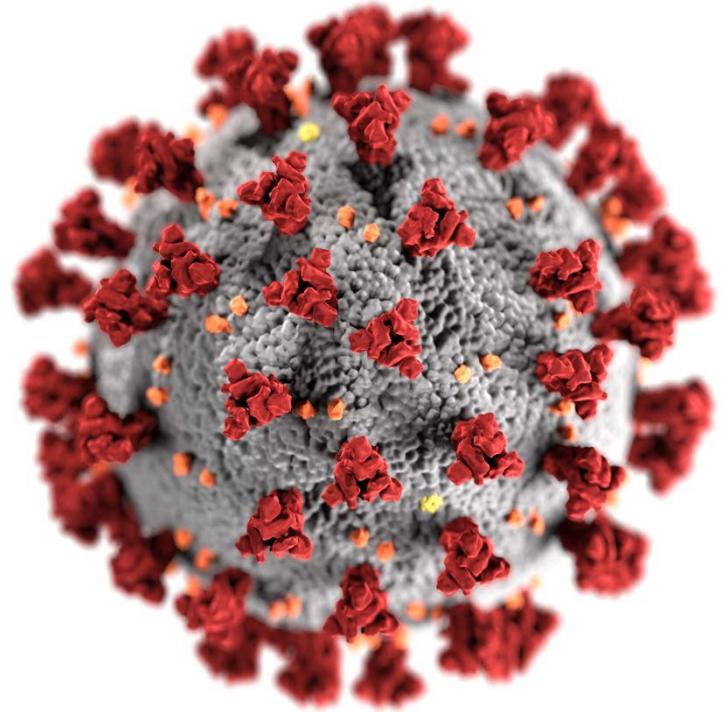


Getting started with Nextstrain

COVID-19 Genomic Epidemiology Toolkit: Module 3.1

Michael Weigand, PhD
Bioinformatician
Centers for Disease Control and Prevention



cdc.gov/coronavirus

Toolkit map

Part 1: Introduction

- 1.1 What is genomic epidemiology?
- 1.2 The SARS-CoV-2 genome
- 1.3 How to read phylogenetic trees

Part 2: Case Studies

- 2.1 SARS-CoV-2 sequencing in Arizona
- 2.2 Healthcare cluster transmission
- 2.3 Community Transmission

Part 3: Implementation

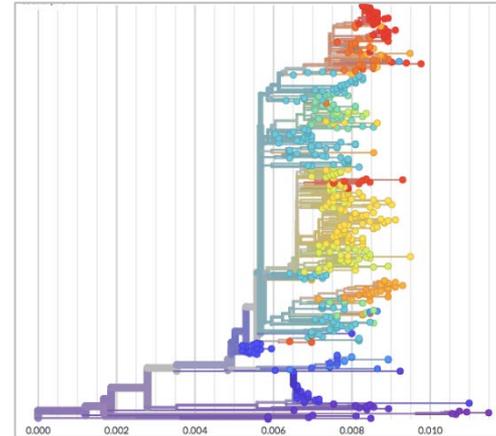
3.1 Getting started with Nextstrain

- 3.2 Getting started with MicrobeTrace
- 3.3 Linking epidemiologic data



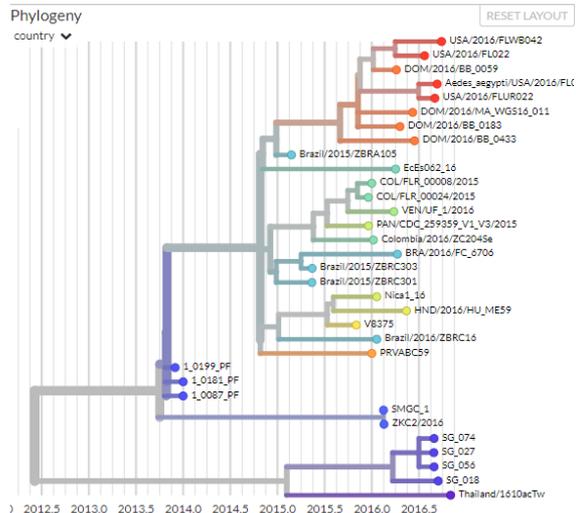
What is Nextstrain?

- Open-source project to harness the power of pathogen genome data
 - Powerful analytics and interactive visualizations
 - Designed to aid epidemiological understanding, improve outbreak response, and provide real-time snapshots of evolving pathogen populations
 - Learn (a lot) more at
 - <https://nextstrain.org>
 - <https://docs.nextstrain.org>
-  [@nextstrain](https://twitter.com/nextstrain)



Nextstrain: Default view

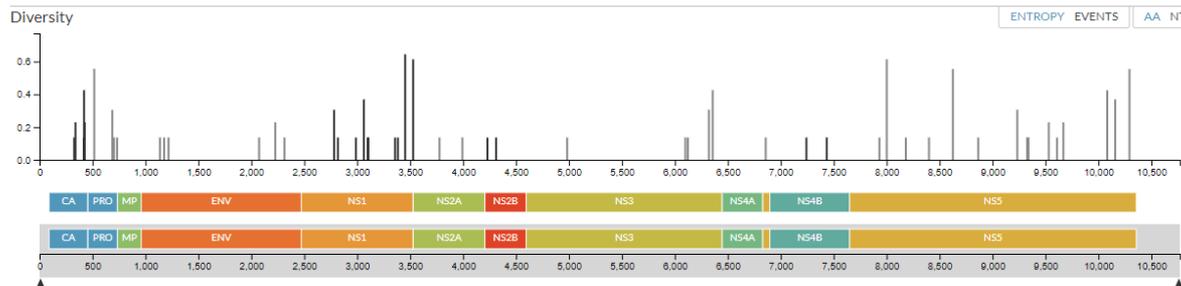
Phylogeny



Map



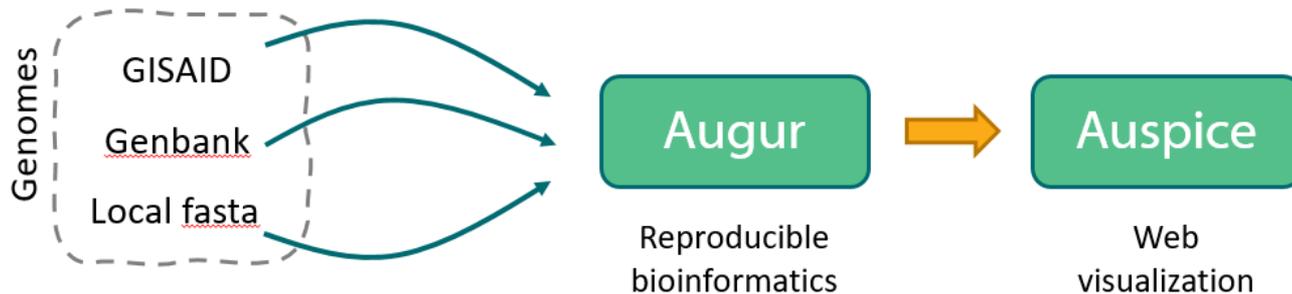
Genome



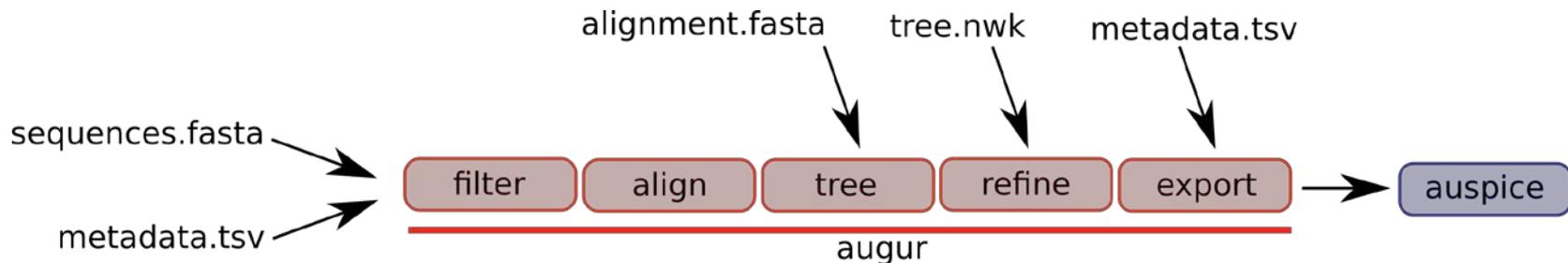
Nextstrain architecture

Two goals, two components

1. Rapid and flexible phylodynamic analysis (*Augur*)
2. Interactive visualization (*Auspice*)



Augur: What does it do?



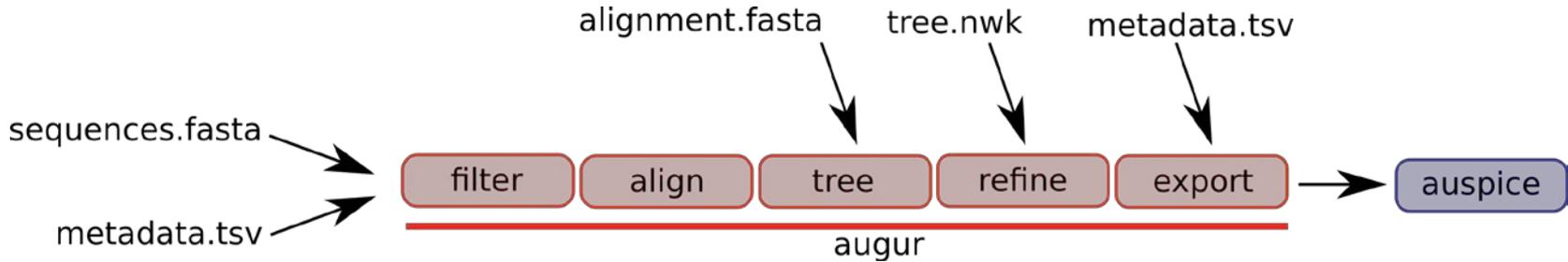
■ Input data

- `sequences.fasta`
- `metadata.tsv`
- *Can also import a tree if already constructed (like a Bayesian tree)*

■ Visualization data for Auspice

- Colors, `lat_longs`, reference genome

Augur: What does it do?

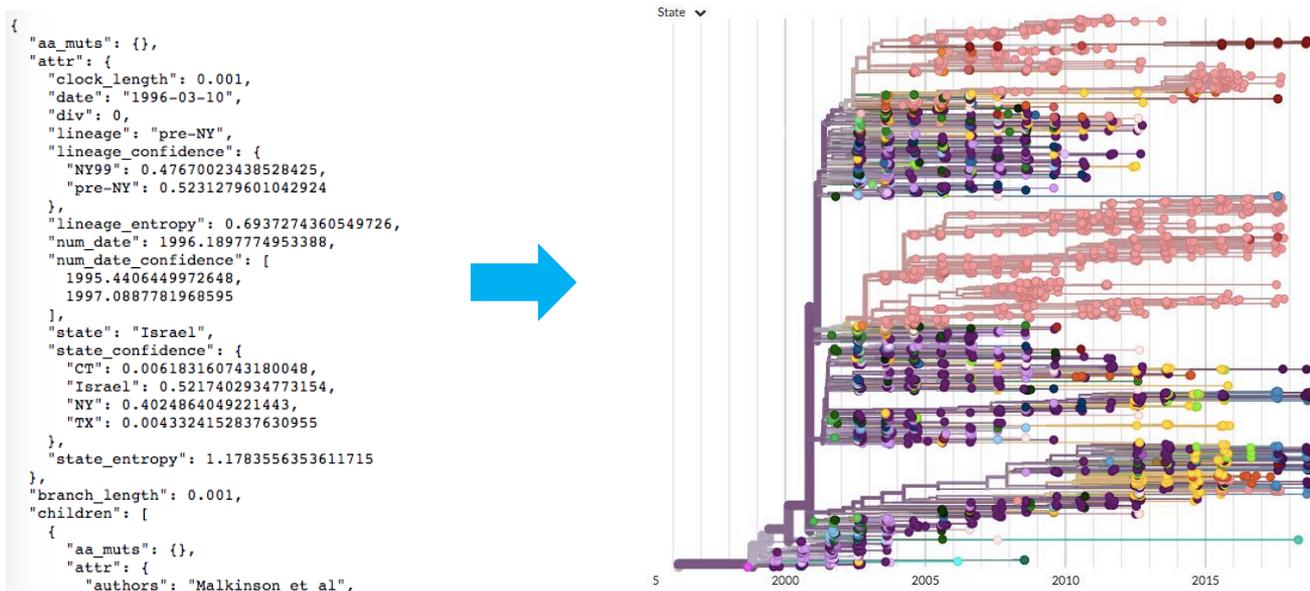


■ What augur does

- Prepare pathogen sequences and metadata
- Align sequences
- Construct a phylogeny from aligned sequences
- Annotate the phylogeny with inferred ancestral pathogen dates, sequences, and traits
- Export the annotated phylogeny and corresponding metadata into auspice-readable text file (JSON)

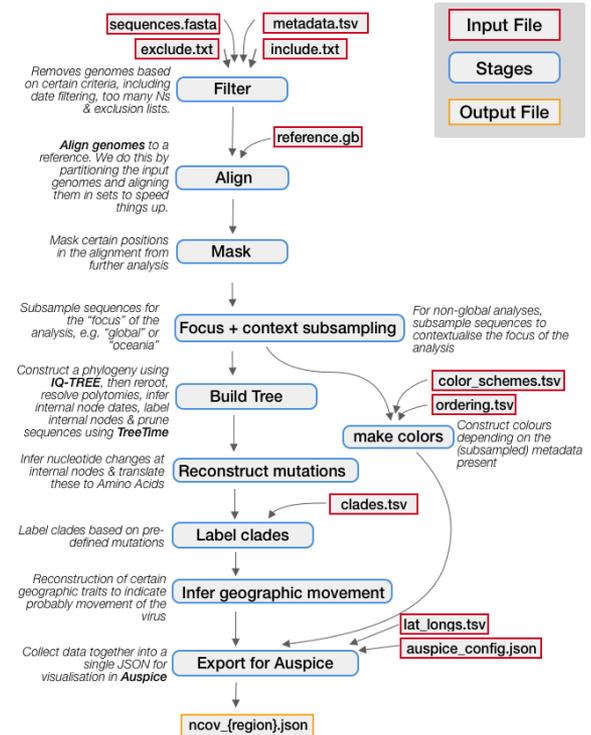
Auspice: What does it do?

- Interactive web-app for tree visualization
 - Translates data text files from augur into trees



What is a Nextstrain 'build' ?

- **Set of commands, parameters, and input files to reproducibly execute bioinformatic analyses and generate an output file for visualization**
- Allows user to frequently run several different analysis workflows or datasets, for example:
 1. Just your lab's data, from your jurisdiction
 2. Your data AND data from public repositories
 3. Data from your jurisdiction AND neighboring counties/states/etc.
- Nextstrain's focus on providing a *real-time* snapshot of evolving pathogen populations necessitates a reproducible analysis that can be rerun when new sequences are available



Nextstrain documentation

- **A Getting Started Guide to the Genomic Epidemiology of SARS-CoV-2**
 - Template and tutorial walks through the process of running a basic phylogenetic analysis on SARS-CoV-2 data, specifically to enable Departments of Public Health to start using Nextstrain to understand their SARS-CoV-2 genomic data
 - <https://docs.nextstrain.org/en/latest/tutorials/SARS-CoV-2/steps/index.html#a-getting-started-guide-to-the-genomic-epidemiology-of-sars-cov-2>

Analysis:

1. Setup and installation
2. Preparing your data
3. Orientation: analysis workflow
4. Orientation: which files should I touch?
5. Running & troubleshooting
6. Customizing your analysis
7. Customizing your visualization

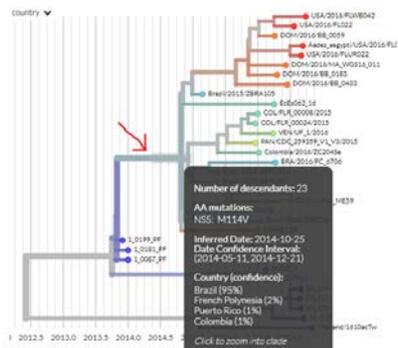
Visualization and interpretation:

1. Options for visualizing and sharing results
2. Interpreting your results
3. Writing a narrative to highlight key findings

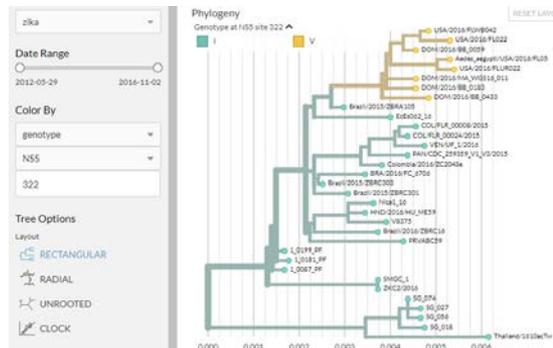
Nextstrain documentation

- Interacting with auspice, the visualization web application
 - Guides through the default phylogeny, map, and genome panels
 - https://neherlab.org/201901_krisp_auspice.html

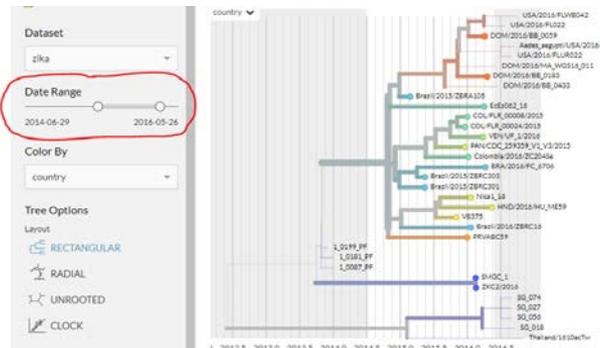
Node details



Highlight variants



Select date ranges



SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology and Surveillance



nextstrain.org/groups/spheres

A screenshot of a web browser displaying the Nextstrain website for the SPHERES group. The browser's address bar shows the URL "nextstrain.org/groups/spheres". The page features the Nextstrain logo in the top left and navigation links for "DOCS", "HELP", and "LOGIN" in the top right. The main content area includes a red and white virus particle icon, the title "SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology and Surveillance (SPHERES)", and a description of the consortium. Below this is a list of goals and a list of member organizations. The "Available datasets:" section lists various state and regional datasets.

Nextstrain / groups / spheres

nextstrain.org/groups/spheres

Nextstrain

DOCS HELP LOGIN

SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology and Surveillance (SPHERES)

A national genomics consortium to coordinate SARS-CoV-2 sequencing across the United States.
<https://www.cdc.gov/coronavirus/2019-ncov/covid-data/spheres.html>

The SPHERES consortium is being led by CDC's Advanced Molecular Detection (AMD) program, which over the past six years has invested in federal, state, and local public health laboratories to expand the use of pathogen genomics and other advanced laboratory technologies to strengthen infectious disease surveillance and outbreak response. SPHERES aims to

- Accelerate the use of real-time pathogen sequence data and molecular epidemiology for the COVID-19 pandemic response.
- Organize and manage public health sequencing and response efforts across the United States.
- Coordinate and support sequencing at state and local public health laboratories across the country.
- Better engage US clinical, academic, and commercial laboratories that are sequencing—or planning to sequence—SARS-CoV-2 data on any scale.
- Improve communication and knowledge-sharing between US laboratories.
- Develop consensus guidance on critical data and metadata standards.
- Reduce barriers to bioinformatic analysis and data sharing.
- Better align sequencing requirements and resource needs with different sources of funding, technology, expertise, and other means of support.

The SPHERES consortium includes 37 state and local public health laboratories, several large regional and national clinical diagnostic corporations, and academic and non-profit leaders in pathogen genomics, bioinformatics, and public health from across the country. Moreover, the consortium aligns federal laboratories and public health agencies with international genomics efforts and engages the private sector in efforts to better understand the genomics and patterns of SARS-CoV-2 transmission across the United States.

Available datasets:

- `groups/spheres/ncov/alabama`
- `groups/spheres/ncov/alaska`
- `groups/spheres/ncov/arizona`
- `groups/spheres/ncov/arkansas`
- `groups/spheres/ncov/california`
- `groups/spheres/ncov/colorado`
- `groups/spheres/ncov/louisiana`
- `groups/spheres/ncov/maine`
- `groups/spheres/ncov/maryland`
- `groups/spheres/ncov/massachusetts`
- `groups/spheres/ncov/michigan`
- `groups/spheres/ncov/minnesota`
- `groups/spheres/ncov/oklahoma`
- `groups/spheres/ncov/oregon`
- `groups/spheres/ncov/pennsylvania`
- `groups/spheres/ncov/puerto-rico`
- `groups/spheres/ncov/rhode-island`
- `groups/spheres/ncov/south-carolina`

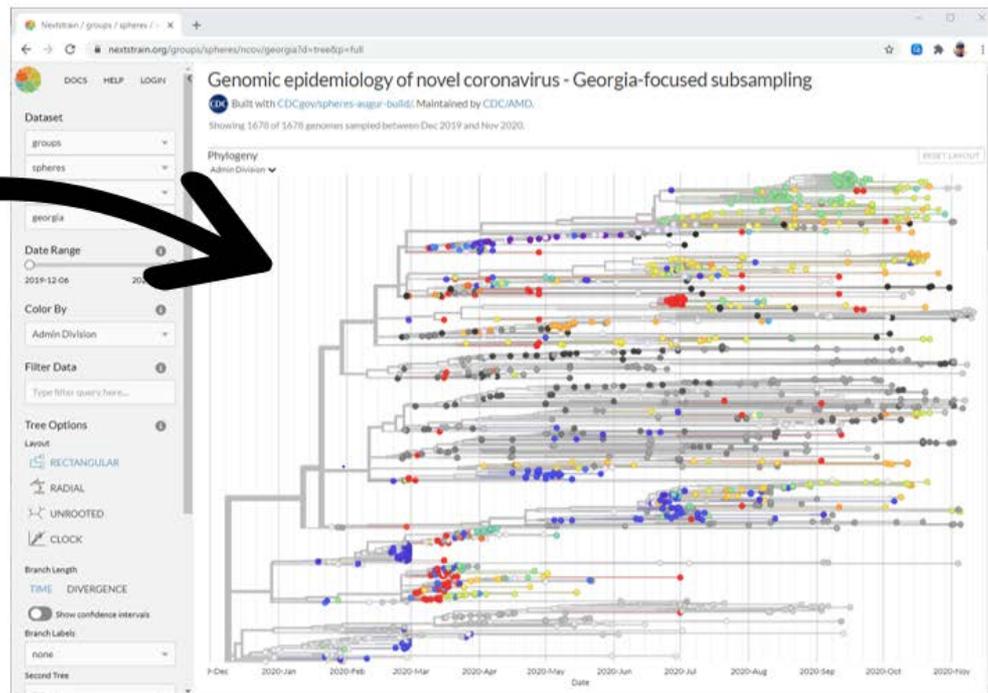
Drag-n-drop metadata

nextstrain.org/groups/spheres/nCoV/georgia



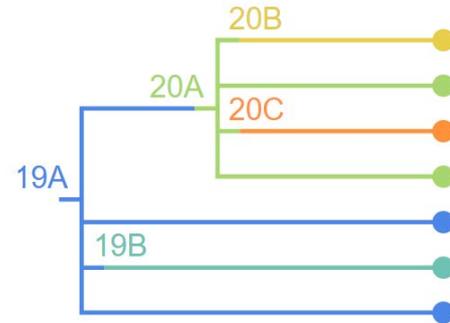
metadata.tsv

Strains	Cluster
Isolate-1	Hospital
Isolate-2	Hospital
Isolate-3	University
...	...



What about Nextclade?

- Web-browser tool to quickly analyze your SARS-CoV-2 genome:
 - Identify mutations compared to a reference used by Nextstrain
 - Assign your sequences to major clades
 - Quality check of your sequence data

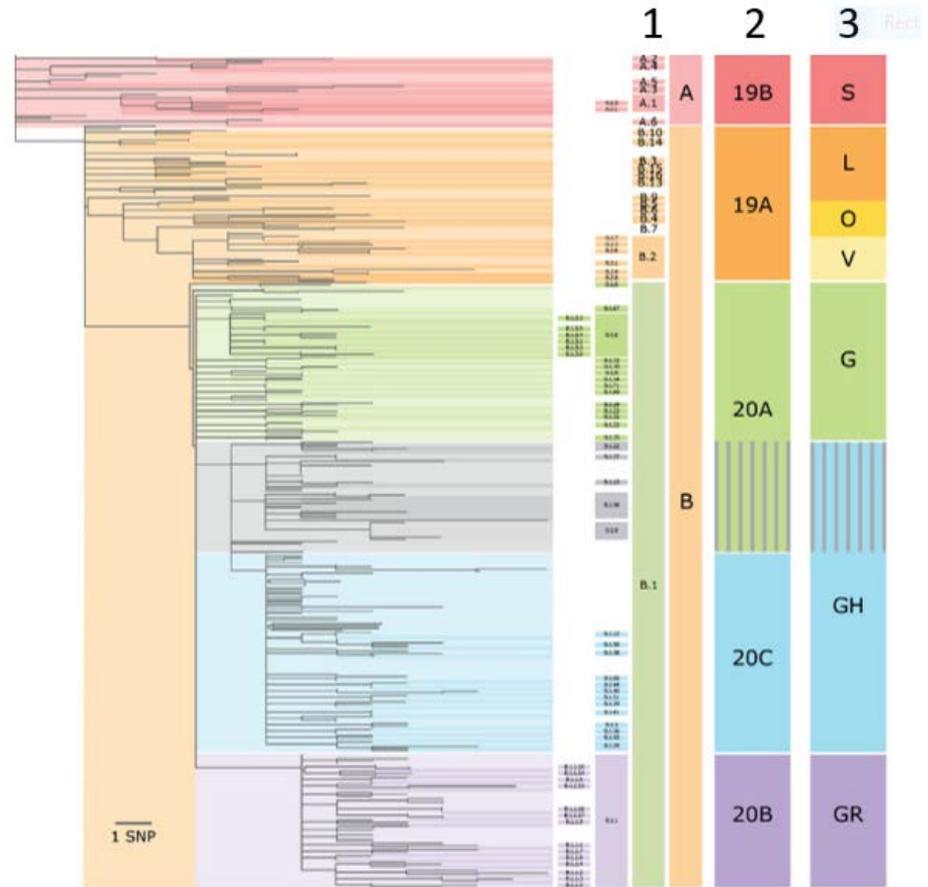


- Drag-and-drop a sequence file or paste sequences into the text box
- All analyses happen in your browser, never leaving your computer

SARS-CoV-2 clades:

Term for clades:

1. Pangolin Lineages
 - cov-lineages.org
2. Clades by Nextstrain ****
 - nextstrain.org
3. Clades by GISAID
 - gisaid.org



Nextclade clades.nextstrain.org

Nextclade ^{beta}

What's new English

Nextclade ^{beta} v0.9.0

Clade assignment, mutation calling, and sequence quality checks

Simple
No installation or setup - drop a file and see the results

Private
No remote processing - sequence data never leaves your computer

Mutation Calling
Find differences of your sequences relative to the reference in standard numbering

Clade Assignment
Find out which Nextstrain clades your samples are from

Phylogenetic Placement
See where on the SARS-CoV-2 tree your sequences fall

Quality Control
Check your data against multiple QC metrics

SARS-CoV-2 Simple mode Advanced mode

* Sequences **required**

Drag & Drop a file here
or

 FASTA

[Show me an Example](#)

Summary

- Nextstrain is a powerful tool to analyze pathogen genomic data and aid epidemiological understanding
- Design focus on *real-time* snapshots of evolving pathogen populations through reproducible analysis (augur)
- Features interactive web application for visualization (auspice)
- Widely used to monitor SARS-CoV-2 genome sequences
- TONS of documentation and tutorials at <https://nextstrain.org>

Acknowledgements

Nextstrain development is lead by

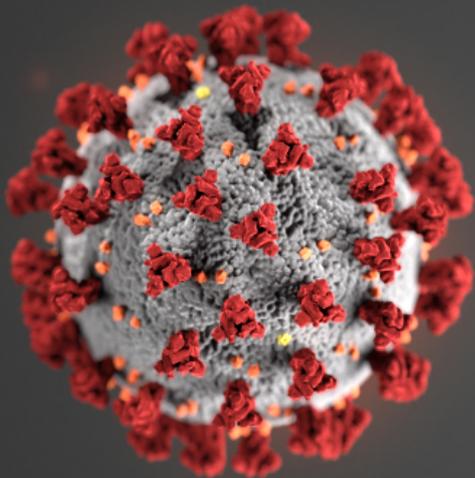
- Trevor Bedford, Fred Hutchinson Cancer Research Center
- Richard Neher, Universität Basel

Nextstrain application to SARS-CoV-2 is lead by

- James Hadfield
- Emma Hodcroft

Learn more

- Next modules
 - 3.2 Getting started with MicrobeTrace
 - 3.3 Linking epidemiologic data
- COVID-19 Genomic Epidemiology Toolkit
 - Find further reading
 - Subscribe to receive updates on new modules as they are released
go.usa.gov/xAbMw



For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

