Clade and lineage assignment with Nextclade

Cornelius Roemer, University of Basel

Viral Sub-Species Classification Workshop

April 9, 2024

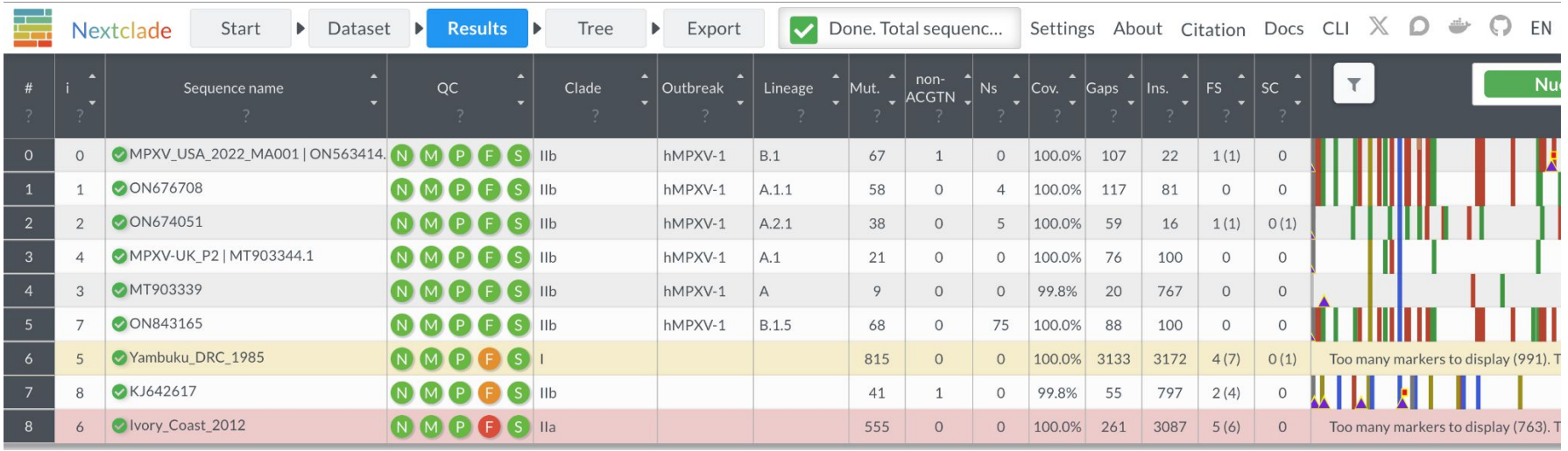
The challenge

Assigning genome sequences to clade

Without requiring bioinformatics skills

Easy to set up for any virus and nomenclature system

Supports hierarchical nomenclature



#	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

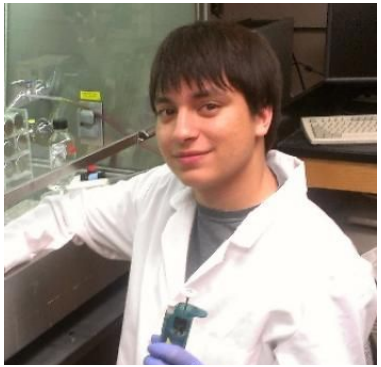
A complete set of config files is called a *dataset*

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



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

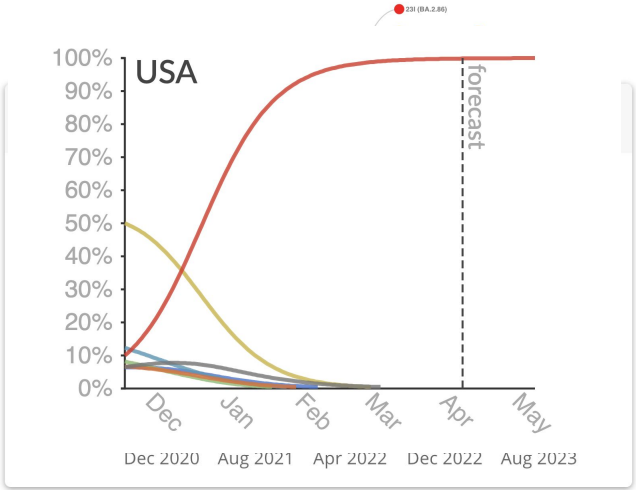
Tutorial available for guidance, reach out for help!

Coarse (Nextstrain clades)

Captures major diversity

Useful for high-level reporting

Year-letter: 23I instead of BA.2.86

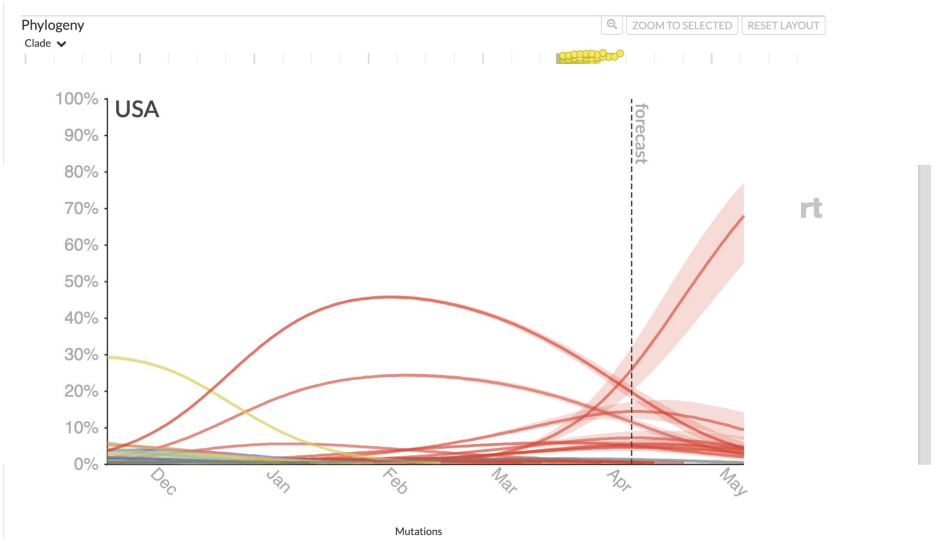


Fine (Pango lineages)

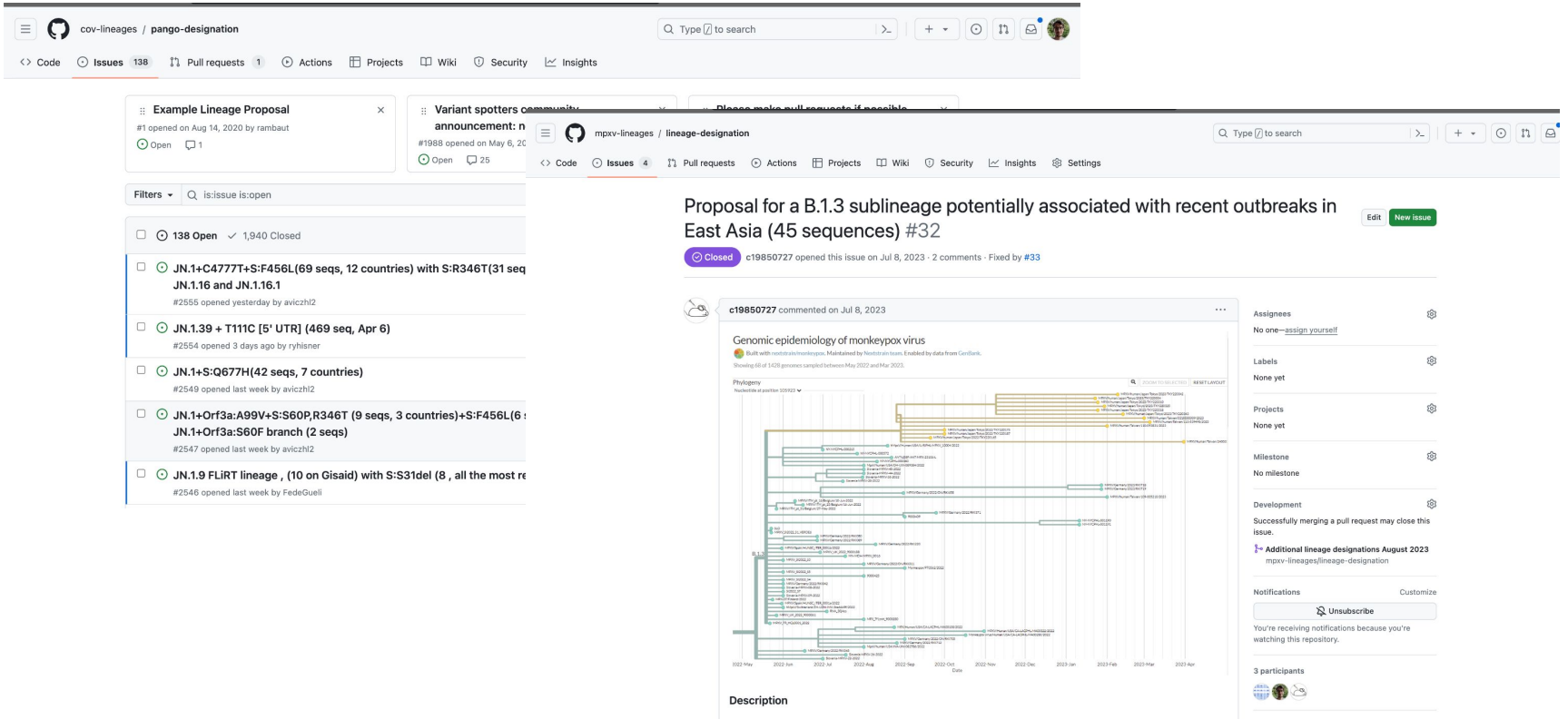
Shared names avoid ad-hoc naming: “that new cluster in X from Y with SNP Z”

Useful for specialists: tracking, papers, forecasting

Can be overwhelming for casual observers



Putting classification ground truth on Github



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

SIB Swiss Institute of Bioinformatics