



Clade and lineage assignment with Nextclade

Cornelius Roemer, University of Basel
Viral Sub-Species Classification Workshop
April 9, 2024

The challenge

Assigning genome sequences to clade

Without requiring bioinformatics skills

Easy to set up for any virus and nomenclature system

The image shows a Microsoft Word document with a table containing numerical data. The table has approximately 10 columns and 10 rows. The data is organized into several sections, with some rows highlighted in yellow and others in pink. A sidebar on the left side of the document displays various charts and graphs, including bar charts and line graphs, which are partially overlapping the table content. The document is titled 'Microsoft Word' and has a 'Print' button visible in the top right corner. The overall layout suggests a data analysis or reporting document.

Supports hierarchical nomenclature

Nextclade Start Dataset Results Tree Export Done. Total sequenc... Settings About Citation Docs CLI X D 🗑️ ↻ EN

#	i	Sequence name	QC	Clade	Outbreak	Lineage	Mut.	non-ACGTN	Ns	Cov.	Gaps	Ins.	FS	SC	
0	0	✓ MPXV_USA_2022_MA001 ON563414.	N M P F S	IIb	hMPXV-1	B.1	67	1	0	100.0%	107	22	1(1)	0	
1	1	✓ ON676708	N M P F S	IIb	hMPXV-1	A.1.1	58	0	4	100.0%	117	81	0	0	
2	2	✓ ON674051	N M P F S	IIb	hMPXV-1	A.2.1	38	0	5	100.0%	59	16	1(1)	0(1)	
3	4	✓ MPXV-UK_P2 MT903344.1	N M P F S	IIb	hMPXV-1	A.1	21	0	0	100.0%	76	100	0	0	
4	3	✓ MT903339	N M P F S	IIb	hMPXV-1	A	9	0	0	99.8%	20	767	0	0	
5	7	✓ ON843165	N M P F S	IIb	hMPXV-1	B.1.5	68	0	75	100.0%	88	100	0	0	
6	5	✓ Yambuku_DRC_1985	N M P F S	I			815	0	0	100.0%	3133	3172	4(7)	0(1)	Too many markers to display (991). T
7	8	✓ KJ642617	N M P F S	IIb			41	1	0	99.8%	55	797	2(4)	0	
8	6	✓ Ivory_Coast_2012	N M P F S	IIa			555	0	0	100.0%	261	3087	5(6)	0	Too many markers to display (763). T

Additional features

Private by design: Data stays on user's computer

Browser based: no installation necessary

Fast: 100 SARS-CoV-2 sequences per second

Works with partial sequences >100bp

Handles >15% nucleotide divergence

High performance CLI handles 16M SC2 in a few hours on a laptop

Nextclade's algorithm in a nutshell

1. Align nucleotide sequence to reference
2. Translate and align protein sequences according to annotation
3. Place on phylogenetic reference tree
4. Present results to user

Underlined are the config files that tailor Nextclade to a particular virus

A complete set of config files is called a *dataset*

Nextclade datasets

Hosted in `nextstrain/nextclade_data` Github repository

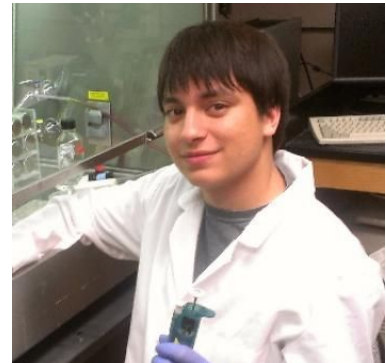
Anyone can create and share new datasets

Example: Michael Zeller (Iowa State) made a PRRSV dataset



PRRSV-2 ORF5 Lineages, Yim-im & Zhang 2023...
community

Reference: PRRSV0004437 (DQ478308.1)
Updated at: 2024-02-22 16:12:03 (UTC)
Dataset name: `community/isuvdl/mazeller/prrsv2/orf5/yimim2023`



How to make your own dataset?

Pick a reference sequence (fasta)

Curate a genome annotation (gff3)

Make a Nextstrain tree annotated with clades (auspice.json)

Put files in folder, make pull request in nextclade_data repository

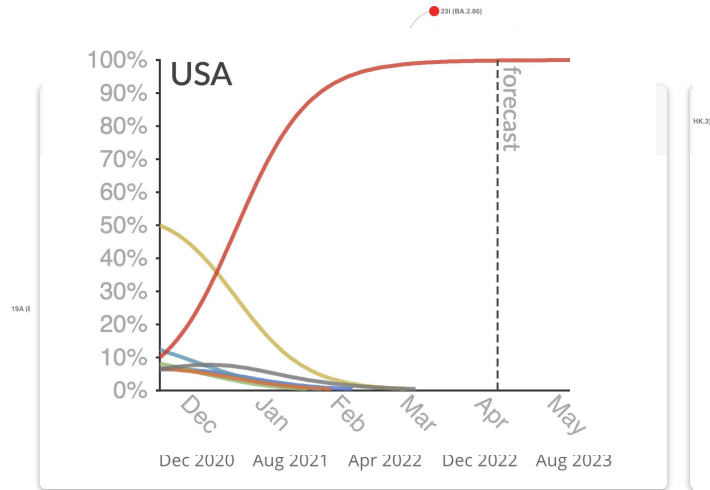
Tutorial available for guidance, reach out for help!

Coarse (Nextstrain clades)

Captures major diversity

Useful for high-level reporting

Year-letter: 23I instead of BA.2.86

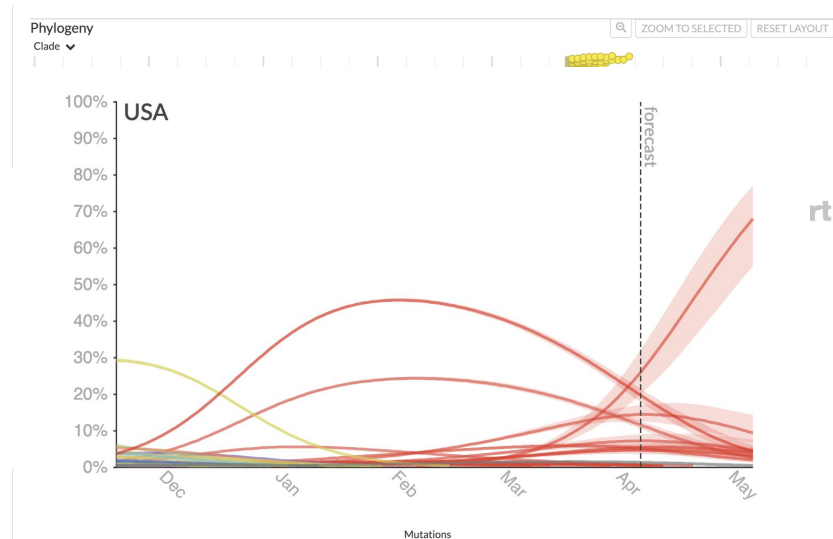


Fine (Pango lineages)

Shared names avoid ad-hoc naming: “that new cluster in X from Y with SNP Z”

Useful for specialists: tracking, papers, forecasting

Can be overwhelming for casual observers



Putting classification ground truth on Github

The screenshot displays the GitHub interface for the repository 'cov-lineages / pango-designation'. The top navigation bar includes links for Code, Issues (138), Pull requests (1), Actions, Projects, Wiki, Security, and Insights. A search bar is present in the top right.

Below the navigation bar, there are two issue cards:

- Example Lineage Proposal**: #1 opened on Aug 14, 2020 by rambaut. Status: Open. 1 comment.
- Variant spotters announcement: n**: #1988 opened on May 6, 2023. Status: Open. 25 comments.

The main content area shows a list of issues with filters set to 'is:issue is:open'. The list includes:

- JN.1+C4777T+S:F456L(69 seqs, 12 countries) with S:R346T(31 seq JN.1.16 and JN.1.16.1**: #2555 opened yesterday by aviczhi2.
- JN.1.39 + T111C [5' UTR] (469 seq, Apr 6)**: #2554 opened 3 days ago by rylisner.
- JN.1+S:Q677H(42 seqs, 7 countries)**: #2549 opened last week by aviczhi2.
- JN.1+Orf3a:A99V+S:S60P,R346T (9 seqs, 3 countries)+S:F456L(6 JN.1+Orf3a:S60F branch (2 seqs)**: #2547 opened last week by aviczhi2.
- JN.1.9 FLIRT lineage , (10 on Gisaid) with S:S31del (8 , all the most r**: #2546 opened last week by FedeGuelli.

The detailed view of issue #32, 'Proposal for a B.1.3 sublineage potentially associated with recent outbreaks in East Asia (45 sequences) #32', is shown. It is marked as 'Closed' and was opened by user c19850727 on Jul 8, 2023. A comment from the same user on Jul 8, 2023, includes a phylogenetic tree titled 'Genomic epidemiology of monkeypox virus'. The tree shows the genetic relationships between various monkeypox virus sequences from May 2022 to March 2023, with a specific lineage highlighted in orange. The x-axis represents time, and the y-axis represents genetic distance.

On the right side of the issue view, there are sections for 'Assignees' (No one—assign yourself), 'Labels' (None yet), 'Projects' (None yet), 'Milestone' (No milestone), 'Development' (Successfully merging a pull request may close this issue), 'Notifications' (Unsubscribe), and '3 participants'.

Maintenance workload and sustainability

Classification/designation as a side effect of studying the virus

Good tooling helps reduce workload

Acknowledgements



Richard Neher



Ivan Aksamentov



University
of Basel



Nextstrain



Swiss Institute of
Bioinformatics