome" versus the "E" gene files

# 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}



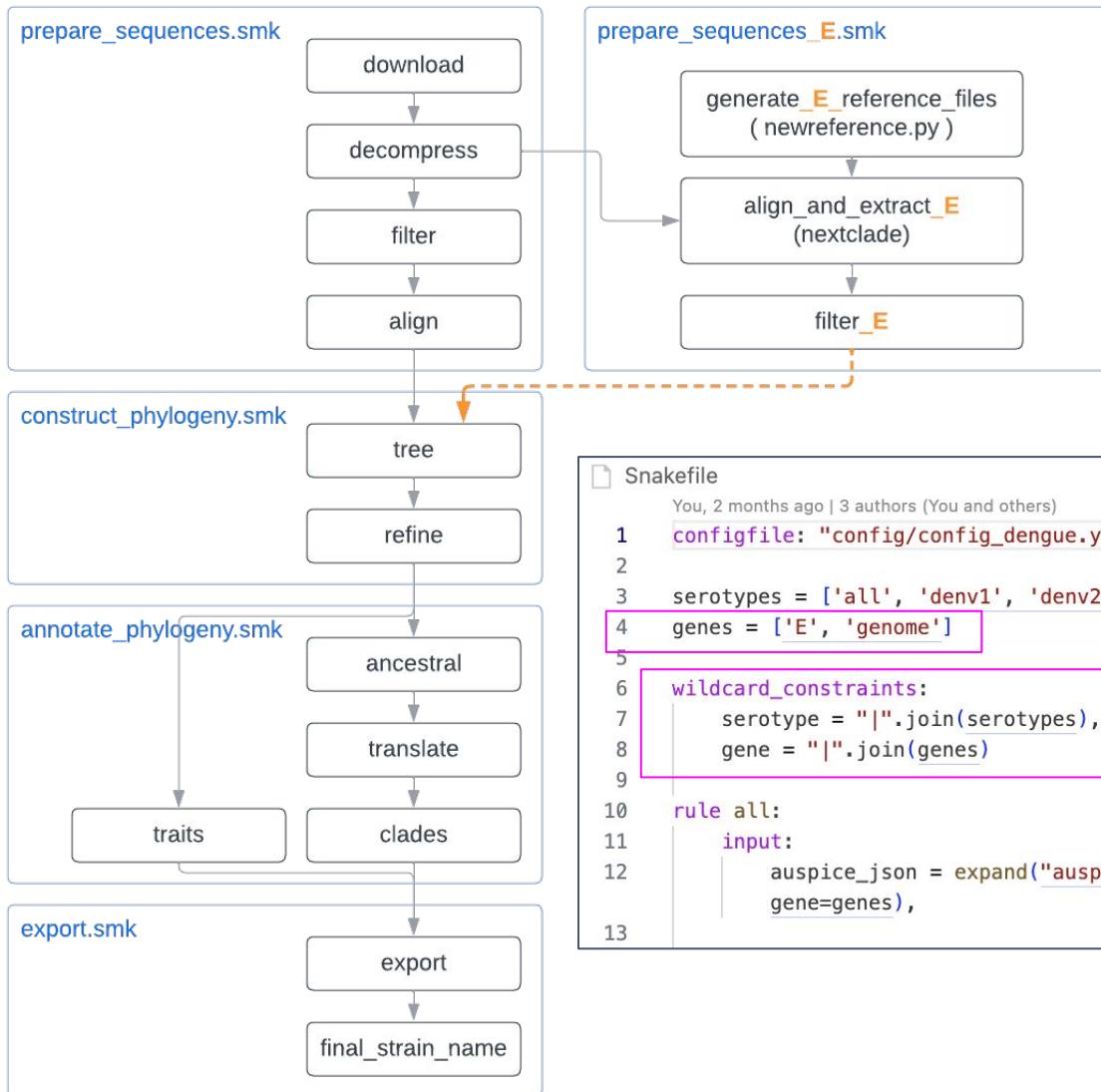
Add wildcards to the phylogenetic/Snakefile

```
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 wildcard_constraints:
6     serotype = "|".join(serotypes),
7     gene = "|".join(genes)

8 rule all:
9     input:
10        auspice_json = expand("auspice/dengue_{serotype}_{gene}.json", serotype=serotypes,
11                               gene=genes),
```

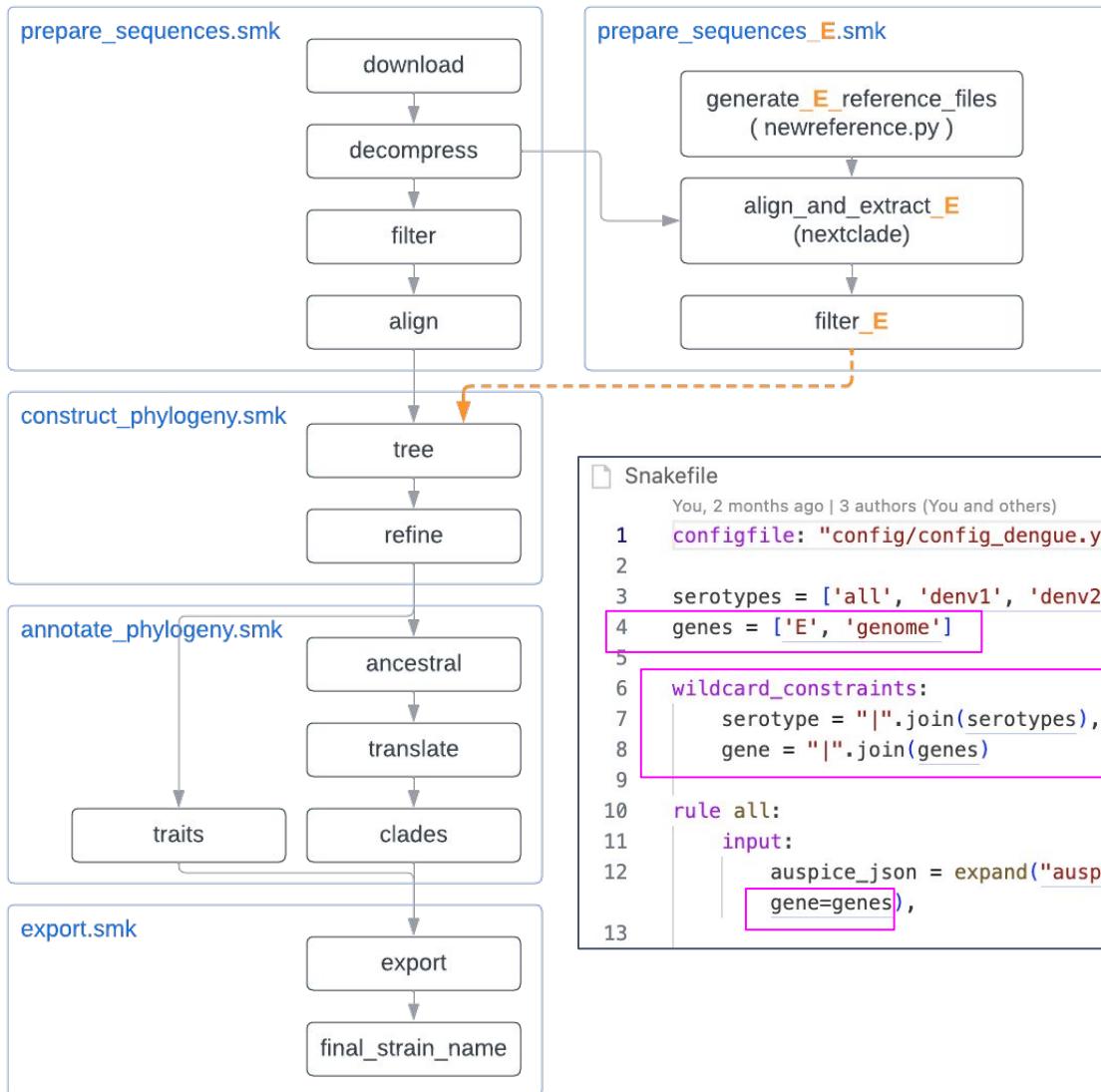
# Sidenote on {wildcards}



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,
13                                     gene=genes),
```

# Sidenote on {wildcards}



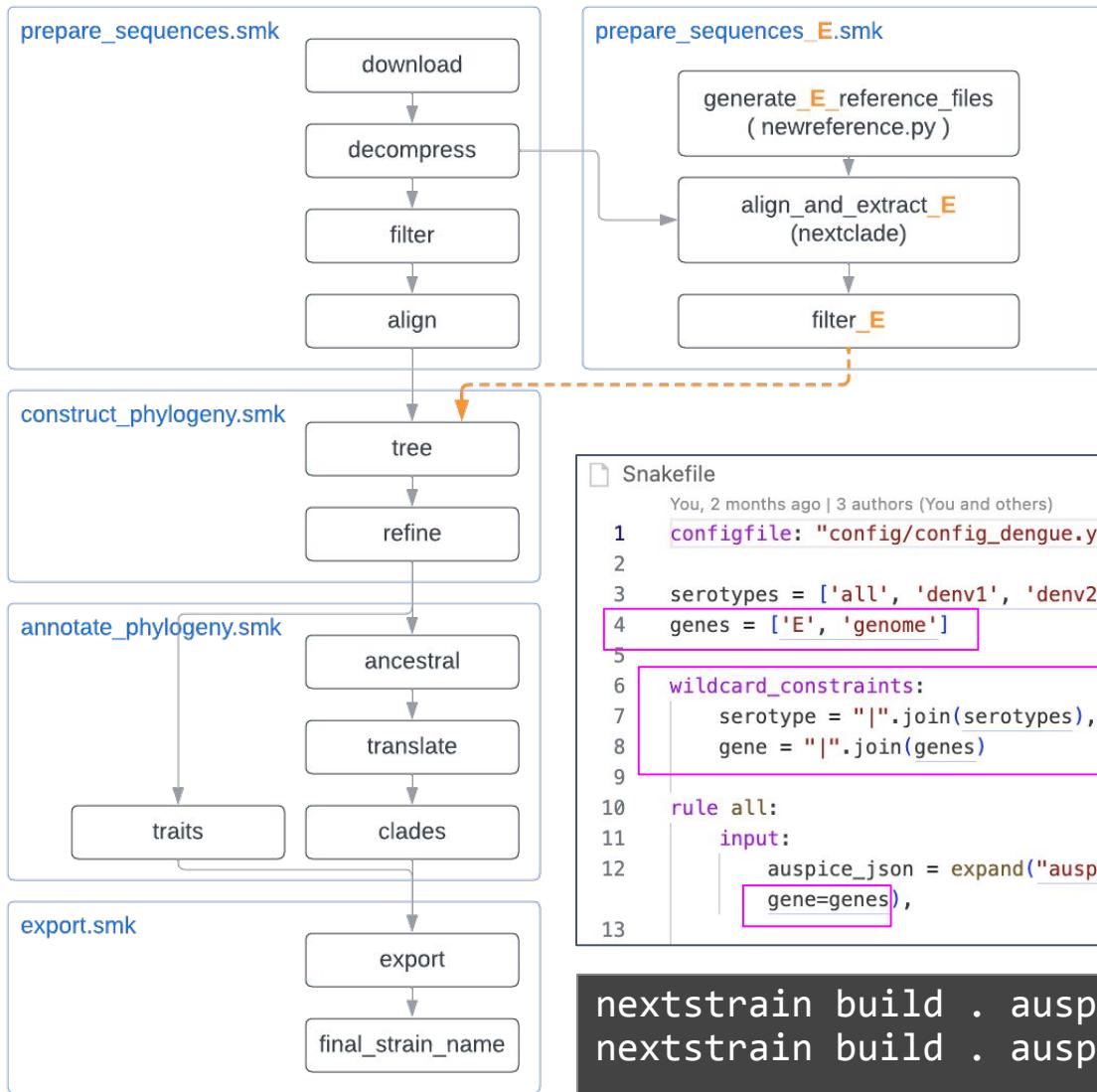
Add wildcards to the phylogenetic/Snakefile

```
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 wildcard_constraints:
6     serotype = "|".join(serotypes),
7     gene = "|".join(genes)

8
9 rule all:
10    input:
11        auspice_json = expand("auspice/dengue_{serotype}_{gene}.json", serotype=serotypes,
12                               gene=genes),
```

# Sidenote on {wildcards}

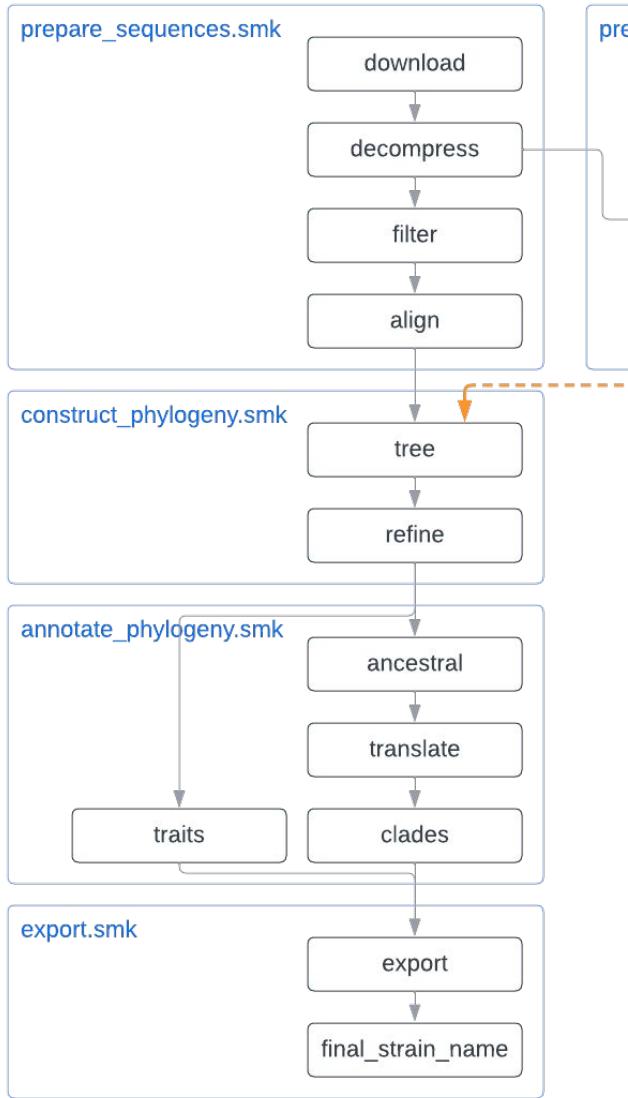


Add wildcards to the phylogenetic/Snakefile

```
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,
13                                     gene=genes),
```

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

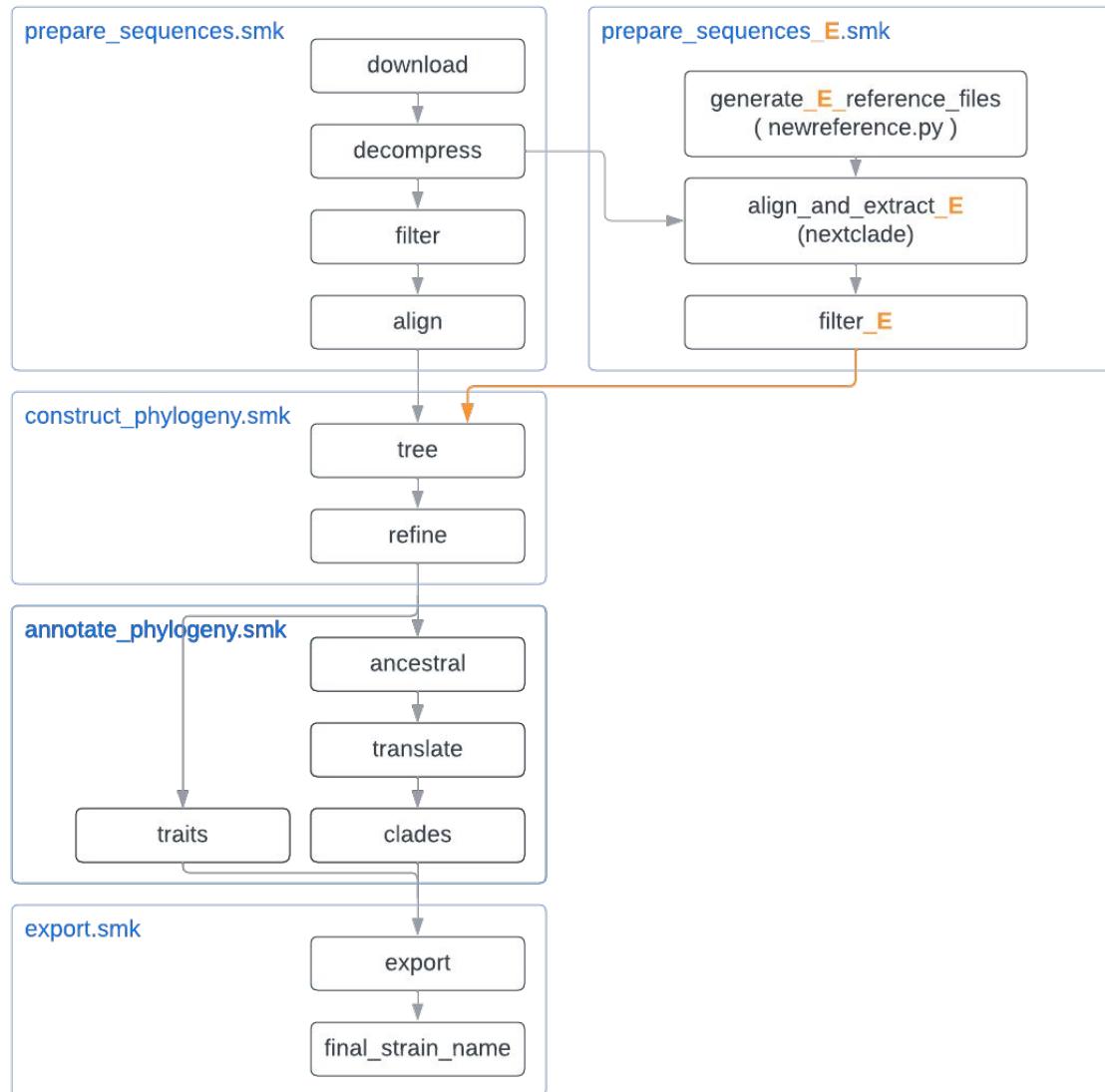
# Sidenote: use \_{gene} wildcards



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

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
38     input:
39         tree = "results/tree-raw_{serotype}_{gene}.nwk",
40         alignment = "results/aligned_{serotype}_{gene}.fasta",
41         metadata = "data/metadata_{serotype}.tsv"
42     output:
43         tree = "results/tree_{serotype}_{gene}.nwk",
44         node_data = "results/branch-lengths_{serotype}_{gene}.json",
45     params:
46         coalescent = "const",
47         date_inference = "marginal",
48         clock_filter_iqd = 4,
```

# Connected! Try running the pipeline



# Complications with "clades"

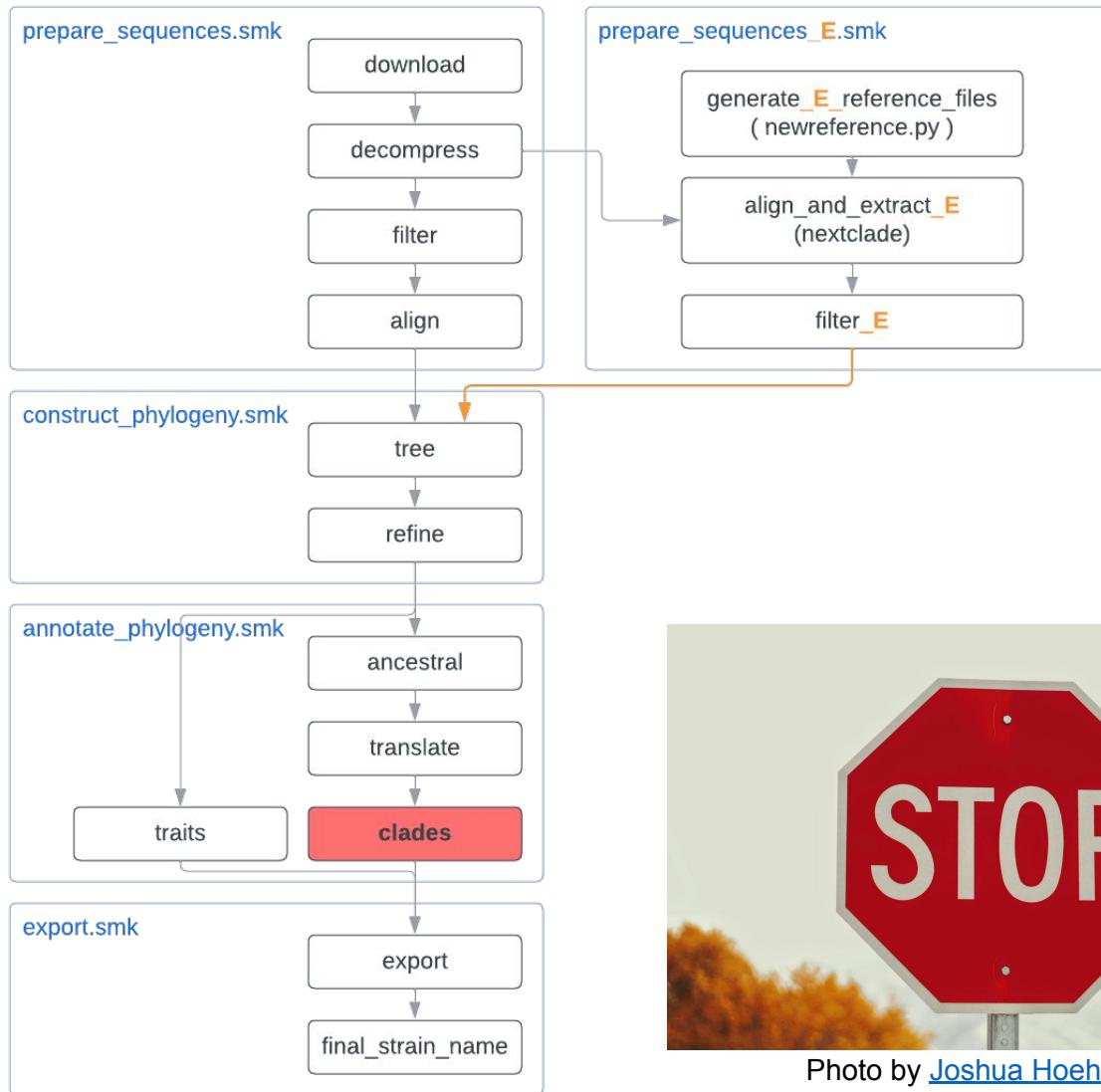
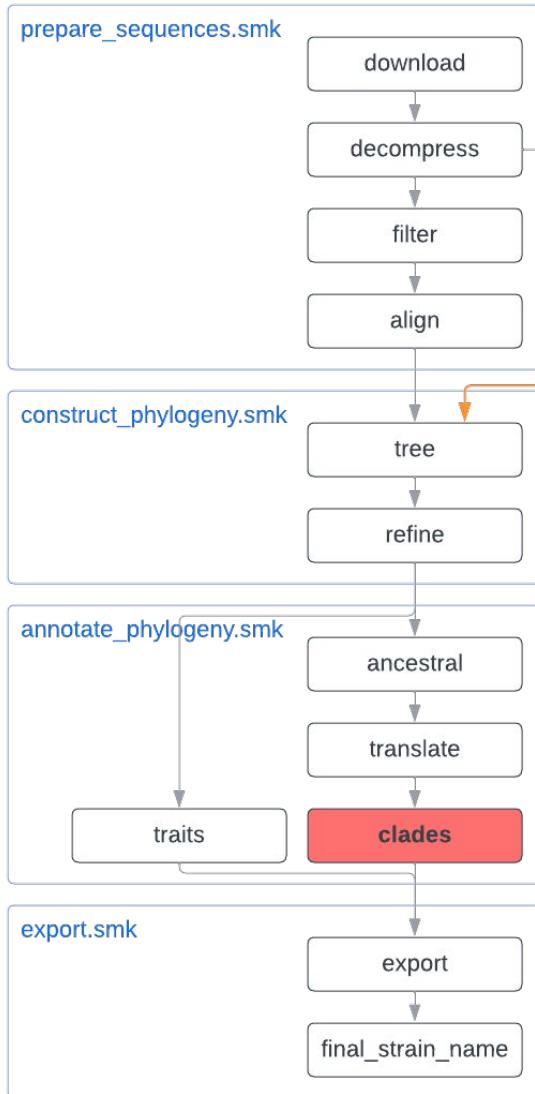


Photo by [Joshua Hoehne](#) on [Unsplash](#)

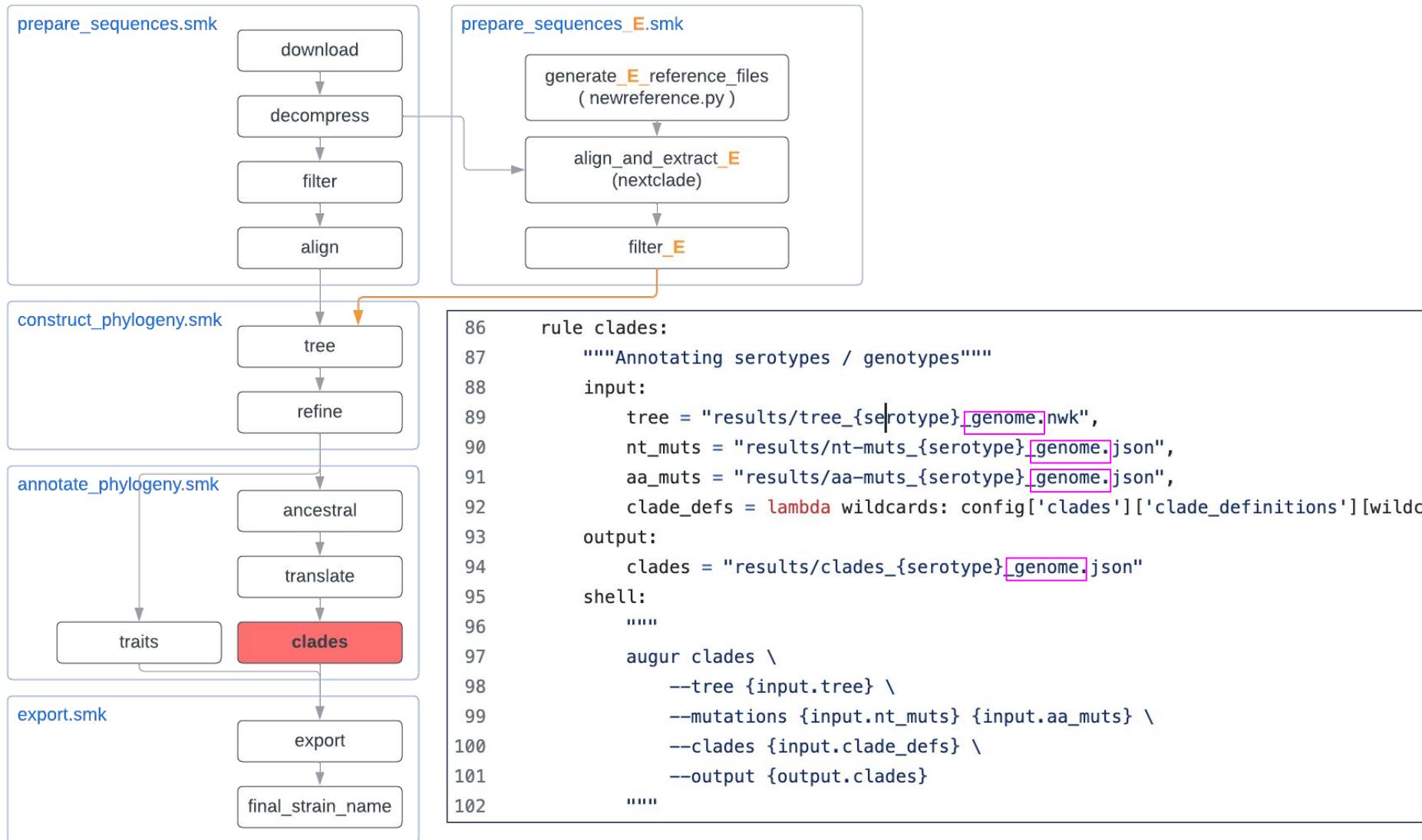
# Complications with "clades"



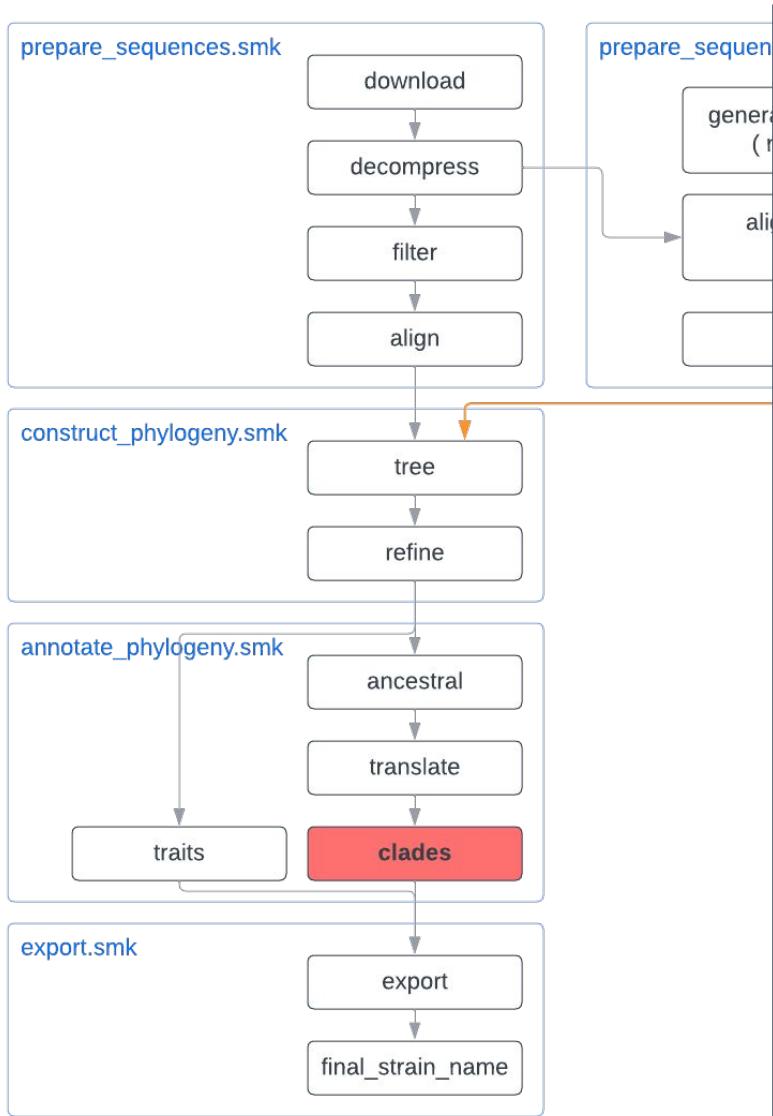
```
augur clades \  
  --tree {input.tree} \  
  --mutations {input.nt_muts} {input.aa_muts} \  
  --clades clades_genotypes.tsv \  
  --output {output.clades}
```

	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					--									

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



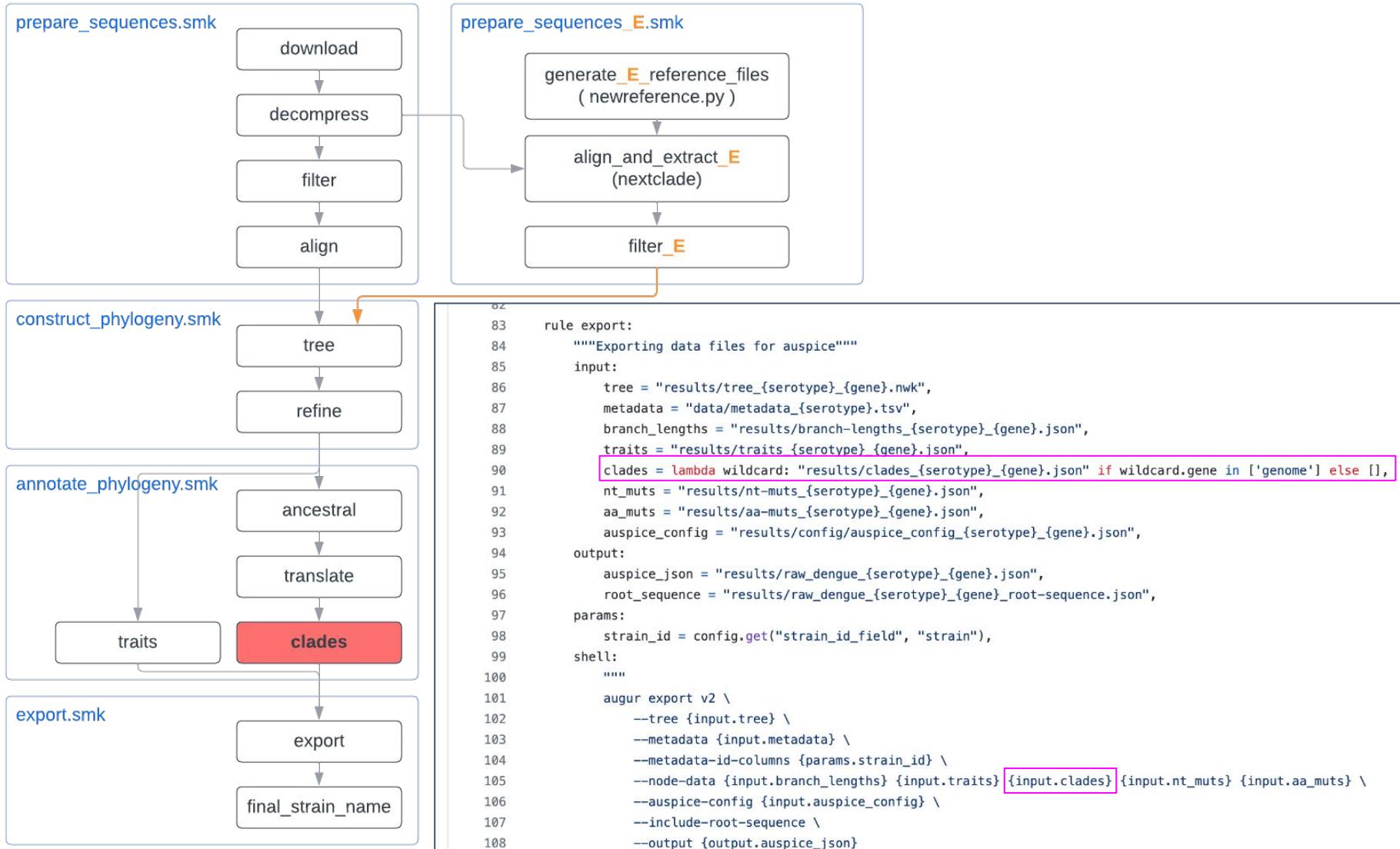
# fix: (2/3) Rely on "Nextclade" metadata



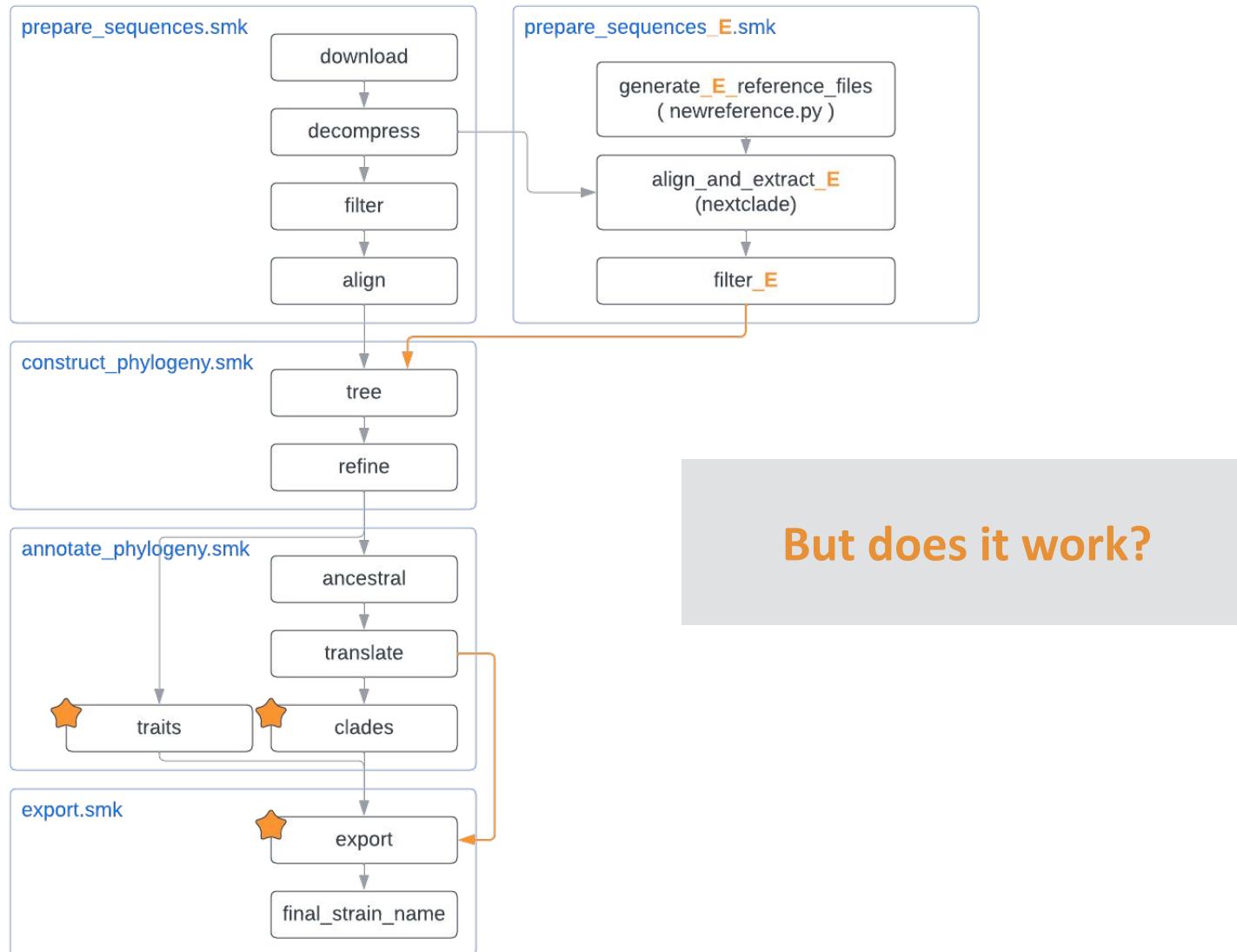
## dengue / phylogenetic / config / config\_dengue.yaml

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	traits:	
23	sampling_bias_correction: '3'	
24	traits_columns:	
25	all_genome: 'region'	
26	denv1_genome: 'country region'	
27	denv2_genome: 'country region'	
28	denv3_genome: 'country region'	
29	denv4_genome: 'country region'	
30	all_E: 'region nextclade_type'	
31	denv1_E: 'country region nextclade_subtype'	
32	denv2_E: 'country region nextclade_subtype'	
33	denv3_E: 'country region nextclade_subtype'	
34	denv4_E: 'country region nextclade_subtype'	
35		
36		

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

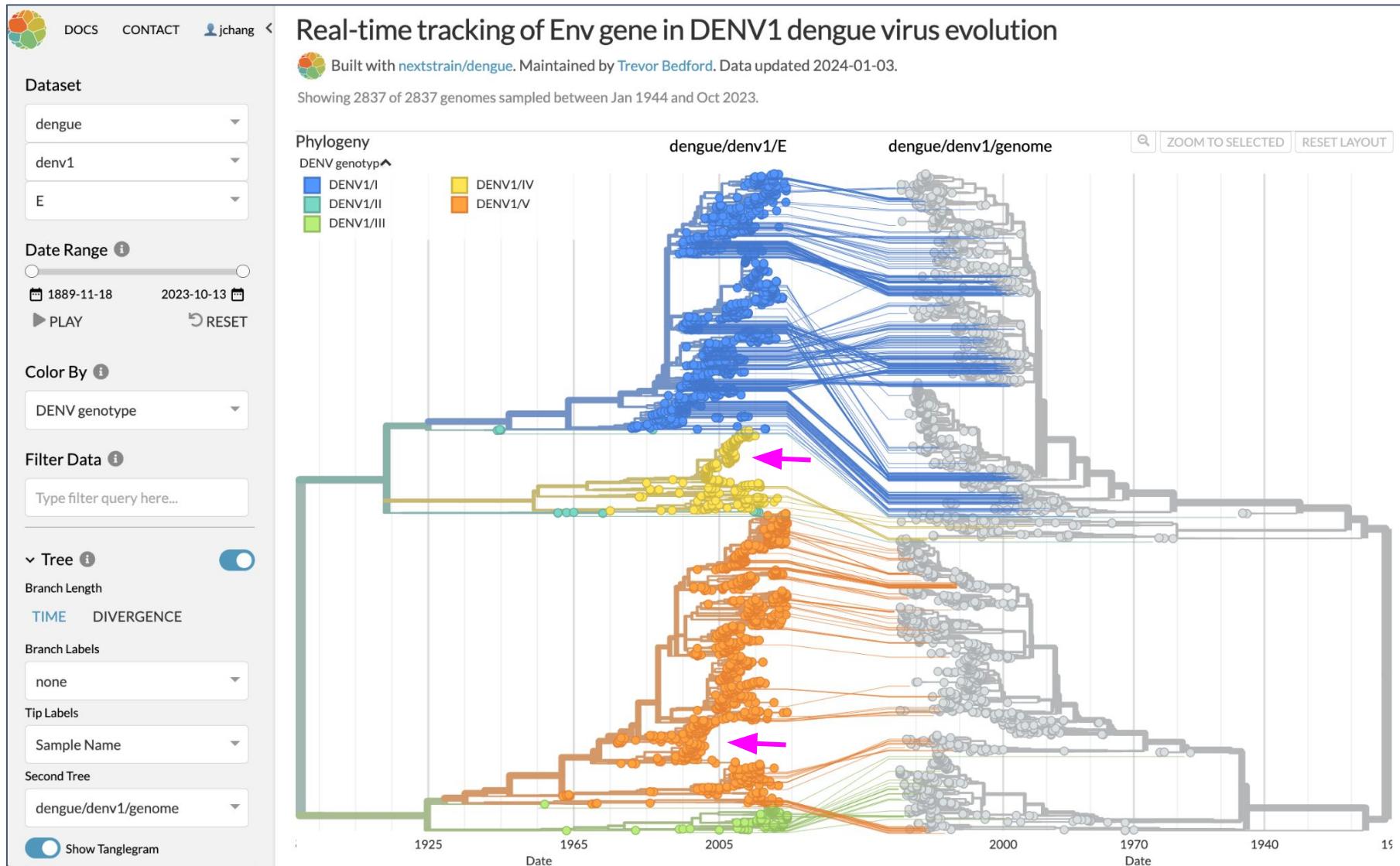


# Dengue pipeline (with "E" gene)



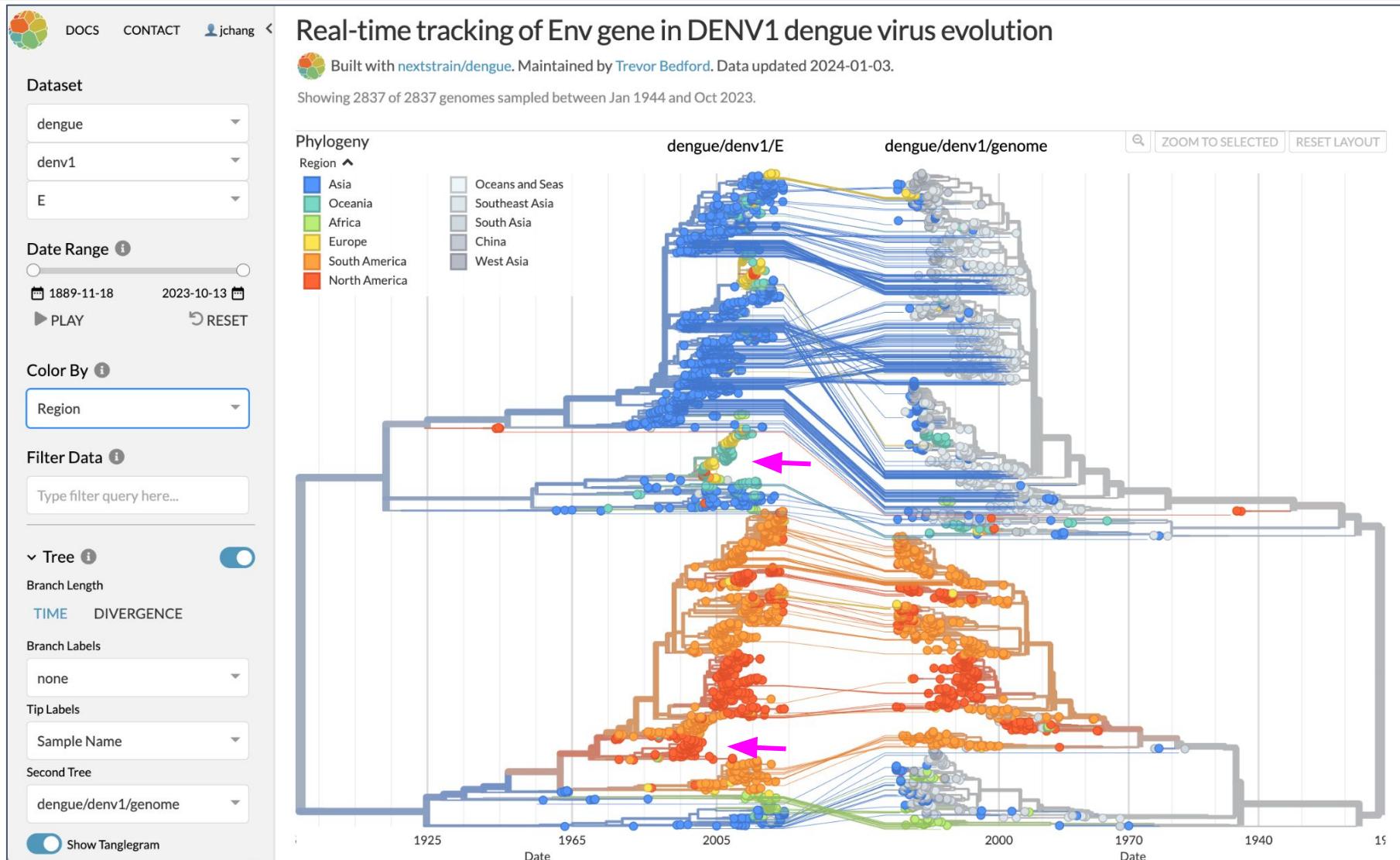
# "E" gene trees

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



# "E" gene trees

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



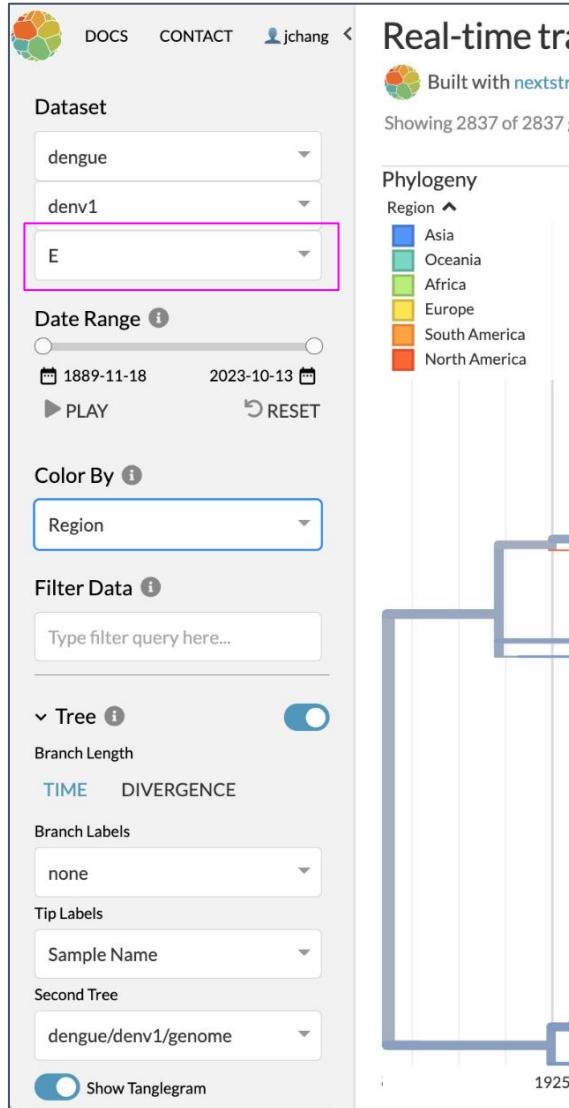
# Outline

- 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



Update manifest with dengue gene datasets #771

Merged j23414 merged 1 commit into master from dengue\_manifest on Jan 2

Conversation 4 Commits 1 Checks 3 Files changed 1 +36 -6 Review changes

Changes from all commits File filter Conversations Jump to

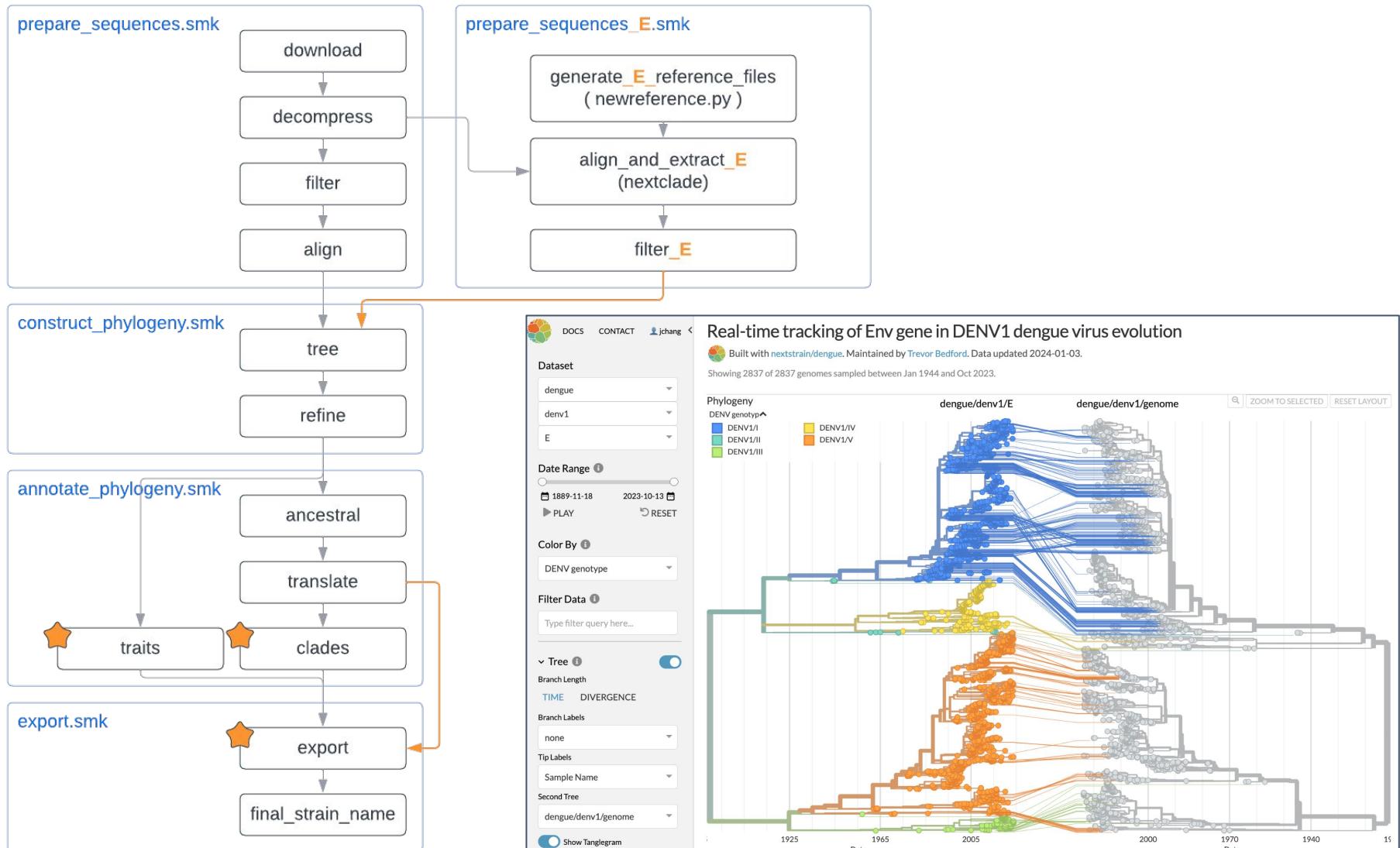
Update manifest with dengue gene datasets

master (#771) j23414 committed on Jan 2 commit 1dc25ac91c1e96d2fcc237765c8e19e36a863e9f

```
@@ -3,12 +3,42 @@
 3     "pathogen": {
 4       "dengue": {
 5         "serotype": {
 6           "-": "denv1": "",
 7           "-": "denv2": "",
 8           "-": "denv3": "",
 9           "-": "denv4": "",
10           "-": "all": "",
11           "-": "default": "denv1"
12         },
13         "+": "denv1": {
14           "+": "segment": {
15             "genome": ""
16           },
17           "+": "E": "",
18           "+": "default": "genome"
19         },
20         "+": "denv2": {
21           "+": "segment": {
22             "genome": ""
23           },
24           "+": "E": "",
25           "+": "default": "genome"
26         },
27         "+": "denv3": {
28           "+": "segment": {
29             "genome": ""
30           },
31           "+": "E": "",
32           "+": "default": "genome"
33         }
34       }
35     }
36   }
37 }
```

<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](#)



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.



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.