Intro to Phylogenetics Using Nextstrain

Day 4 of Training Workshop

Norman Hassell

Disclaimer

- The findings and conclusions in this presentation are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.
- Use of trade names and commercial sources is for identification only and does not imply endorsement by the U.S. Department of Health and Human Services.
- References to non-CDC sites on the Internet do not constitute or imply endorsement of these organizations or their programs by CDC or the U.S. Department of Health and Human Services. CDC is not responsible for the content of pages found at these sites.

Section 1: Overview

Getting Started

Before we get started, let's confirm your computer has `nextstrain` setup properly. Open a terminal and run the following command:

```
nextstrain setup --set-default conda
```

You should see the following output:

```
Checking setup…
✔ yes: operating system is supported
✔ yes: runtime data dir doesn't have spaces
✔ yes: runtime appears set up
✔ yes: snakemake is installed and runnable
✔ yes: augur is installed and runnable
✔ yes: auspice is installed and runnable

Setting default runtime to conda.

All good!  Set up of conda complete.
```

Getting Started (Cont.)

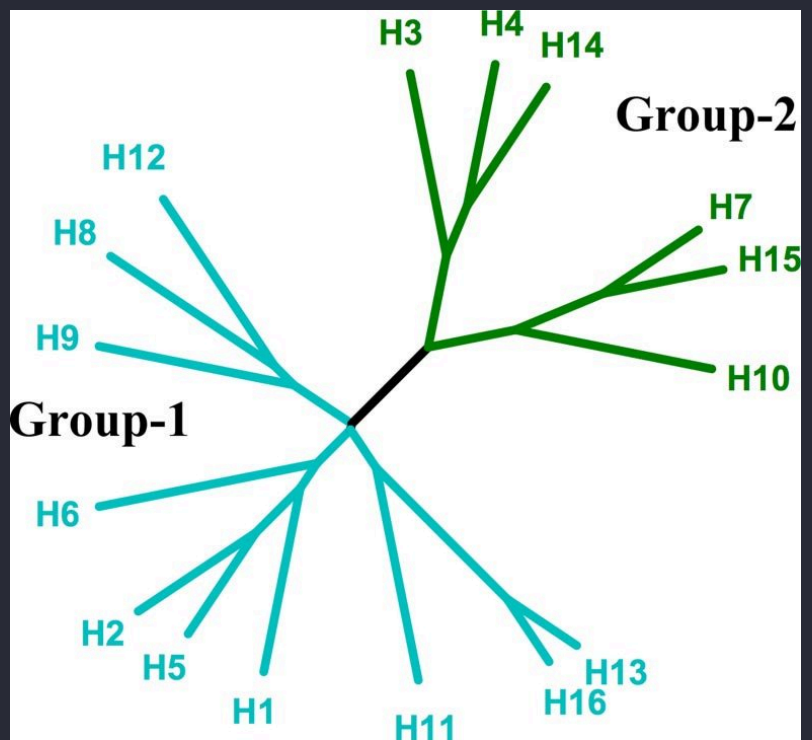
If you ran into an error when running the command, go to the following website.

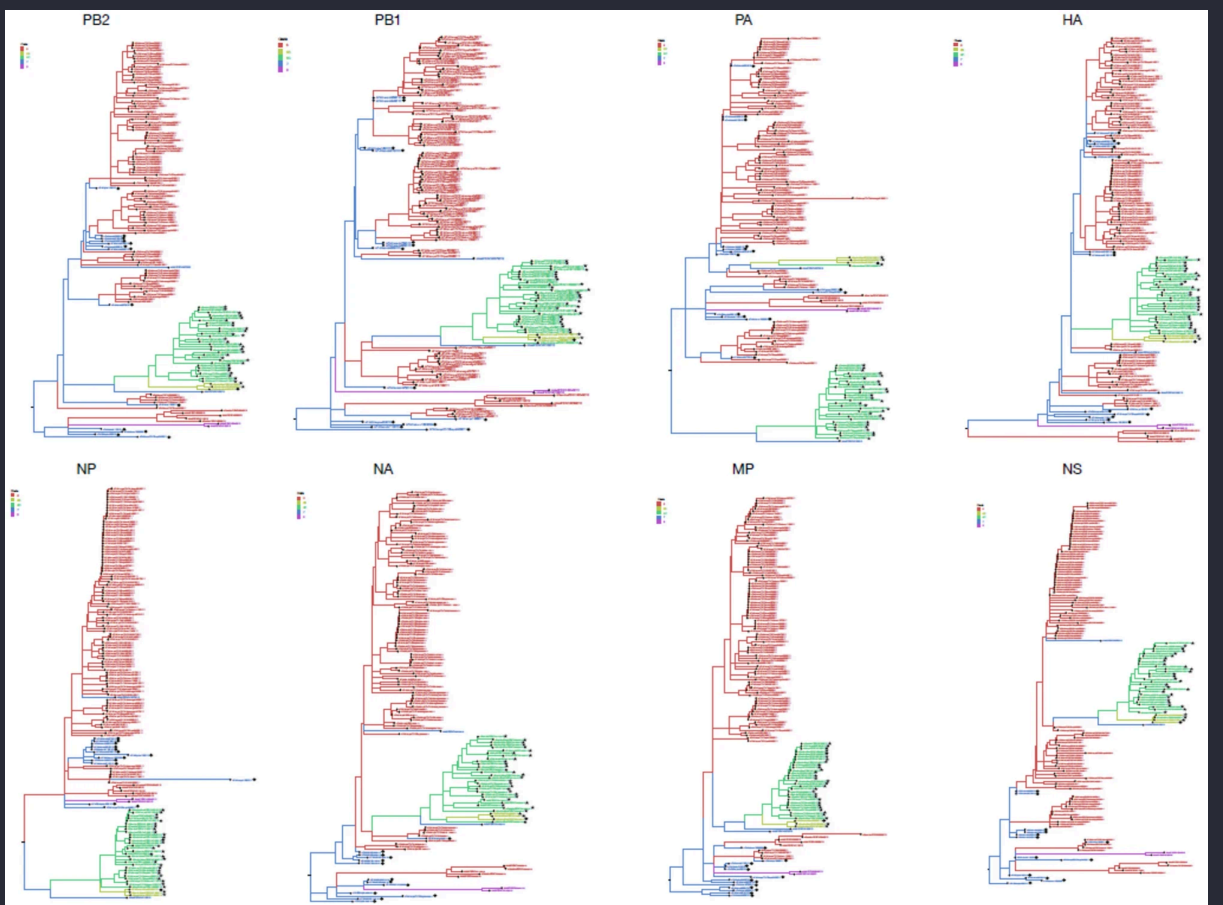
Install the `nextstrain-cli` using the instructions for your operating system.

For the “Set up a Nextstrain runtime” section, follow the “Conda” instructions for your operating system. We’ll go around to help if you are running into issues.

Intro to the Intro

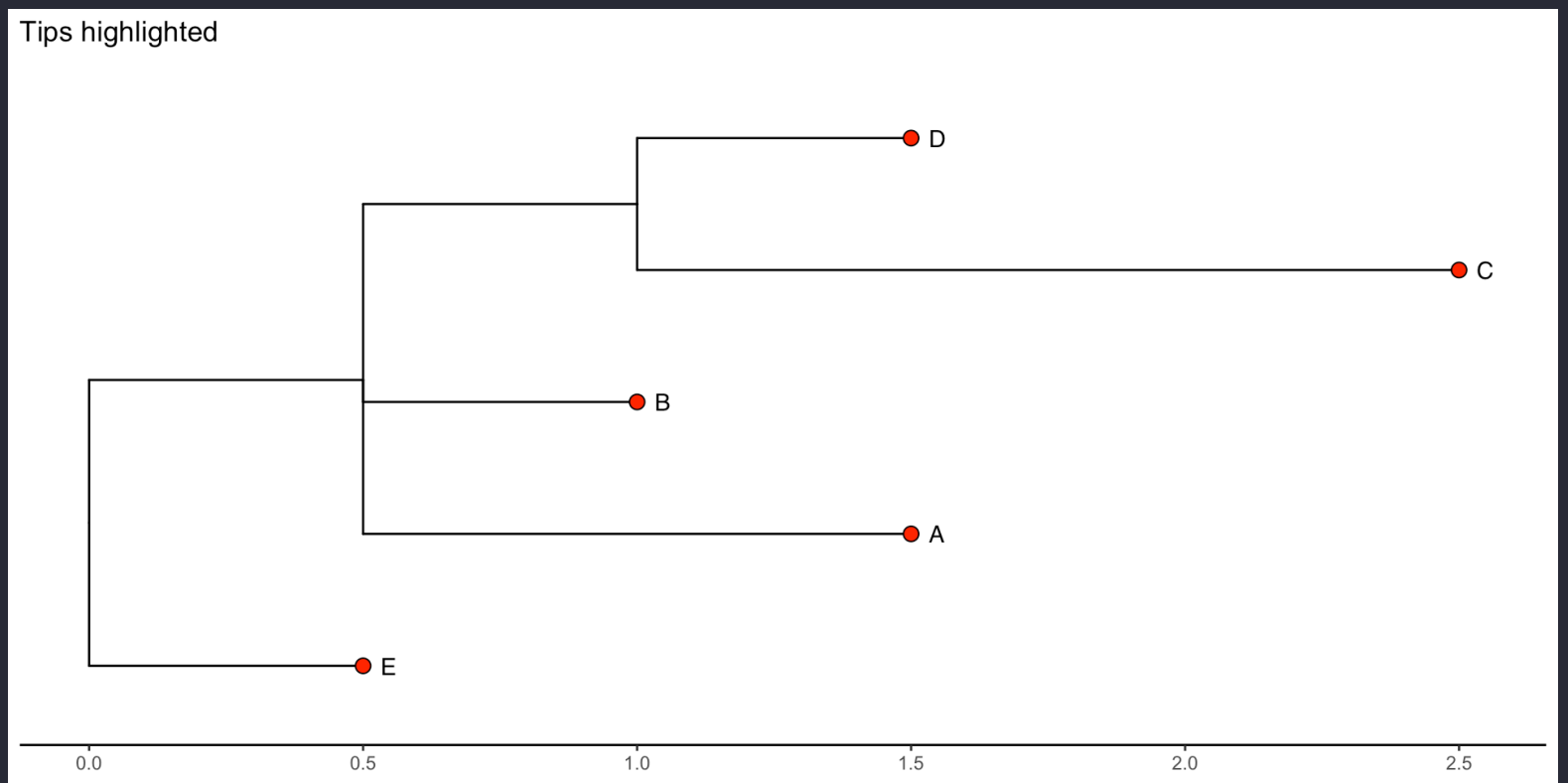
Phylogenetics is the study of evolutionary relationships among organisms.





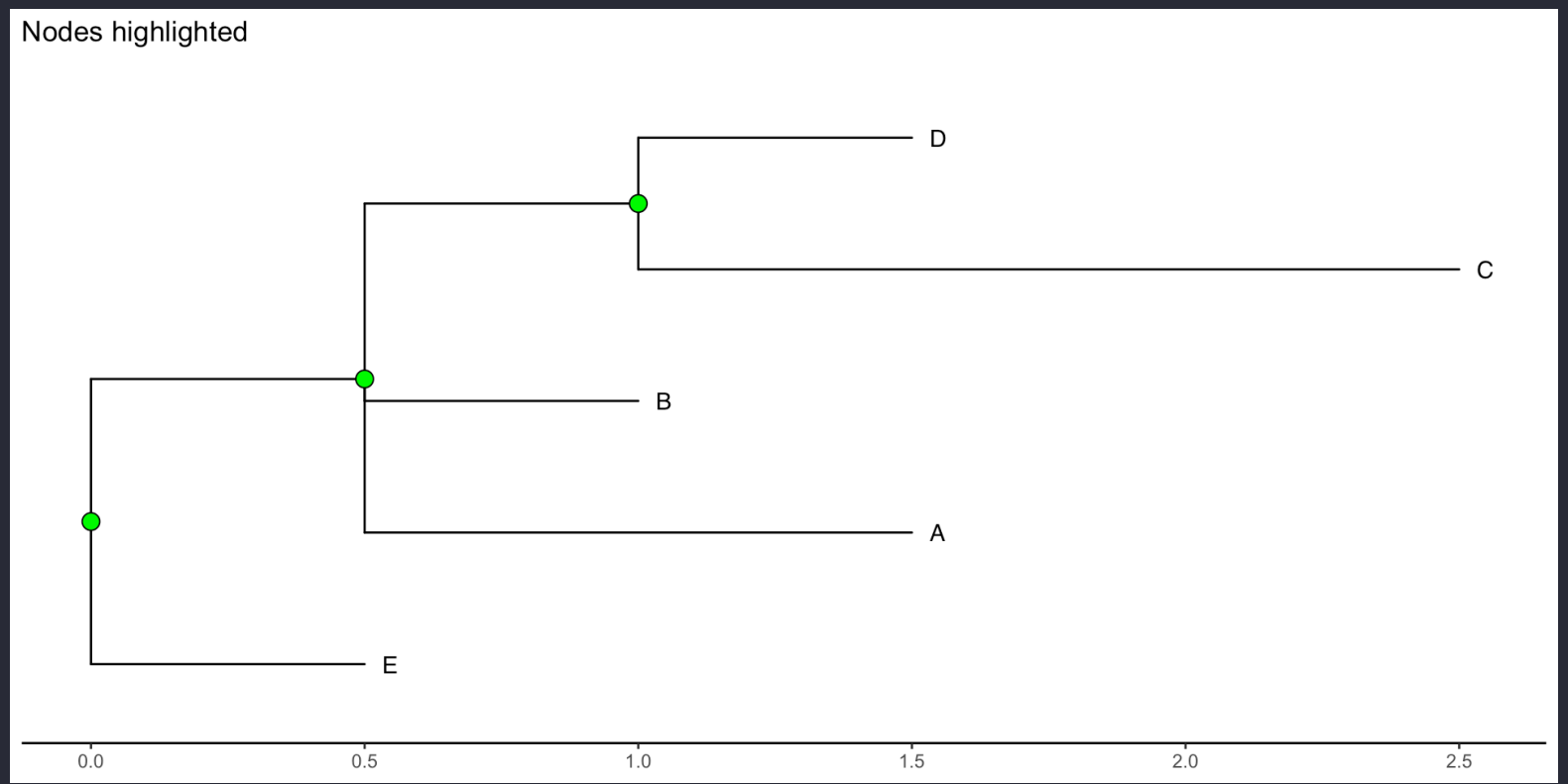
Basic Components of a Phylogenetic Tree (Tips)

- Code



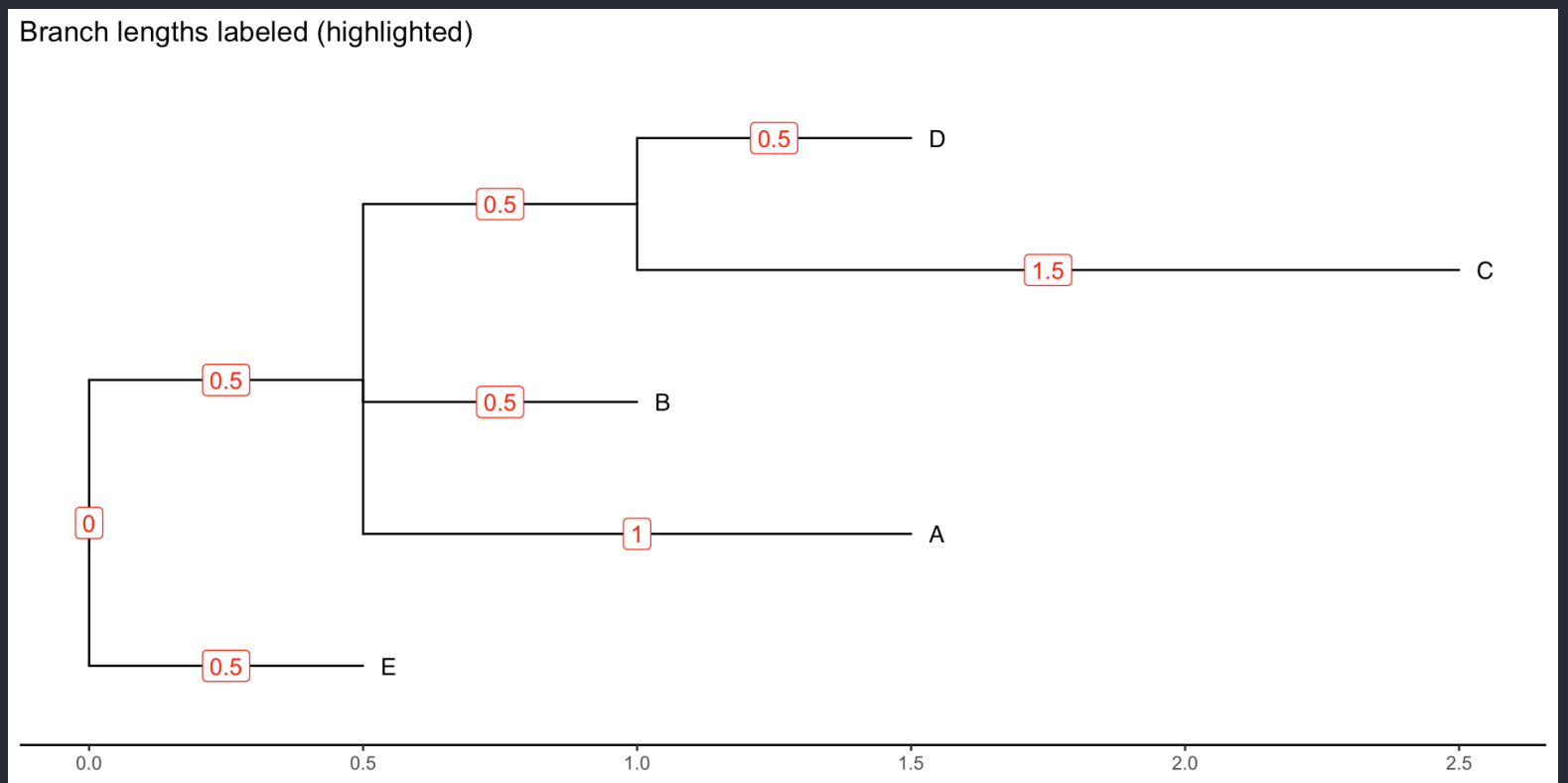
Basic Components of a Phylogenetic Tree (Nodes)

- Code



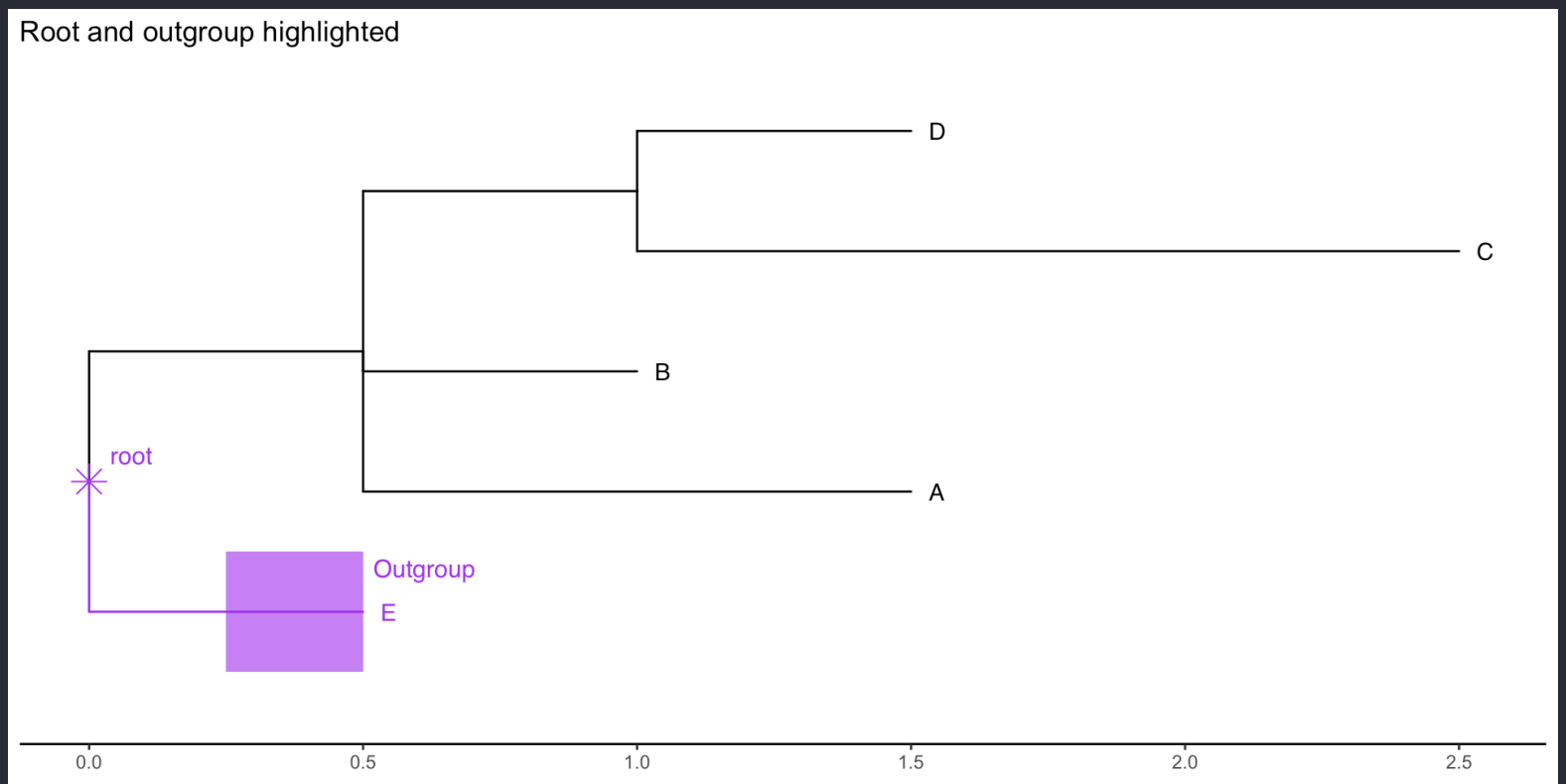
Basic Components of a Phylogenetic Tree (Branches)

- Code



Basic Components of a Phylogenetic Tree (Root)

- Code



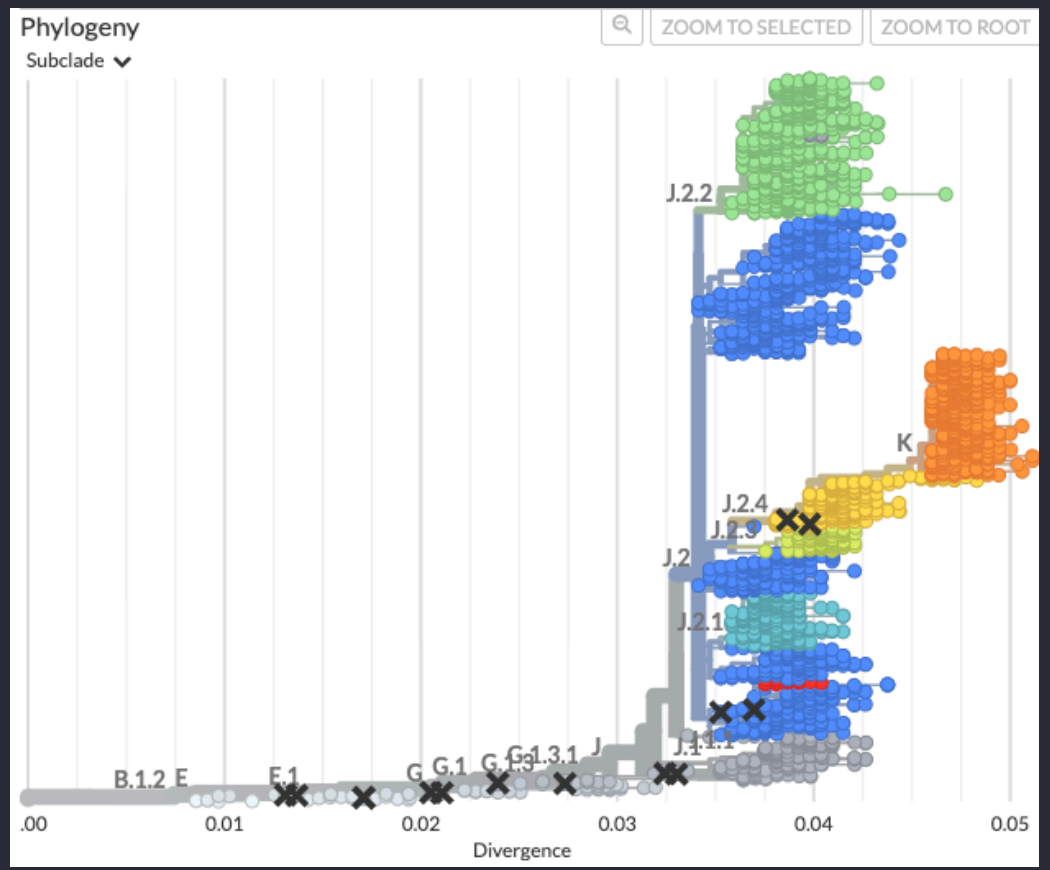
Notes on Tree Rooting

Conceptually, tree rooting should be done:

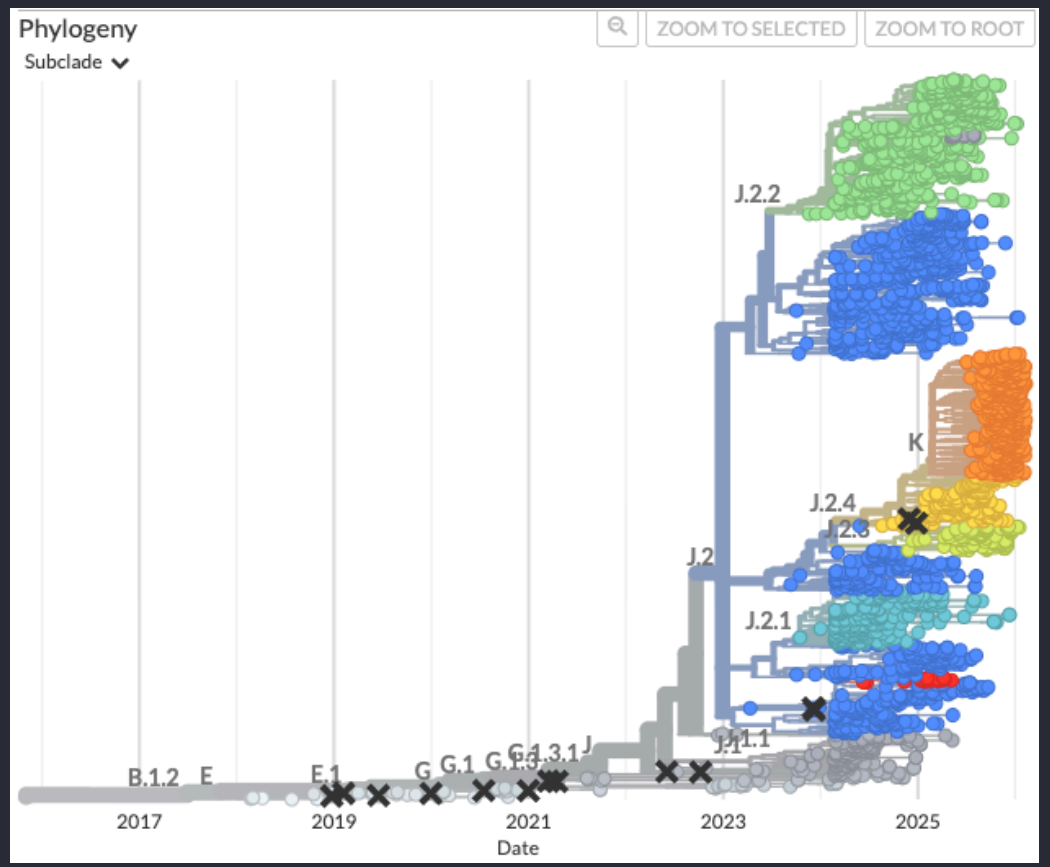
- Using a known outgroup (a taxon closely related to, but outside, the study group) to place the ancestor
- Or using midpoint rooting, which sets the root at the midpoint of the longest path between any two taxa

Outgroup rooting provides a more robust evolutionary history, and is generally preferred.

Types of Trees

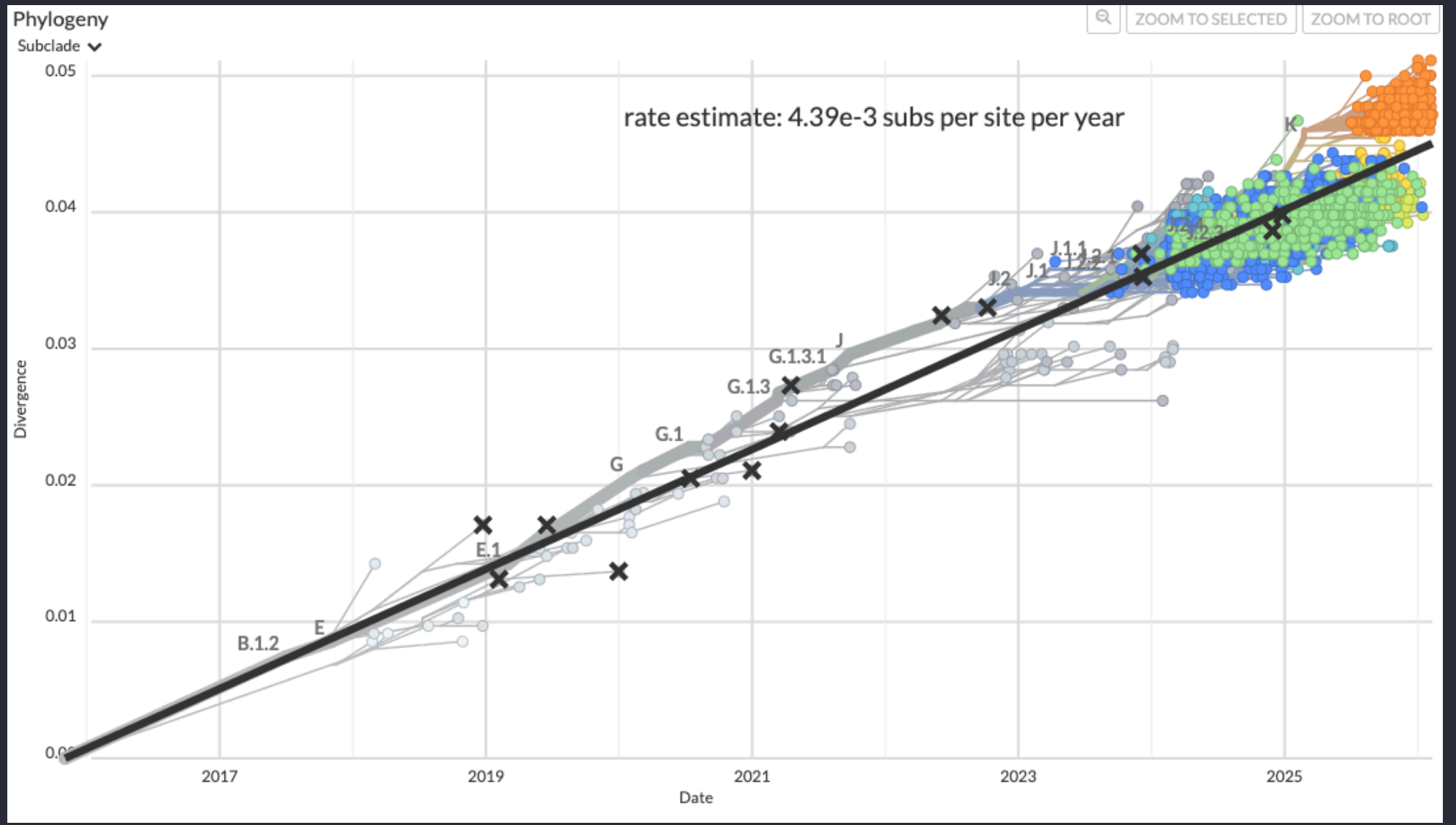


Nextstrain Tree in Divergence Display



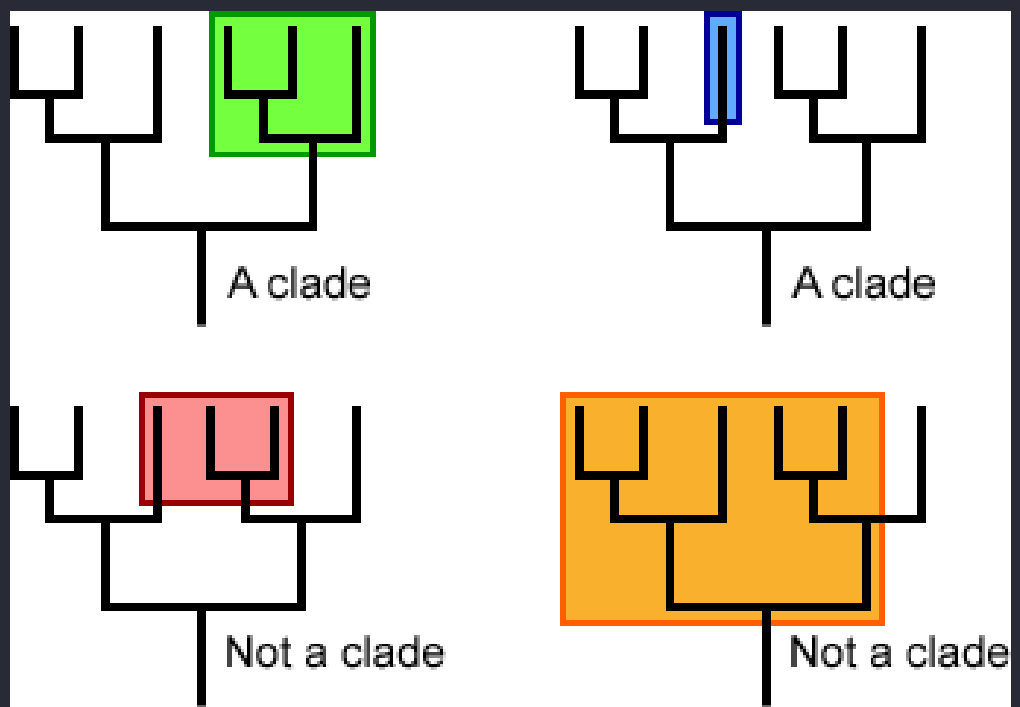
Nextstrain Tree in Time Display

Molecular Clock Concept



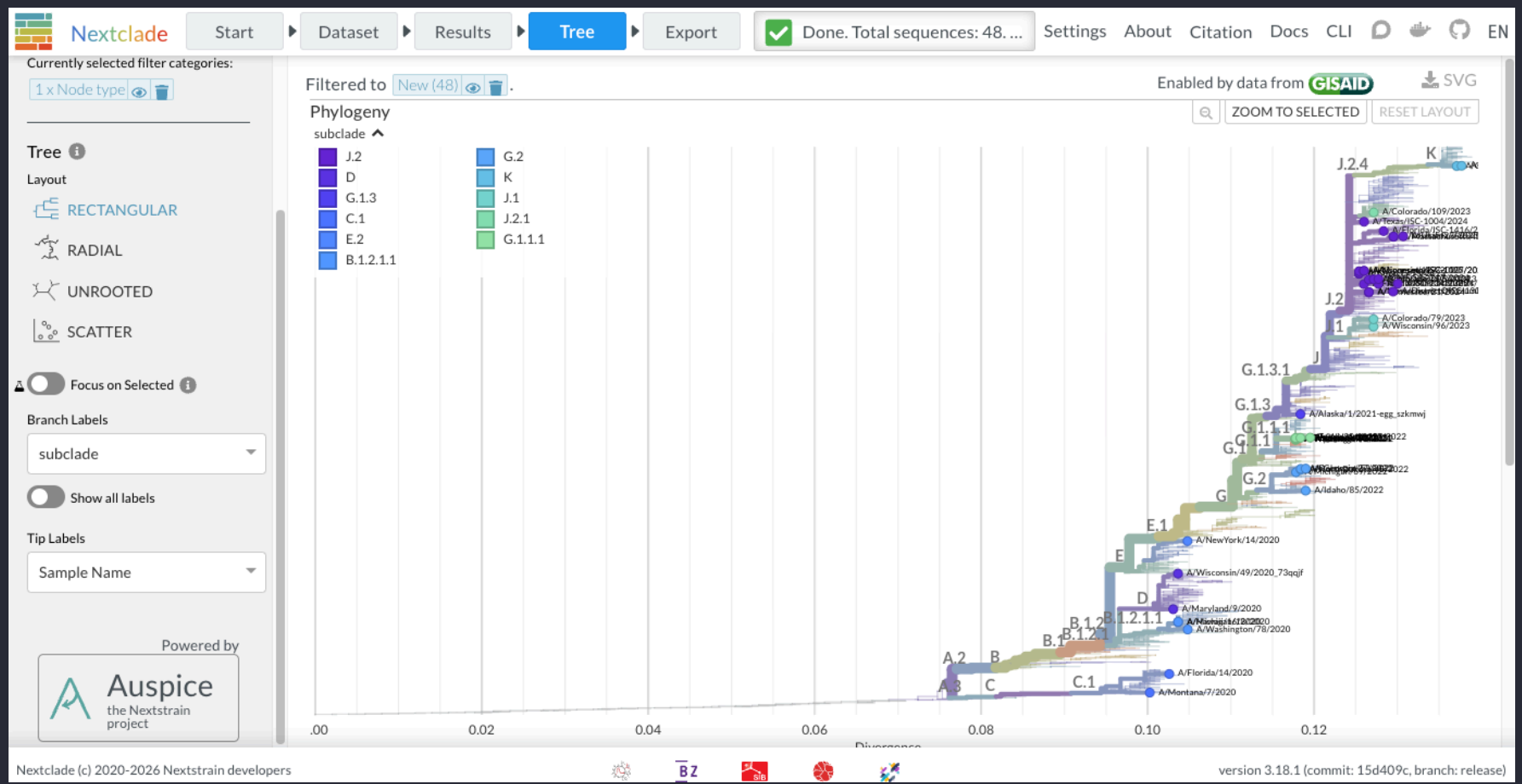
The Concept of Clades

Clades are the fundamental grouping of phylogenetic trees. They are defined as a monophyletic group of taxa in a tree that includes a single common ancestor and all its descendants.

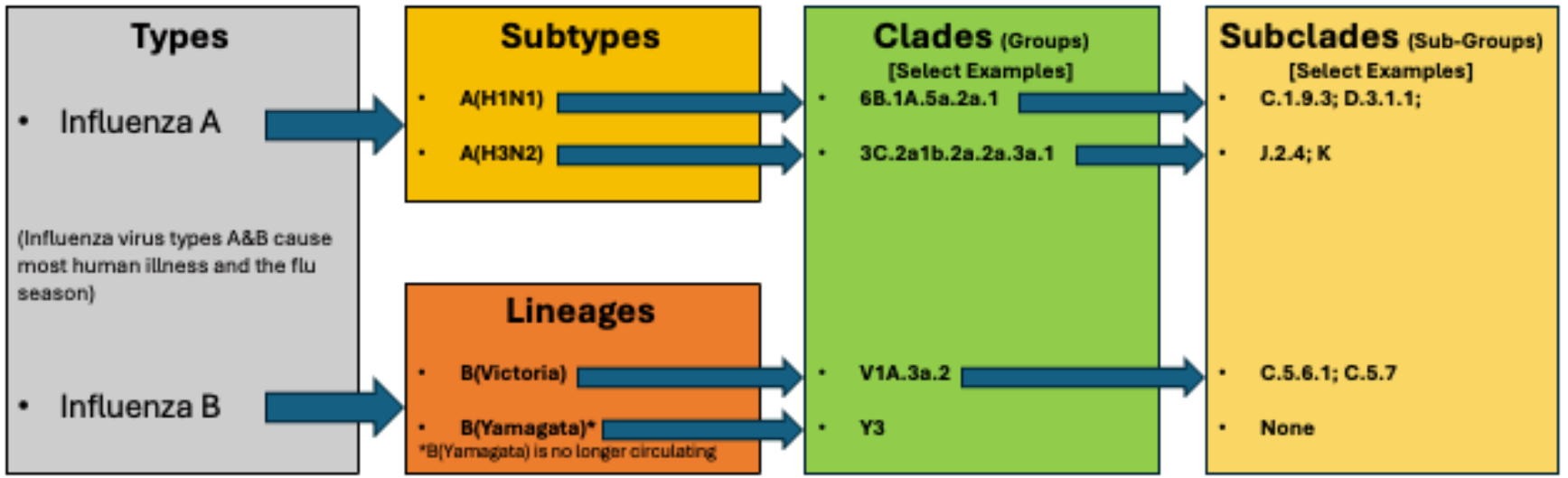


Clades in Seasonal Influenza

The seasonal flu research community recently implemented a nomenclature and clade proposal system for tracking important emerging clades. These correspond to the "subclade" designations from Nextclade outputs.



Clades in Seasonal Influenza (cont.)



Clades in Seasonal Influenza (cont.)

Defining a subclade in Nextstrain (`augur clades` command):

clade	gene	site	alt
A	HA1	45	N
A	HA1	48	I
A	nuc	473	T

clade	gene	site	alt
K	clade	J.2.4	
K	HA1	2	N
K	HA1	144	N
K	HA1	158	D
K	HA1	160	K
K	HA1	173	R

Clades in Seasonal Influenza (cont.)

Subclade K example:

Subclades vs. Lineages (Pango 🤦)

Influenza subclades are **not** equivalent to Pango (SARS-CoV-2) lineages.

Pango lineage designations only requirement is a group of viruses having a shared ancestry.

Lineage proposals are through public submission

This has led to the designation of 5.7K+ lineages. 🤯

Subclade Proposal Process

Subclade proposal criteria:

1. Size: large groups should have a higher priority for designation.
2. Divergence: the more mutations have accumulated relative to the break point of the parent clade, the higher the priority of a novel clade
3. Specific mutations: Ideally, breakpoints sit on long branches with significant mutation. Such mutations will be better defined for well studied segments/genomes.

Each of these components feeds into the calculation of normalized branch scores across a phylogenetic tree.

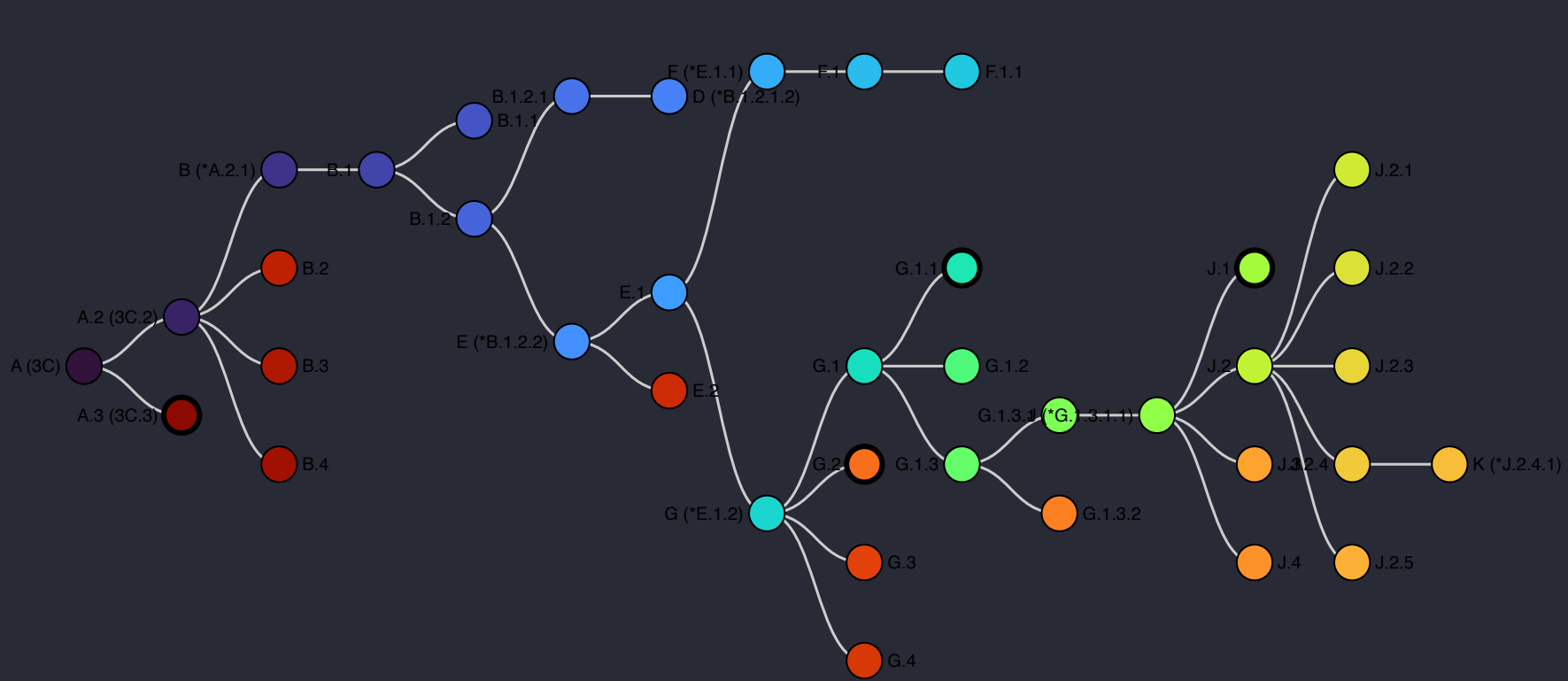
A threshold score of 1.0 is set.

Subclade Naming

- The system of subclade naming uses a variation of the Pango nomenclature system.[1]
- Capitol letters are used as aliases and numbers separated by periods for the retained part of the hierarchical name.
- Shortening (aliasing) of names after three hierarchical levels (i.e. D for C.1.1.1).
- However, a new alias may be given earlier in the case of a rapid expansion or special designation.

Example of Implementation Using H3N2 HA

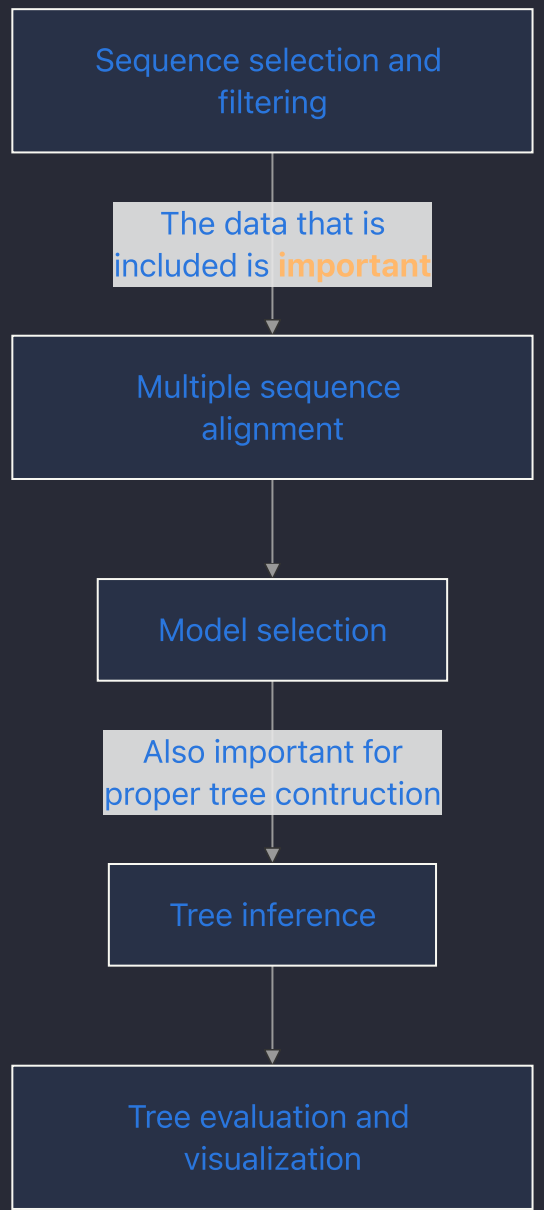
- Code



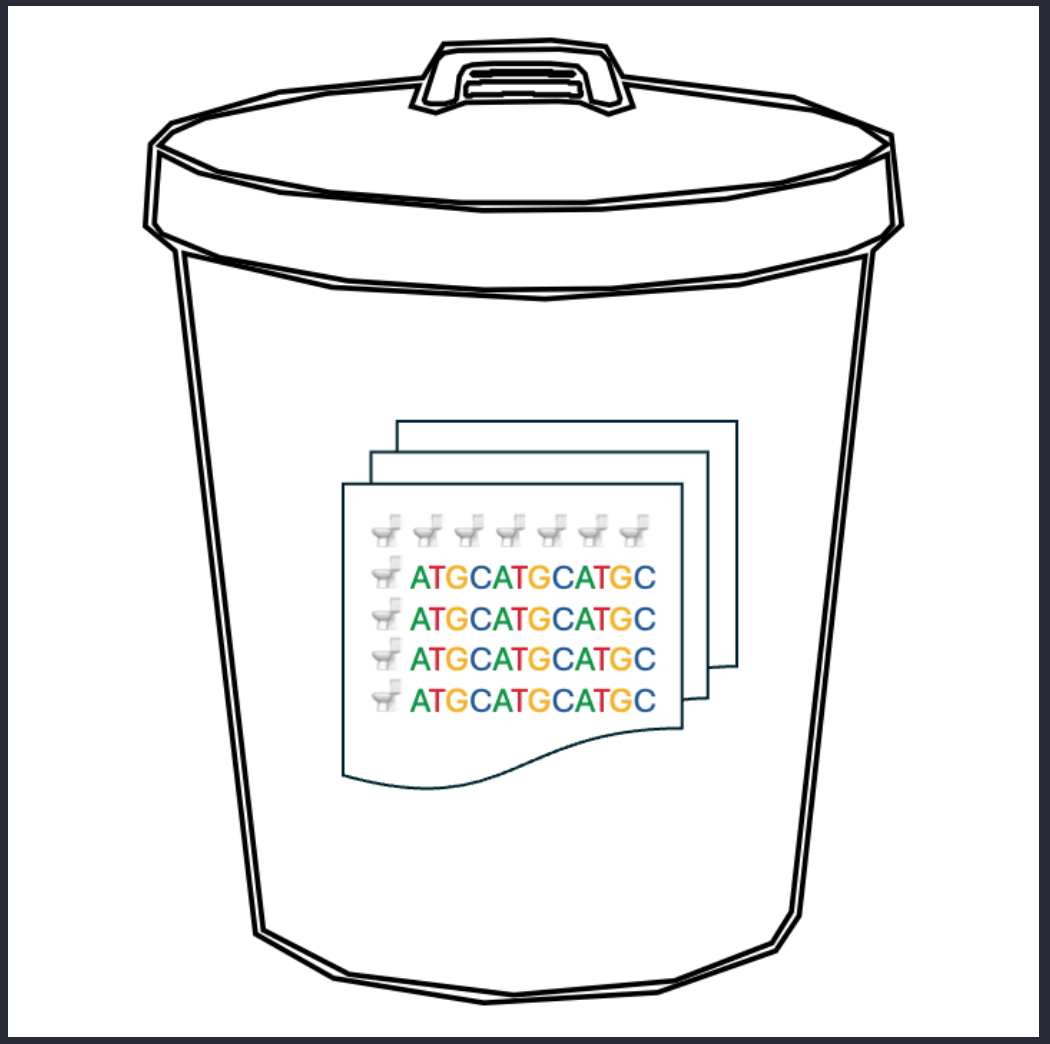
Section 2: Tree Building Overview

Quick break.

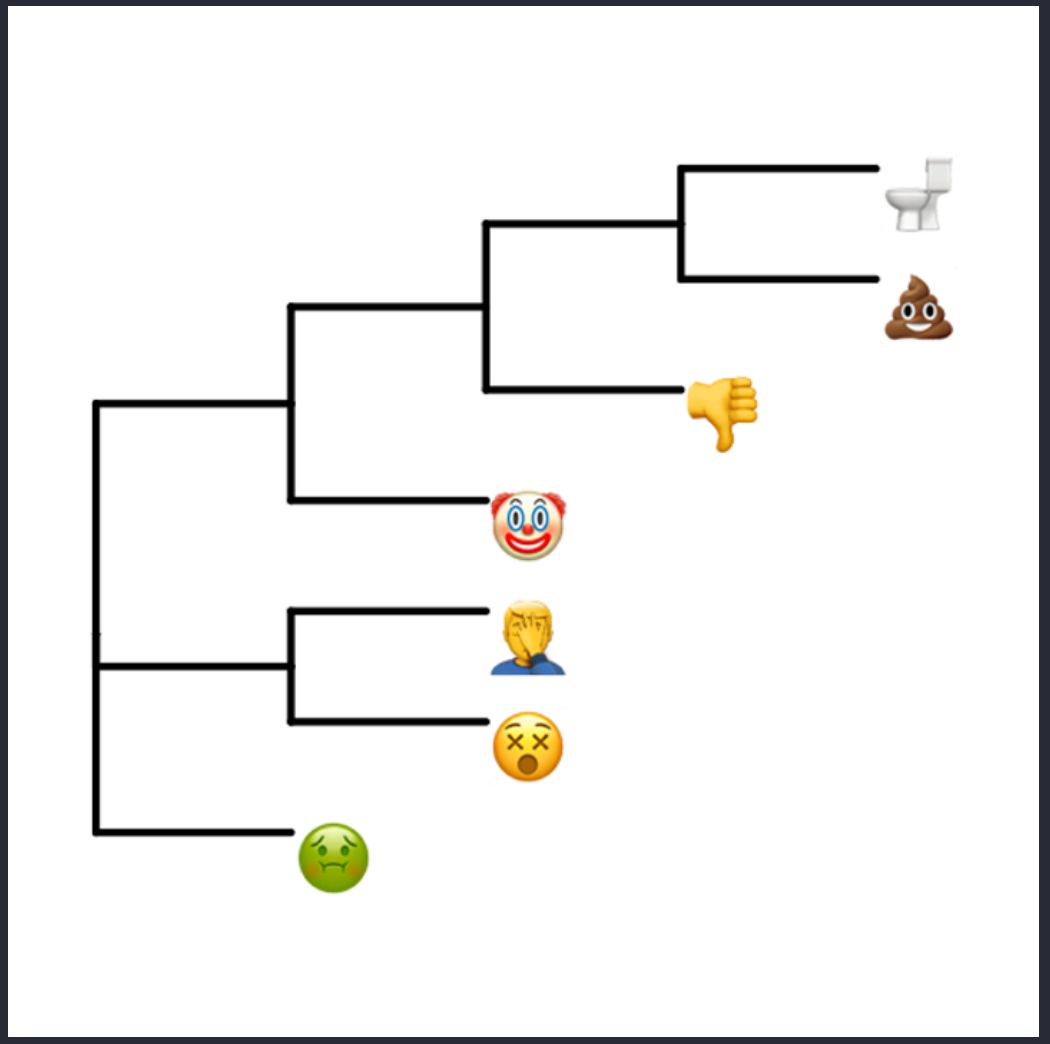
Phylogenetic Tree Building Steps



Sequence Selection



Garbage in



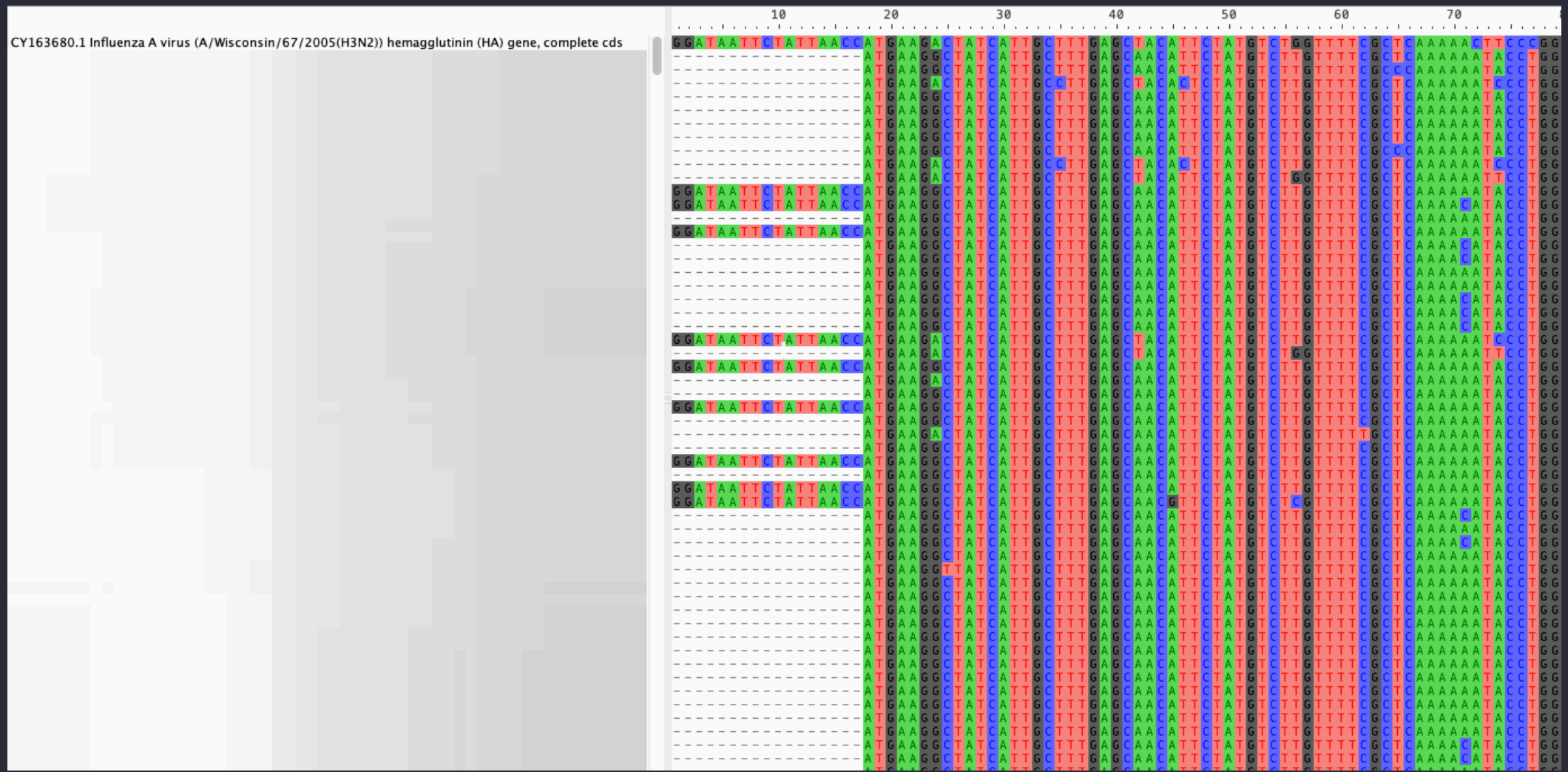
Garbage out

Sequence Selection

When starting the process of building a tree:

- Consensus sequence data quality (not just NGS coverage)
- What is the question you are trying to address?
  - Global surveillence?
  - Targeted regional analysis?
  - Country specific questions?
  - Timing of a newly emerged variant?
    - *Be especially careful here. This requires intensive analysis and scrutiny. Understand the limitations of analysis.*

Multiple Sequence Alignment (MSA)



MSA Algorithms

Examples of MSA software:

- CLUSTAL Omega
- MAFFT
- MUSCLE
- **nextclade** aligner (previously nextalign)

Model Selection

Assuming nucleotide sequence data as input:

- Jukes-Cantor (simplest model)[1]
- HKY
- TN93
- GTR (Model implemented in Nextstrain flu pipelines)

Tree Inference - Algorithms

Distance Based:

- Neighbor joining
- UPGMA

Character Based:

- Maximum Parsimony
- Maximum Likelihood (used in Nextstrain pipelines)
- Bayesian Inference

Tree Inference - Software

- MEGA
- RAxML
- FastTree
- IQ-TREE (Nextstrain Default)
- cmaple (Standalone program and within IQ-TREE using `--pathogen` flag)

Tree Evaluation and Visualization

- Nextstrain (auspice)
- auspice.us (browser based auspice)
- IcyTree (browser based)
- FigTree
- ITOL
- Dendroscope
- archaeopteryx
- Programmatic libraries:
  - ETE Toolkit
  - ggtree

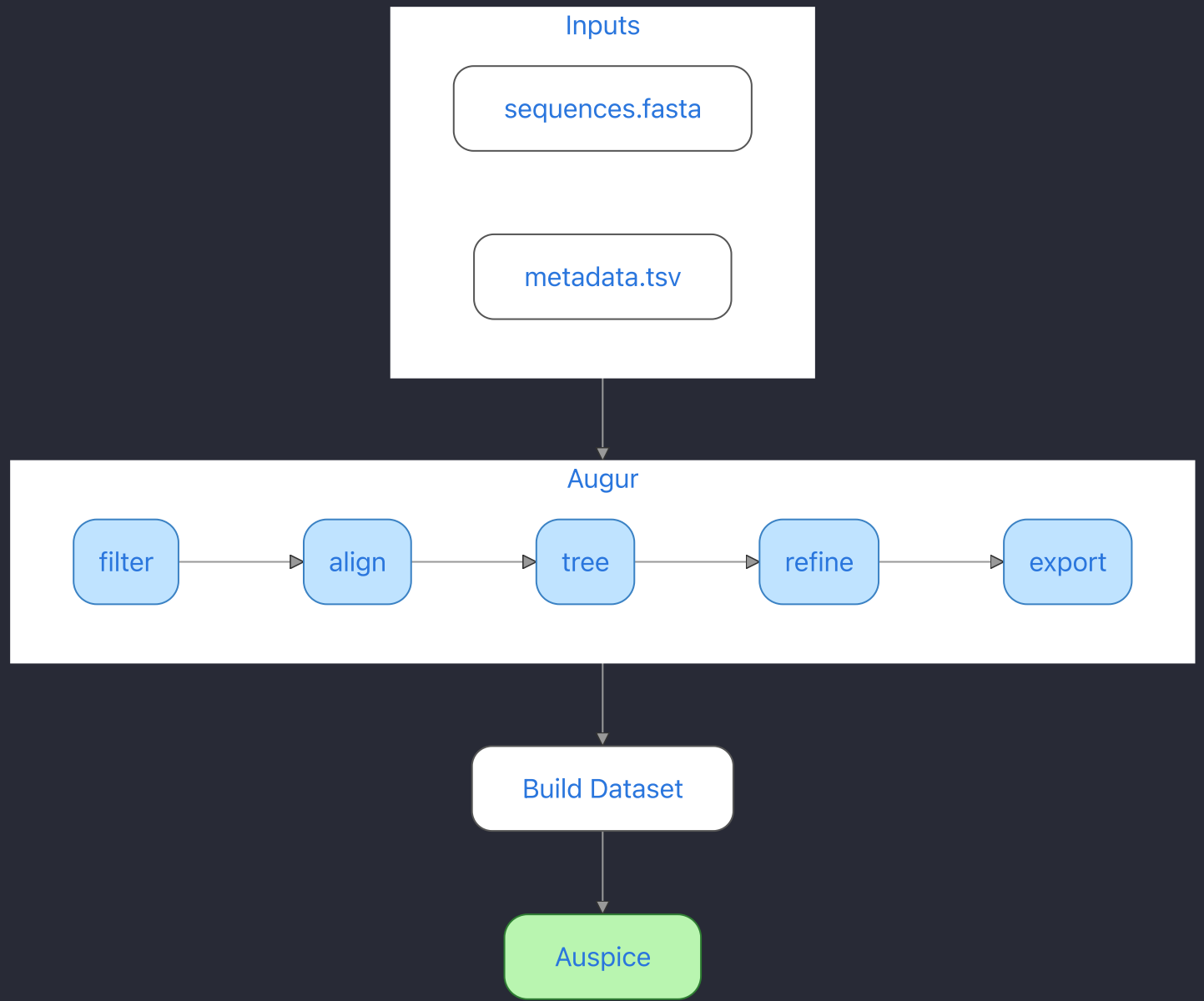
Intro to Nextstrain

Nextstrain is an open source toolkit for analyzing and visualizing pathogen genomic data.

The two core parts of Nextstrain are:

- Augur (modular bioinformatics toolkit)
- Auspice (browser based vizualization)

Nextstrain process overview



Section 3: Nextstrain Build Exercise

Click on the following link which will direct you to the exercise we will be going through together.

Section 4: Overview of Nextstrain Features and Navigation

We'll go over features of your build and navigate it together.

Section 5: Communicating Results Using Nextstrain Narratives

Introduction to Nextstrain Narratives

Nextstrain narratives are markdown (.md) files that allow the communication of results interactively in a slide-like format.

Click on the link above for detailed documentation on narratives.

For an example of a narrative, take a look at the Twenty years of West Nile virus narrative.

Components of a Narrative: Header

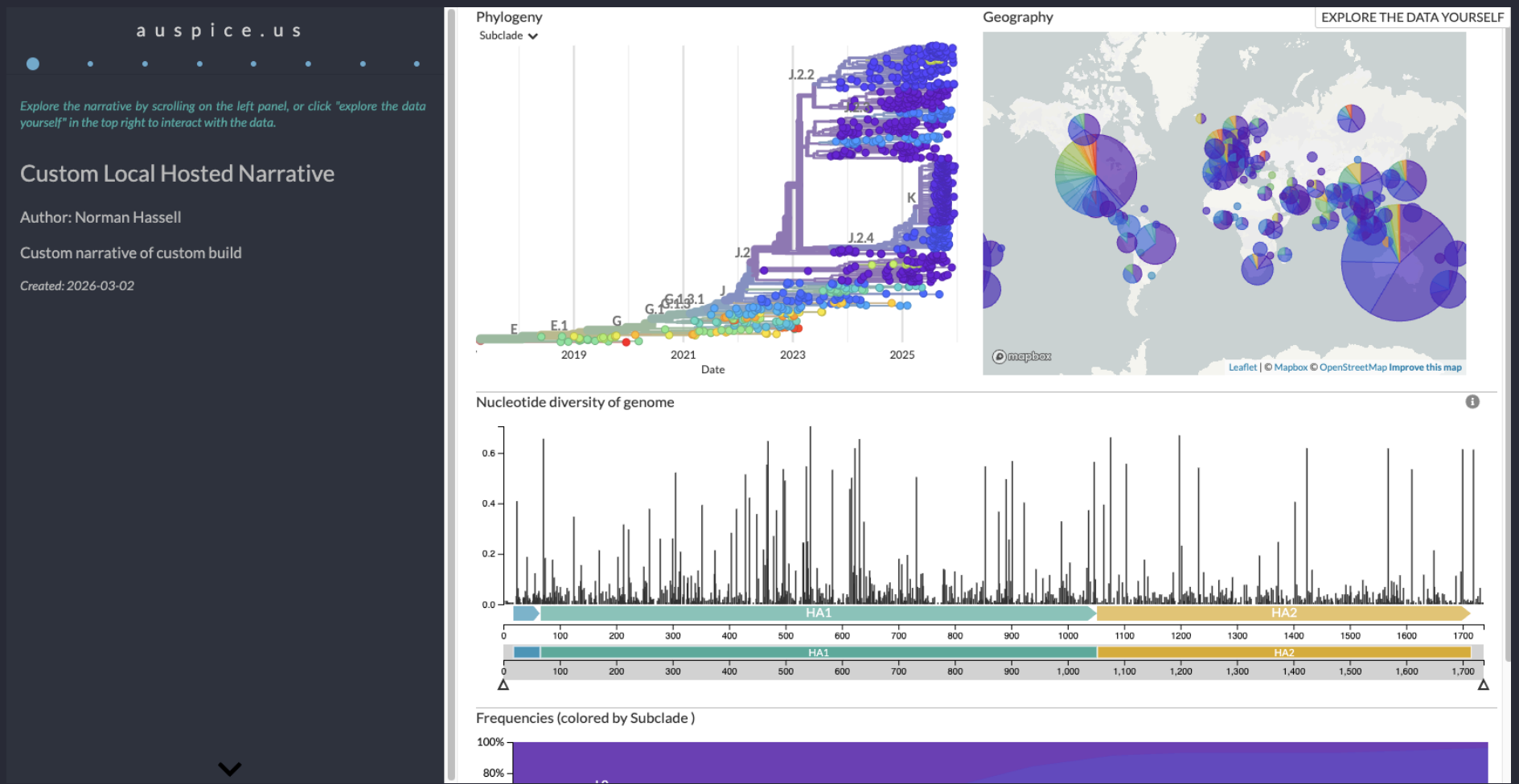
Open the example narrative file `/profiles/gisaid/custom.md` in VS Code.

At the top of the file is the narrative header:

```
---
title: Custom Local Hosted Narrative
authors: Norman Hassell
date: "2026-03-02"
dataset: "https://nextstrain.org/custom/h3n2/ha"
abstract: "
Custom narrative of custom build
"
---
```

For local builds, you will have the prefix `https://nextstrain.org` followed by your build file name with underscores replaced by the `/` character (ie. `custom_h3n2_ha.json` dataset becomes `https://nextstrain.org/custom/h3n2/ha`)

Components of a Narrative: Header (Result)



Components of a Narrative: Basic Slides

Basic slides are set up by providing copied links as first level headers:

```
# [Custom Dataset Slide](https://nextstrain.org/custom/h3n2/ha?d=tree,map,f

Example of a basic slide view.

The slide is filtered to a "Sample Type" of "sample".
```

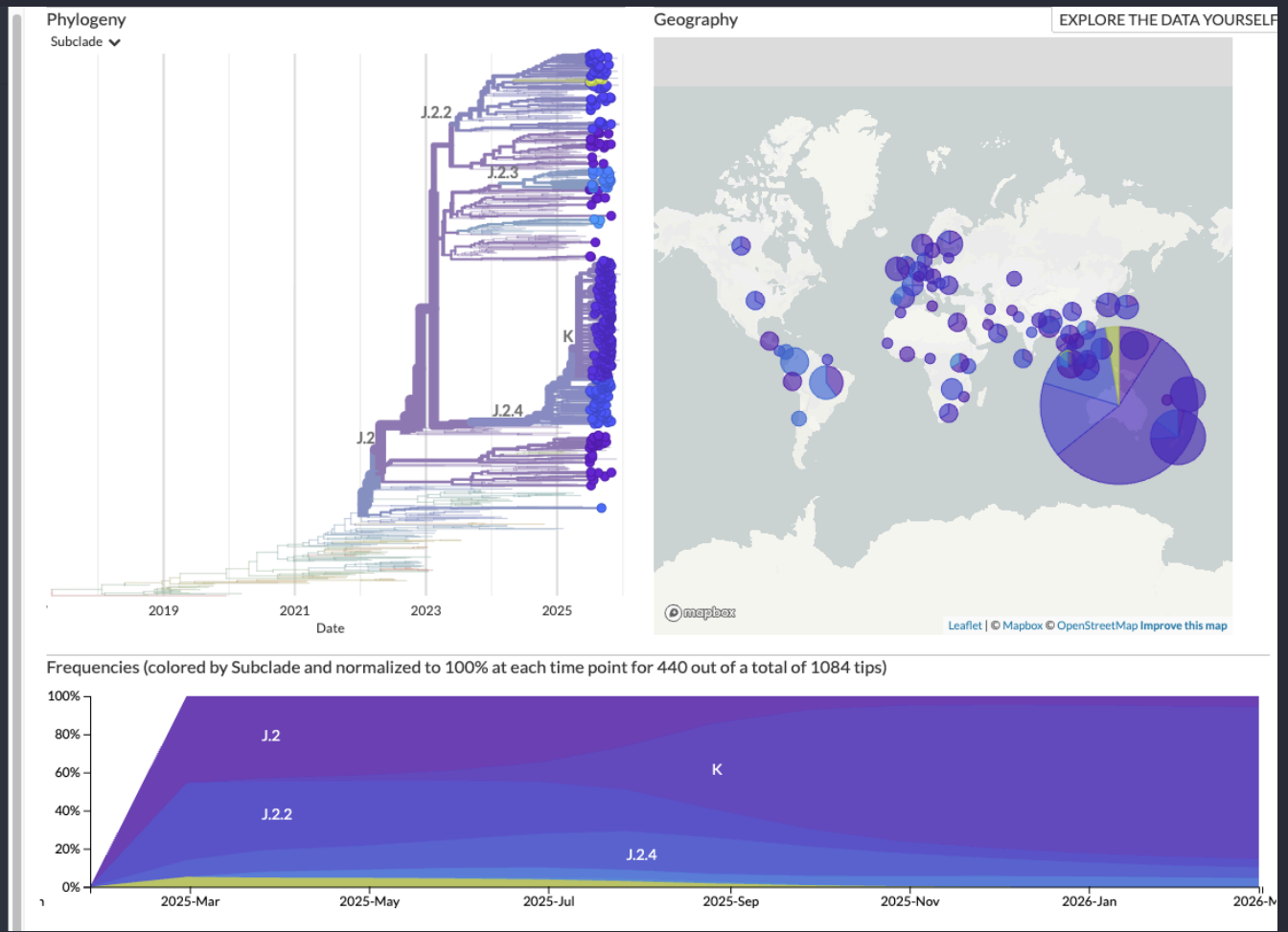
Components of a Narrative: Basic Slides (Result)

auspice.us

Custom Dataset Slide

Example of a basic slide view.

The slide is filtered to a "Sample Type" of "sample".



Components of a Narrative: Animations

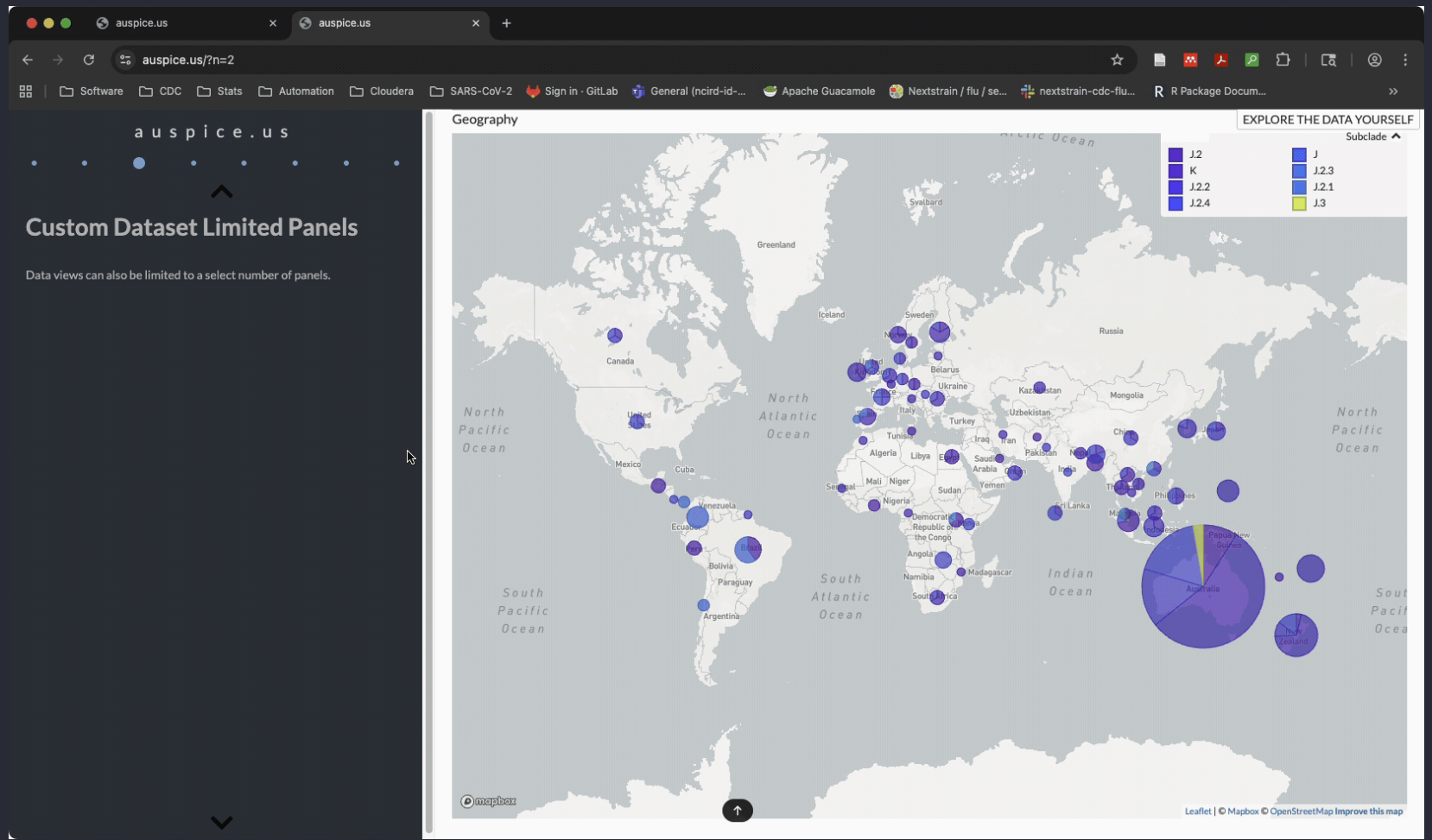
Timeline animations can be used in narratives by copying the link associated with the animation.

The link will briefly show after clicking the “Play ▶️” button underneath the “Date Range” toolbar.

```
# [Custom Dataset Animations](https://nextstrain.org/custom/h3n2/ha?animate

Animations can be captured.
```

Components of a Narrative: Animations (Result)



Components of a Narrative: Images

Images can be imbedded into the narrative sidebar or main display from direct links or base64 conversion.

```
# [Image Embedding](https://nextstrain.org/custom/h3n2/ha)

Images can be embedded from the web in the side panel.

![](https://www.esrf.fr/files/live/sites/www/files/news/general-old/general

```auspiceMainDisplayMarkdown
### Or in the main body display via the `auspiceMainDisplayMarkdown` sectio

For main body markdown display.

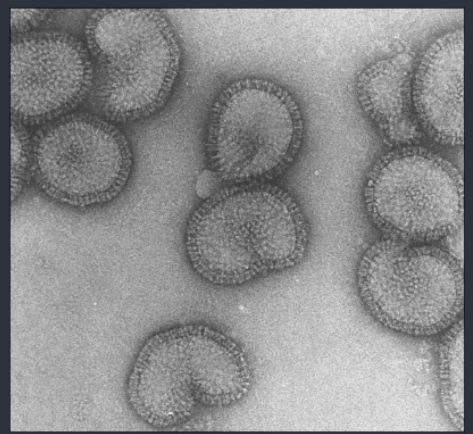
![Flu](https://www.embl.org/news/wp-content/uploads/2020/04/INFLUENZA_EDIT.
```
```

Components of a Narrative: Images (Results)

auspice.us

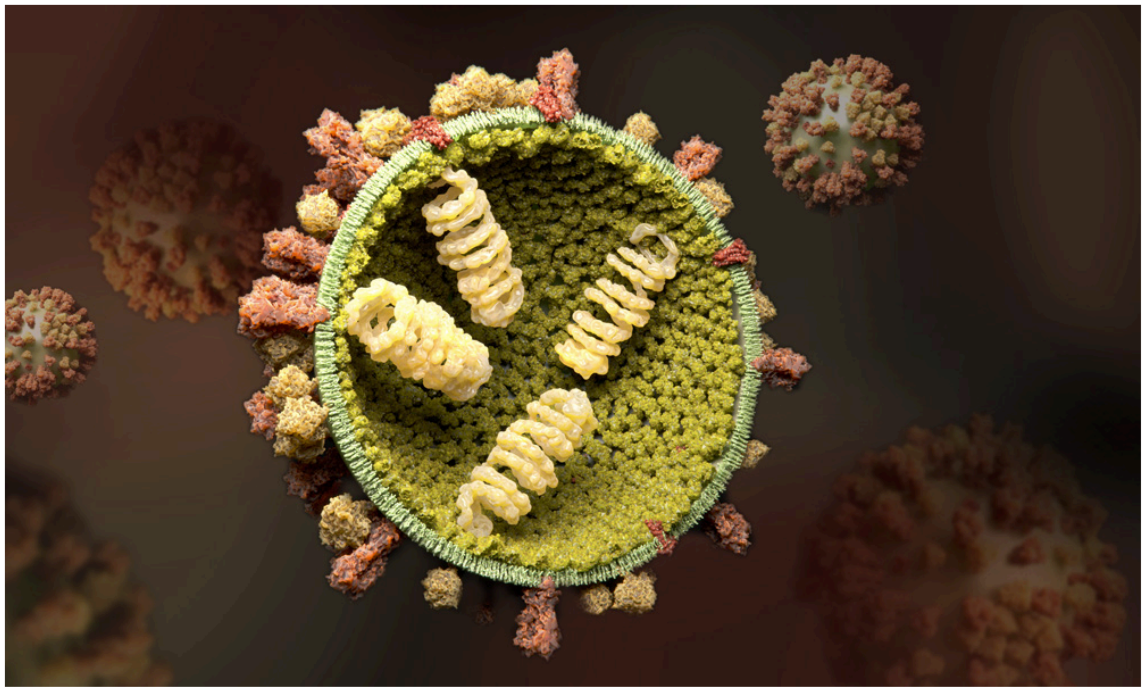
Image Embedding

Images can be embedded from the web in the side panel.



Or in the main body display via the `auspiceMainDisplayMarkdown` section.

For main body markdown display.



Narratives Continued

Normal features of markdown such as:

- Tables
- Lists
- Text formatting
- HTML embedding

Can all be used in both the narrative side panel and main display.

Again, the main display requires the `auspiceMainDisplayMarkdown` block for use.

Narrative Exercise

Before moving on to the exercise open a new terminal in the `seasonal-flu-demo` folder and enter into a new nextstrain shell session:

```
nextstrain shell .
```

Then, run another build using the following command:

```
nextstrain build . --configfile profiles/gisaid/custom_gisaid_global.yaml
```

Repeat this process for a third build.

Open a new terminal, enter a nextstrain shell, and execute the following:

```
nextstrain build . --configfile profiles/gisaid/custom_gisaid_norefs.yaml
```

Narrative Exercise

Develop a narrative walking through your current build focusing on viruses circulating in Oceania/Australia (`custom_h3n2_ha`).

- What were the most prominent subclades for samples?
- What subclade is exhibiting the highest local branching index (LBI)?
- From the identified subclade with the highest LBI, what is the estimated emergence date and range?
- What regions is this subclade present in?
- What is the rate calculation for the tree overall vs. the rate for the subclade?
- Are there any reagent viruses in this subclade? Highlight them.
- Add another narrative slide of your choice, highlighting something you've found.
- Close out your narrative with a summary slide.

Section 6: Build Comparisons

We'll take some time to compare the three different builds we've created.