pangolin --usher

Angie Hinrichs
StaPH-B - Nov. 19, 2021

Outline

- Background
  - UShER
  - UCSC's big tree built by UShER
  - Pango lineages
- Pangolin v3
- **`pangolin --usher`**
  - How does it work?
  - How is it different from pangoLEARN?
- Looking ahead

Ultrafast Sample placement on Existing tRees (UShER)

https://github.com/yatisht/usher/

Yatish Turakhia, UCSD

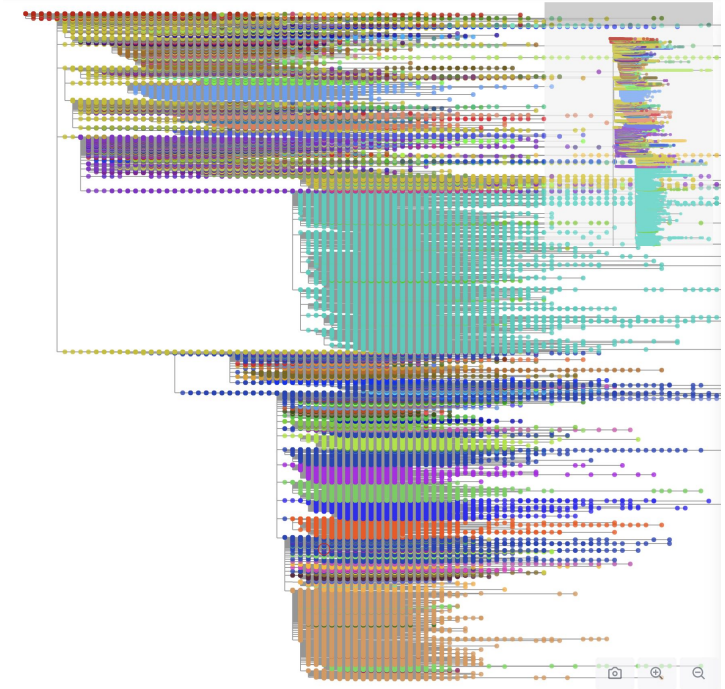
- Precomputed Mutation Annotated Tree (MAT) data structure
- Place new sequence in tree by Maximum Parsimony
- Fast! (just seconds to place on tree of >5M sequences)
- Web interface, matUtils, matOptimize, workflows, ...

https://www.nature.com/articles/s41588-021-00862-7

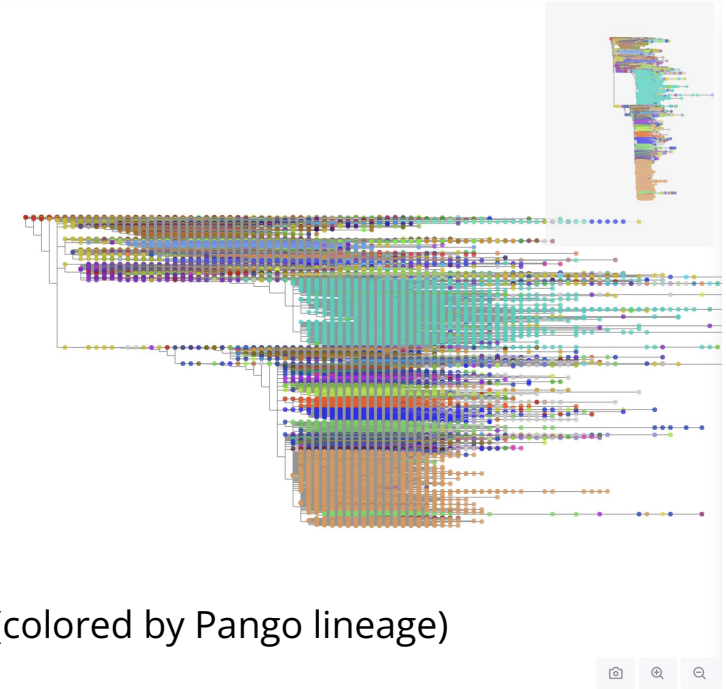
UCSC's Big Trees

UCSC's Big Trees

>5M: GISAID, GenBank, COG-UK
Not publicly shareable 😒



>2.5M: GenBank, COG-UK
Public downloads 😁



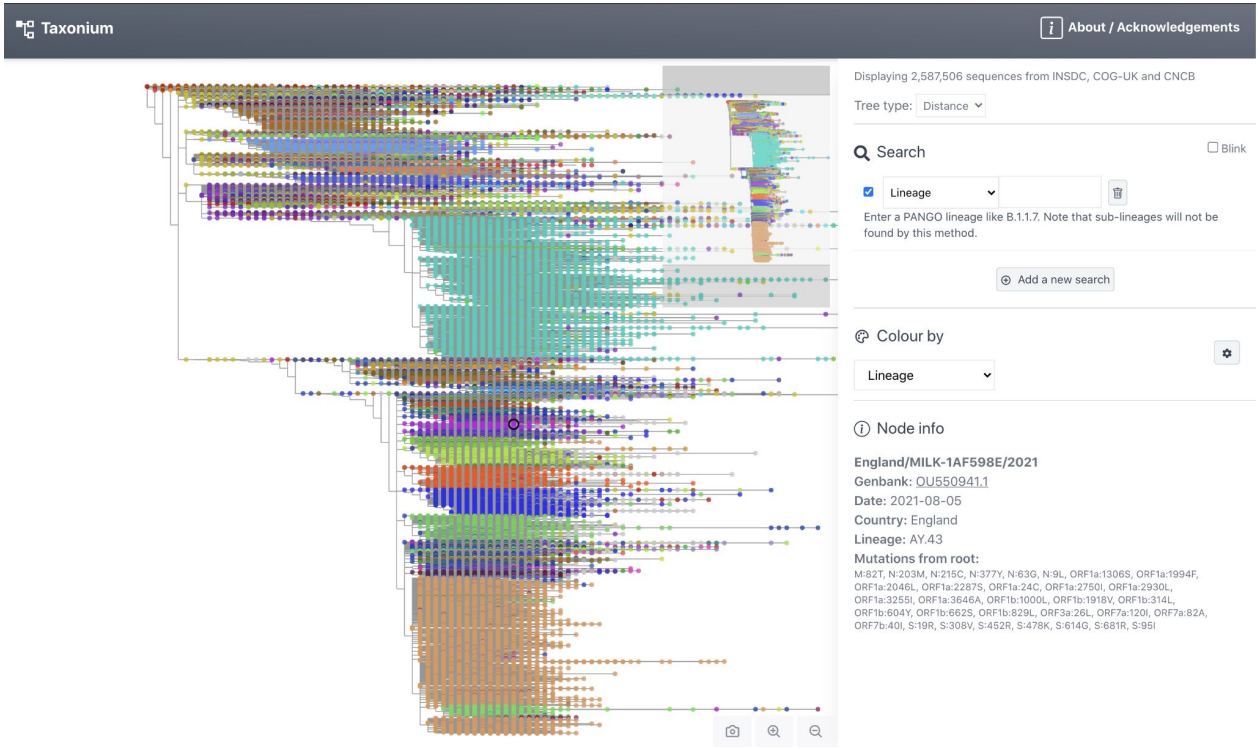
(colored by Pango lineage)

UCSC's Big Trees

- Daily update
  - Aggregate & deduplicate sequences and metadata
  - QC: remove sequences with <20000 non-N bases
  - Align new sequences to reference
  - Mask Problematic Sites
  - Use UShER to add new sequences to yesterday's tree
  - QC: Remove sequences with too many equally parsimonious placements
  - Extract public tree

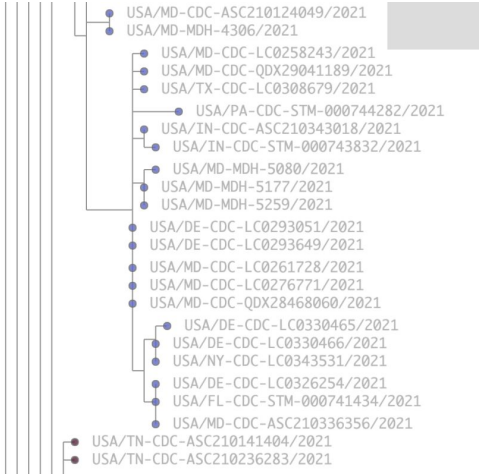
Browse the public tree with Taxonium

cov2tree.org



Theo Sanderson

Francis Crick Institute /
Wellcome Sanger Institute



Pango lineages

- A, B, B.1, B.1.1.7, B.1.351, P.1, B.1.617.2, AY.∞, ...
- Initially proposed in bioRxiv April 2020
- cov-lineages.org
- Pango.network
  - Criteria for new lineages
  - How to suggest a new lineage
  - News (AY.* updates!)
- pango-designation github repository

> Nat Microbiol. 2020 Nov;5(11):1403-1407. doi: 10.1038/s41564-020-0770-5. Epub 2020 Jul 15.

A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology

Andrew Rambaut [1], Edward C Holmes [2], Áine O'Toole [3], Verity Hill [3], John T McCrone [3], Christopher Ruis [4], Louis du Plessis [5], Oliver G Pybus [6]

Affiliations + expand

PMID: 32669681 PMCID: PMC7610519 DOI: 10.1038/s41564-020-0770-5 Sign in

What defines a Pango lineage?

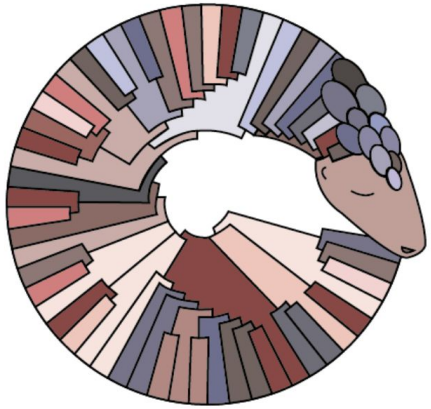
Not a set of mutations!

lineages.csv in the pango-designation github repository (>1M lines):

```
...
India/GJ-ICMR-NIV-INSACOG-GSEQ-3045/2021,B.1.617.2
India/PY-SEQ_294_S22_R1_001/2021,B.1.617.2
Malaysia/IMR_682164/2021,B.1.617.2
Japan/IC-1175/2021,B.1.617.2
USA/TX-CDC-ASC210037740/2021,B.1.617.2
England/WSFT-25C6539/2021,B.1.1.7
USA/MI-UM-10039543606/2021,AY.3
USA/KS-KHEL-1922/2021,AY.3
USA/KS-KHEL-1923/2021,AY.3
USA/MO-MSPHL-002099/2021,AY.3
USA/MO-MSPHL-002132/2021,AY.3
...
```

A Brief History of Pangolin

Phylogenetic Assignment of Named Global Outbreak LINeages



Contributors 23

+ 12 contributors

- v1.0 (April 29, 2020): phylogenetic model (iqtree... not fast enough)
- v2.0 (July 22, 2020): pangoLEARN model (fast! sensitive to noise)
- v3.0 (May 27, 2021): pangoLEARN + scorpio/constellations + `--usher` option

How does pangoLEARN work?

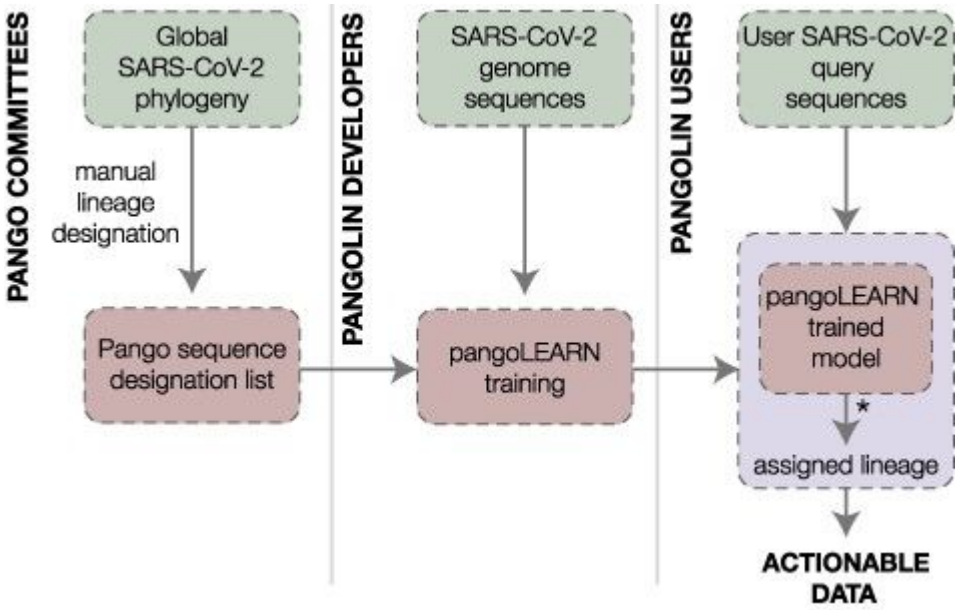
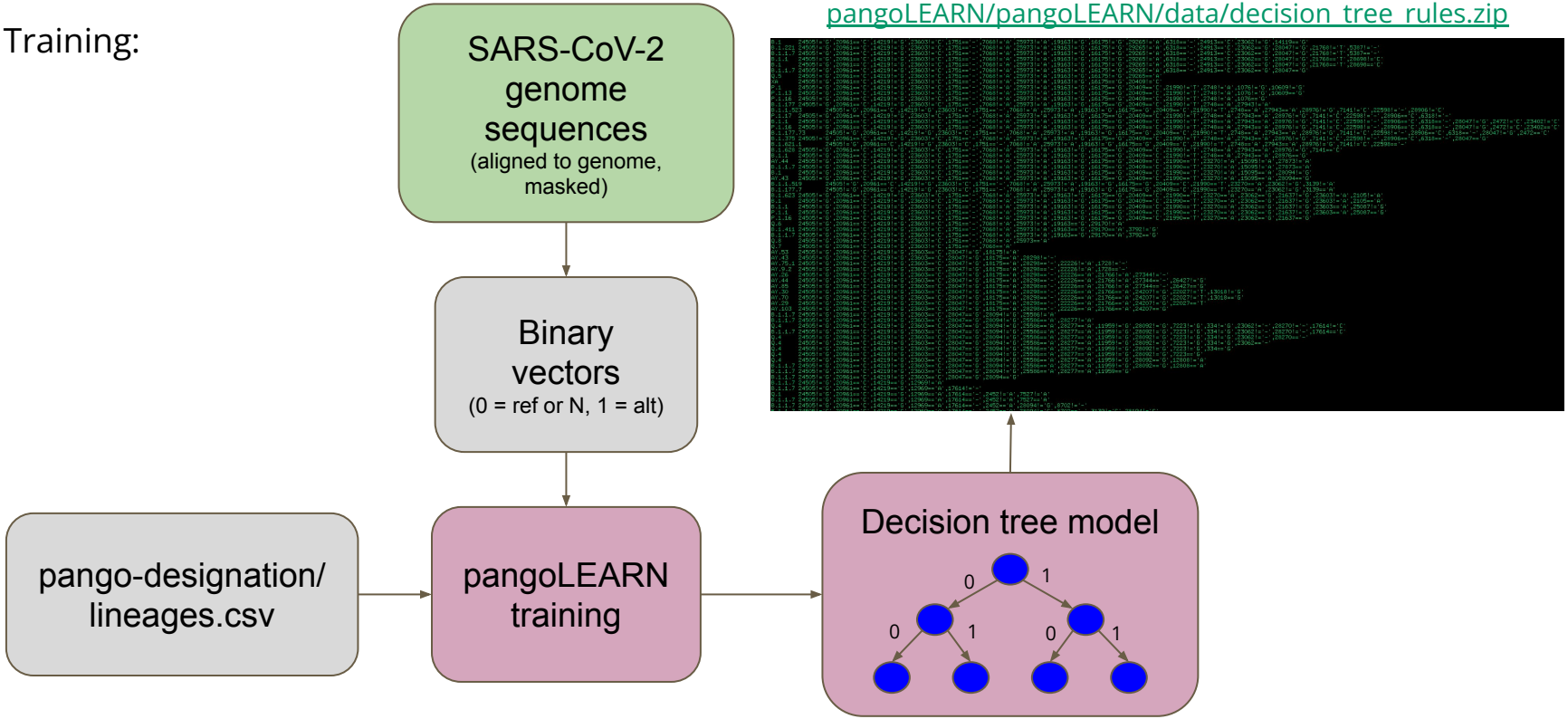


Figure 2, Áine O'Toole, Emily Scher, *et al.*, Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool, Virus Evolution, Volume 7, Issue 2, November 2021, veab064, https://doi.org/10.1093/ve/veab064

How does pangoLEARN work?

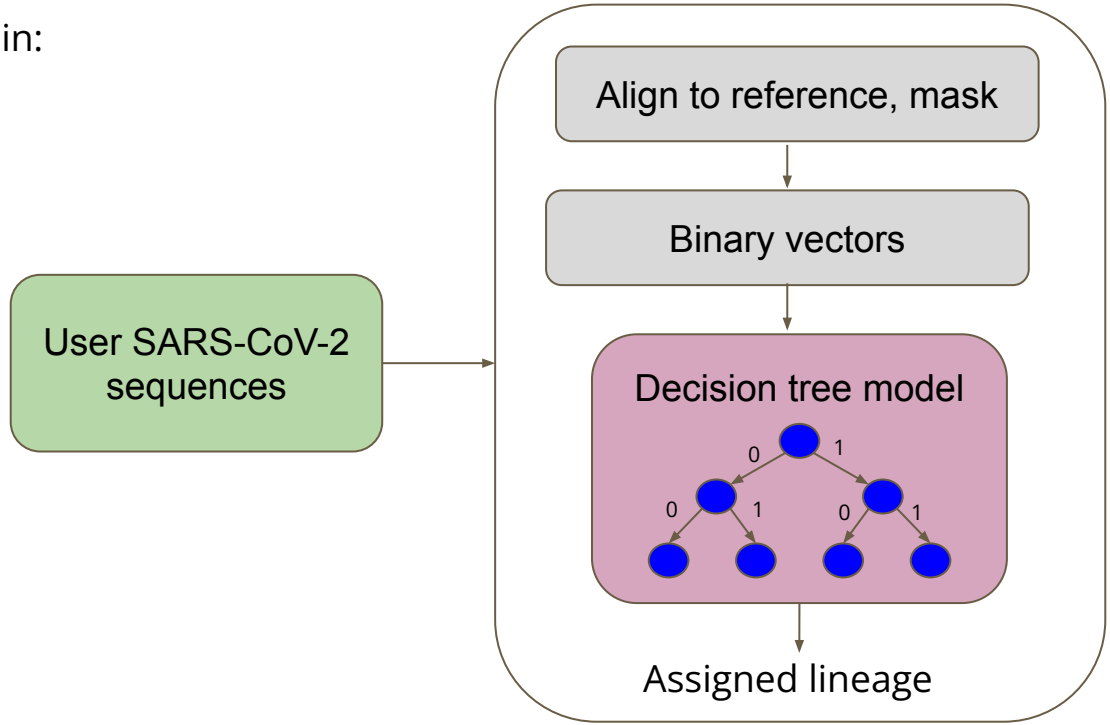
Training:



pangoLEARN/pangoLEARN/data/decision_tree_rules.zip

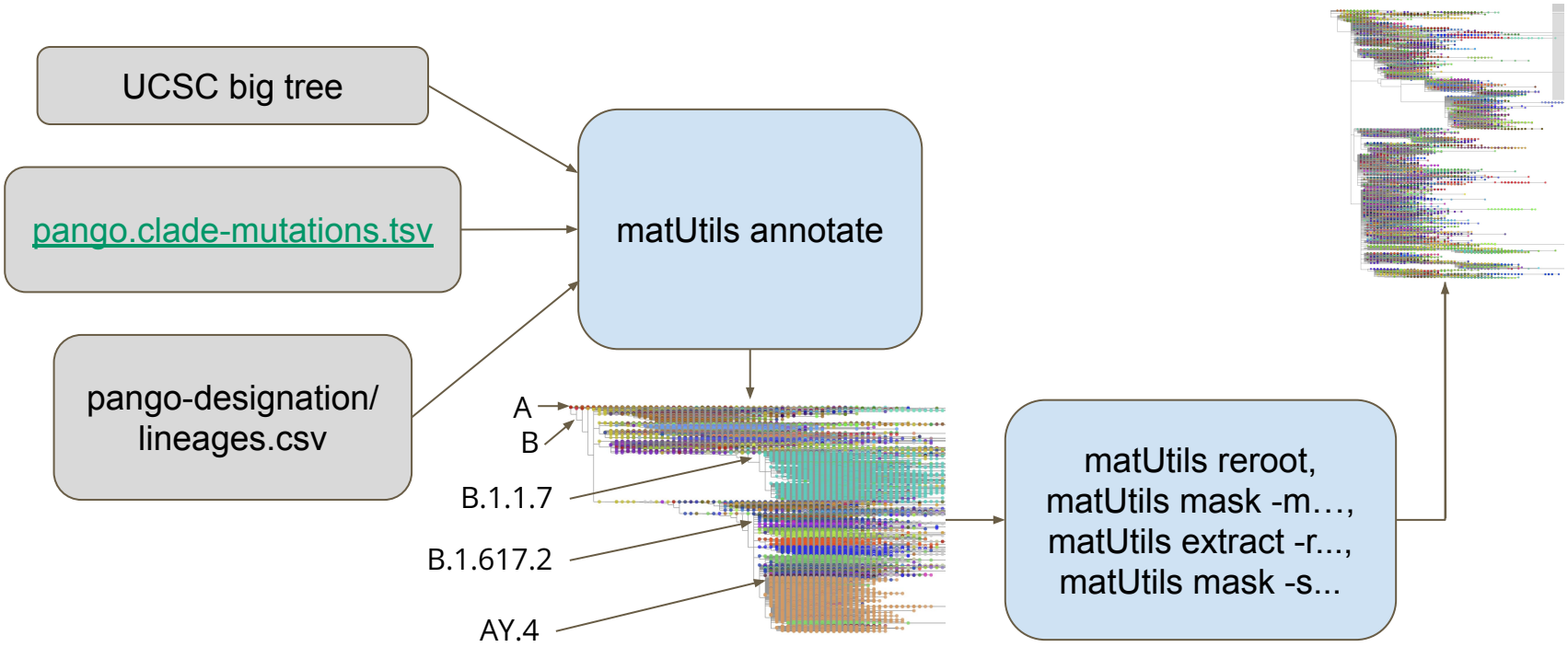
How does pangoLEARN work?

Running pangolin:



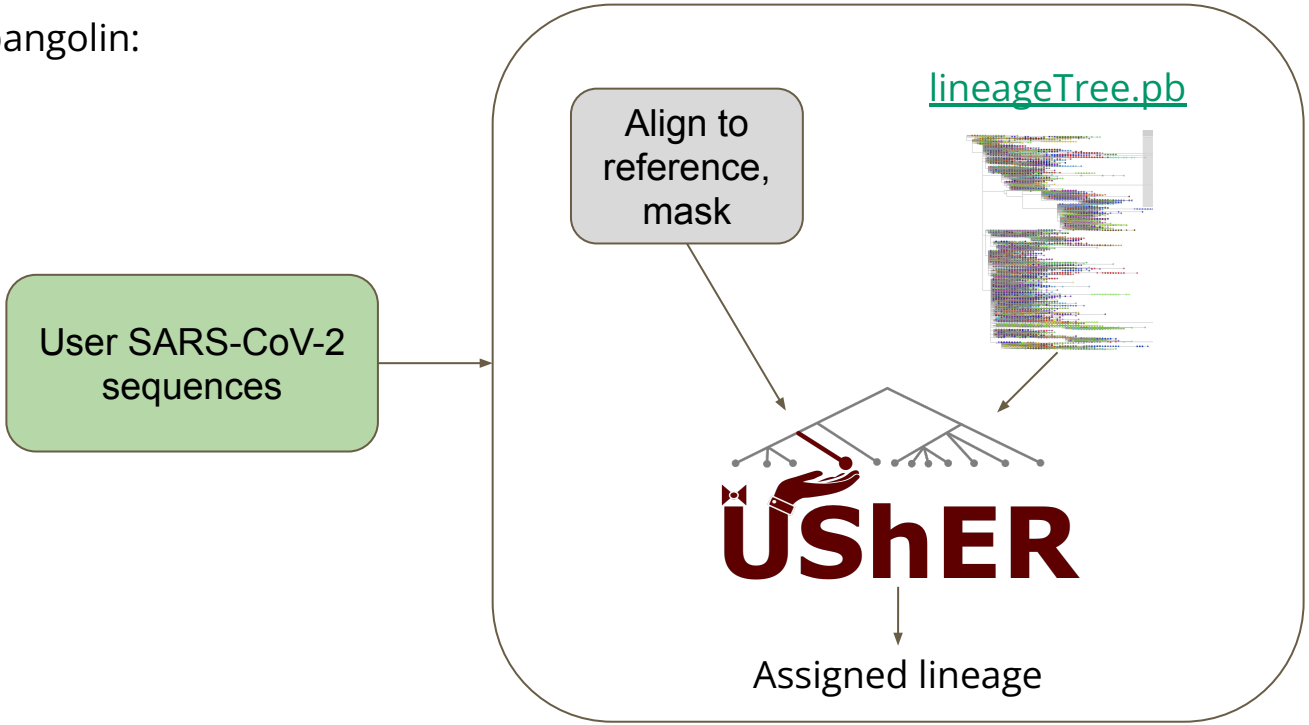
How does `pangolin --usher` work?

Making lineage-annotated tree:



How does `pangolin --usher` work?

Running pangolin:



What's the difference?

- pangoLEARN is ~16x faster
- UShER uses a mutation-annotated phylogenetic tree

Not all assignments come from pangoLEARN/UShER

- Designated sequences: directly assigned, no pangoLEARN/UShER

1002005561,AY.44,,,,,,**PANGO-v1.2.93**,3.1.16,2021-11-09,v1.2.93,passed_qc,**Assigned from designation hash.**

- Scorpio/constellations: overrides pangoLEARN/UShER

2000051407,**B.1.617.2**,0.0,0.9288622754491018,Delta (B.1.617.2-like),0.384600,0.076900,**PLEARN-v1.2.93**,3.1.16,2021-11-09,v1.2.93,passed_qc,scorpio call: Alt alleles 5; Ref alleles 1; Amb alleles 6; Oth alleles 1; **scorpio replaced lineage assignment B.1.1.7**

3000136426,**None**,,,,,,**PLEARN-v1.2.93**,3.1.16,2021-11-09,v1.2.93,passed_qc,**pangoLEARN lineage assignment AY.4.5 was not supported by scorpio**

3000137678,**B.1.617.2**,0.5,,Delta(B.1.617.2-like),1.000000,0.000000,**PUSHER-v1.2.93**,3.1.16,,v1.2.93,passed_qc,scorpio call: Alt alleles 13; Ref alleles 0; Amb alleles 0; **scorpio replaced lineage assignment AY.4**; Usher placements: AY.4(1/2) B.1.617.2(1/2)

7000000606,**None**,,,,,,**PUSHER-v1.2.93**,3.1.16,,v1.2.93,passed_qc,**usher lineage assignment AY.13 was not supported by scorpio; Usher placements: AY.13(5/6) B.1.617.2(1/6)**

Looking forward...

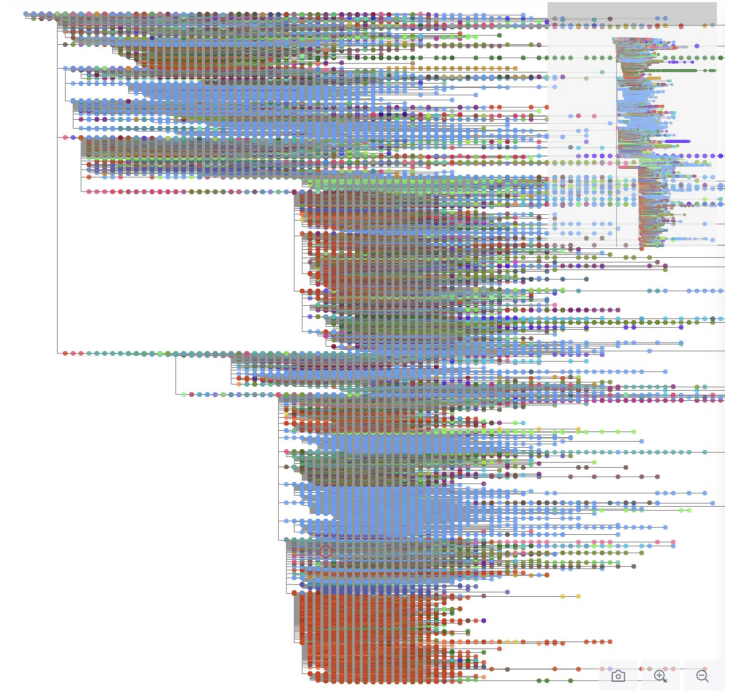
- Definitely: Ongoing updates with new lineages
- Probably: Precomputed assignments
- Maybe?: Expanded use of Scorpio

Acknowledgements

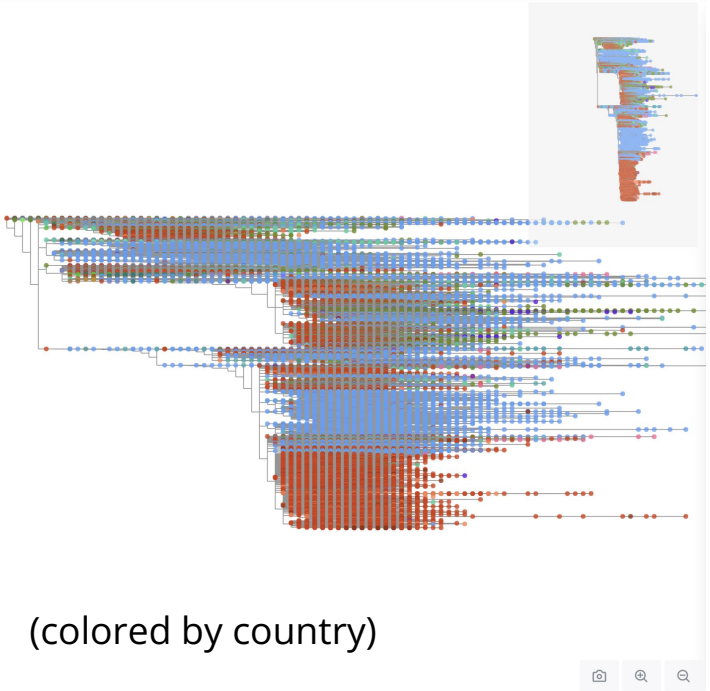
- UCSD: Yatish Turakhia, Cheng Ye (UShER, matOptimize)
- U. of Edinburgh: Àine O'Toole, Emily Scher, Rachel Colqhoun, Andrew Rambaut (pangolin)
- UCSC: Russ Corbett-Detig, Jakob McBroome, Bryan Thorlow, Alex Kramer, Marc Perry (matUtils, evaluation)

UCSC's Big Trees

>5M: GISAID, GenBank, COG-UK
Not publicly shareable



>2.5M: GenBank, COG-UK
Public downloads



(colored by country)

Comparison of pangoLEARN vs UShER in discriminating Delta and sublineages

Michelle Su

11.19.2021

NYC Health

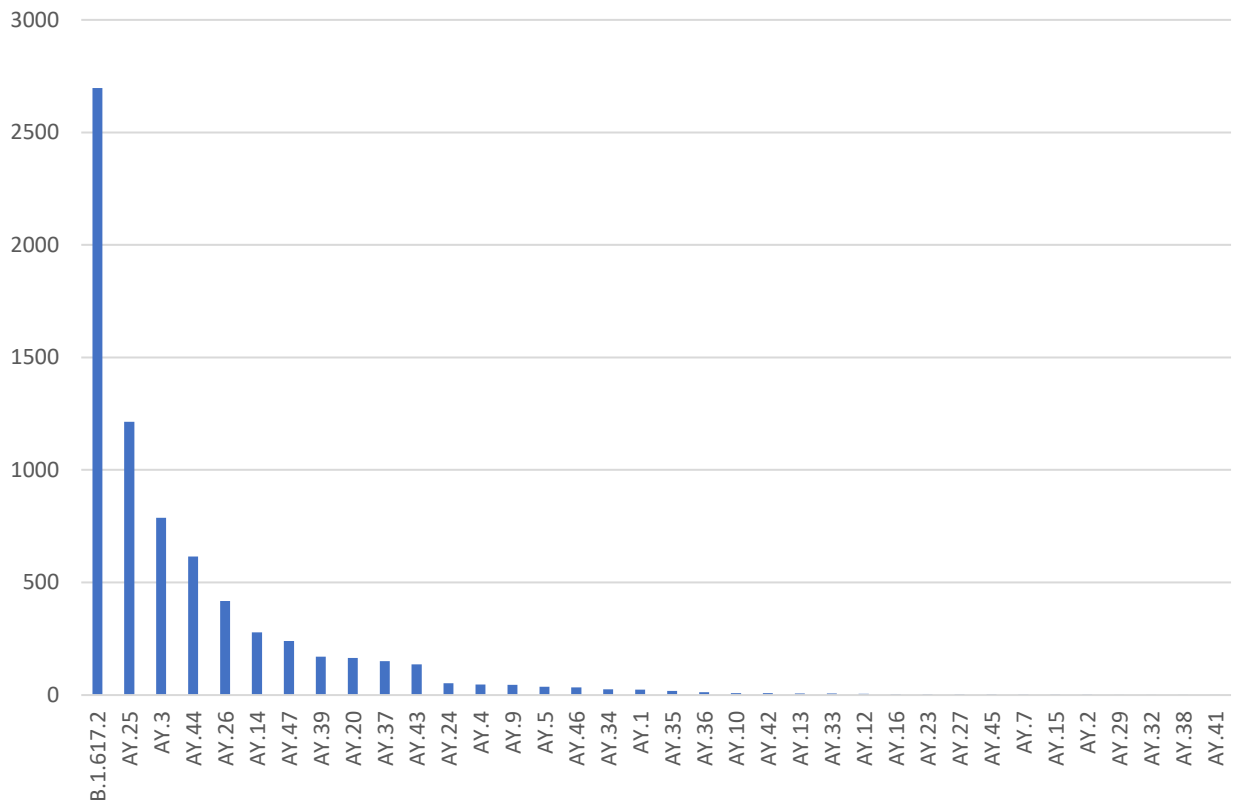
Background

Some pangoLEARN lineage calls were not monophyletic on a phylogenetic tree.

What is the extent of this issue in the Delta lineage?

Sample set

- Delta(B.1.617.2) and sublineages
  - 7229 samples
  - 36 lineages – pangoLEARN
    - *Sub sub lineages removed
  - Majority are parental lineage
    - 11 lineages have > 100 sequences



NYC Health

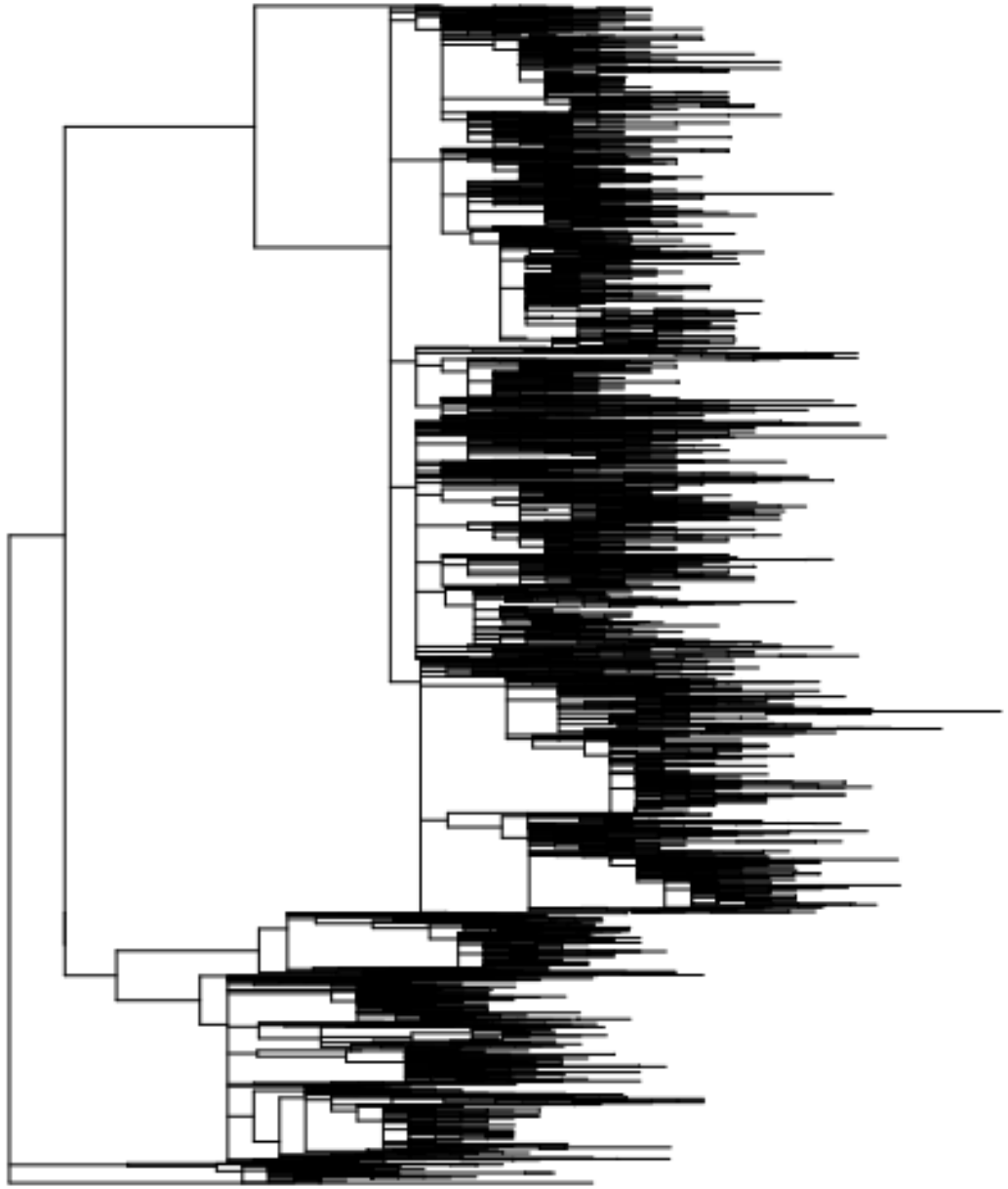
Methods

1. Phylogenetic tree building
    1. Take consensus genomes
    2. Align genomes using MAFFT (Multiple Alignment using Fast Fourier Transform)
    3. Build phylogenetic tree with IQ-TREE

2. Pangolin lineage calling
    1. Run pangolin using pangoLEARN and UShER
    2. Seven timepoints from 9.22 – 11.04
        1. Pango designation v.1.2.76-91, pangolin v.3.1.11-16, pangoLEARN 2021-09-17 to 2021-10-18

NYC Health

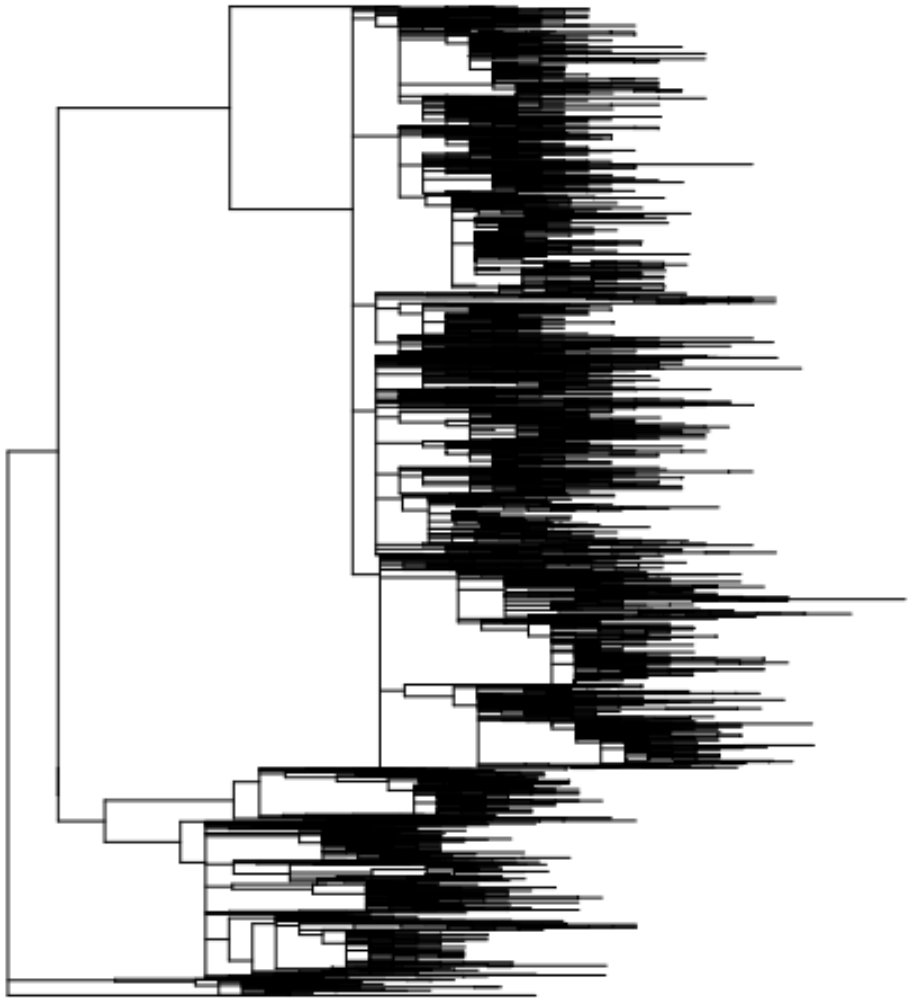
Phylogenetic tree of 7229 Delta samples

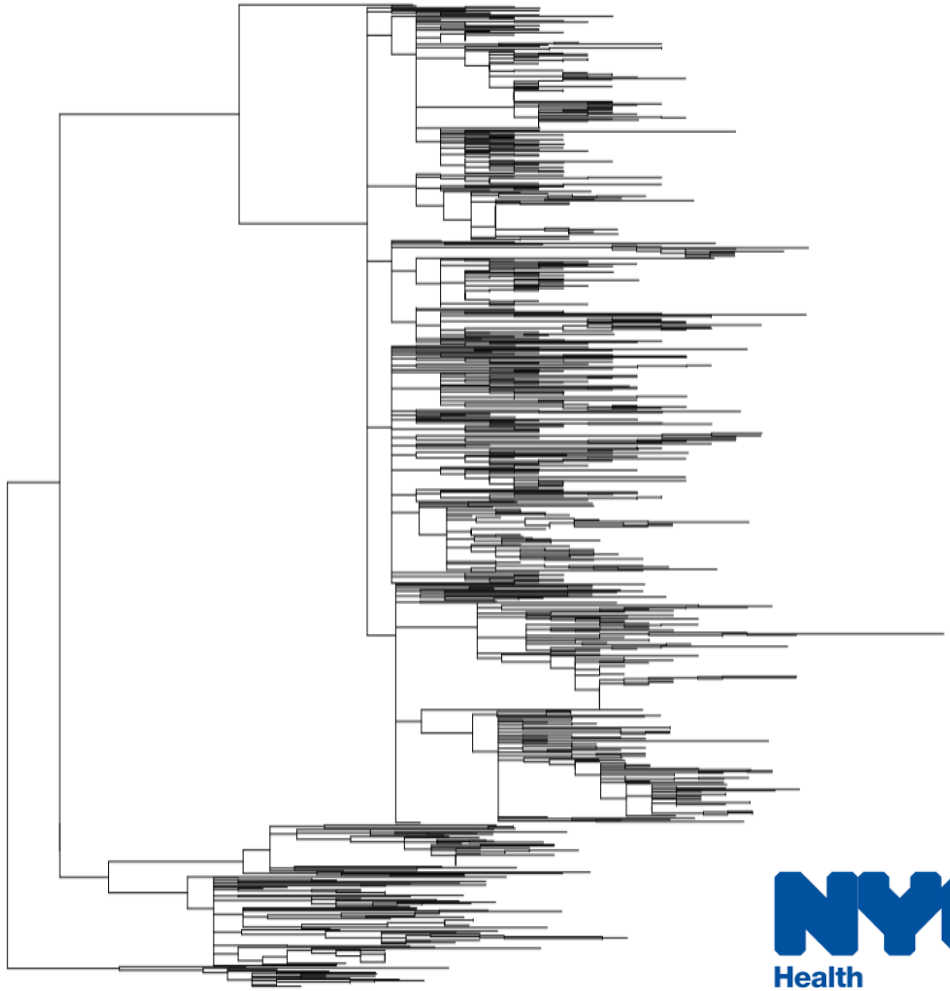


Downsample tree

- For all nodes in the tree, check if tips are the same lineage and have same history of pangoLEARN calls.
  - Of the unique lineages, pick one sample and discard the rest
- Repeat if necessary

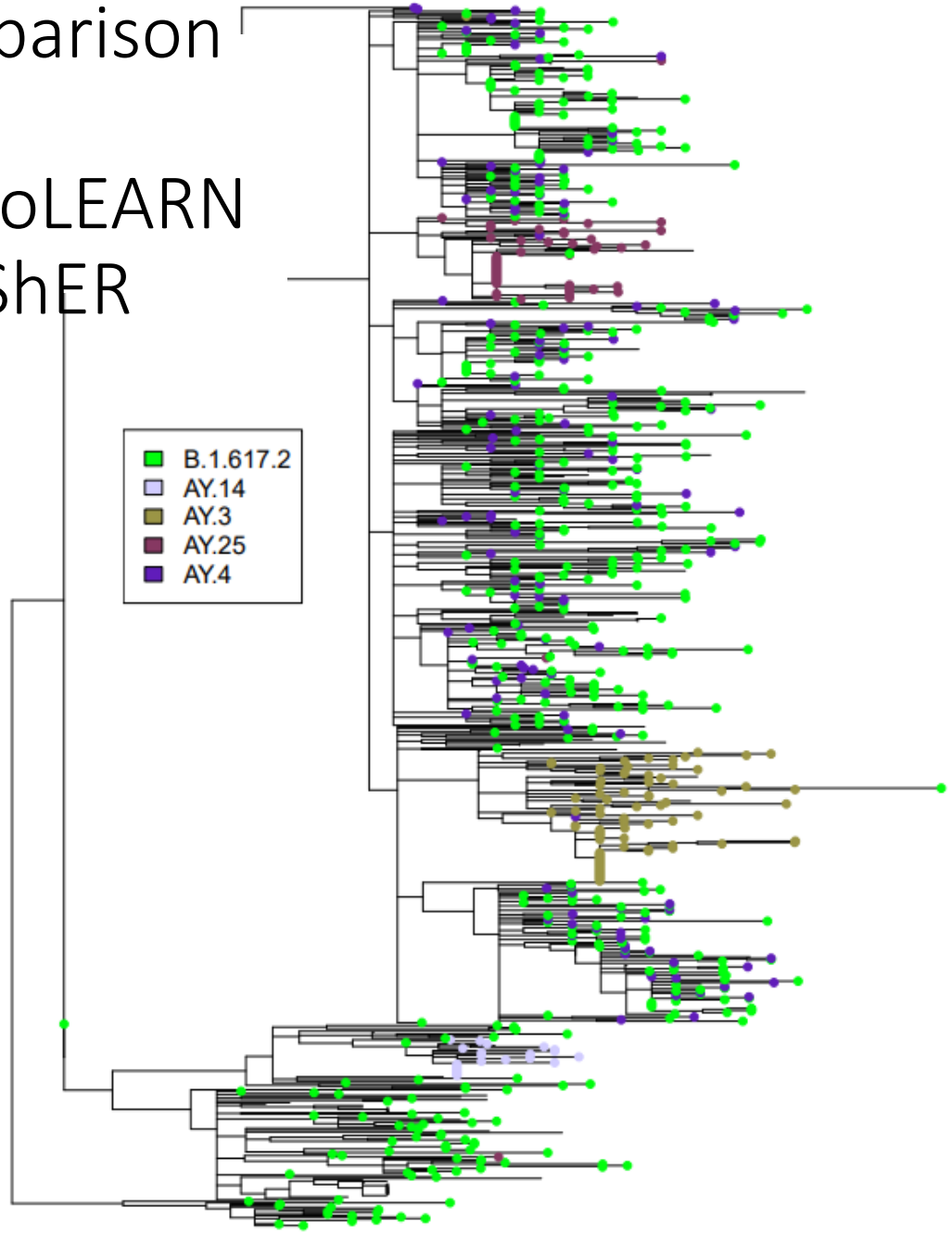
6 rounds of pruning: 853 samples



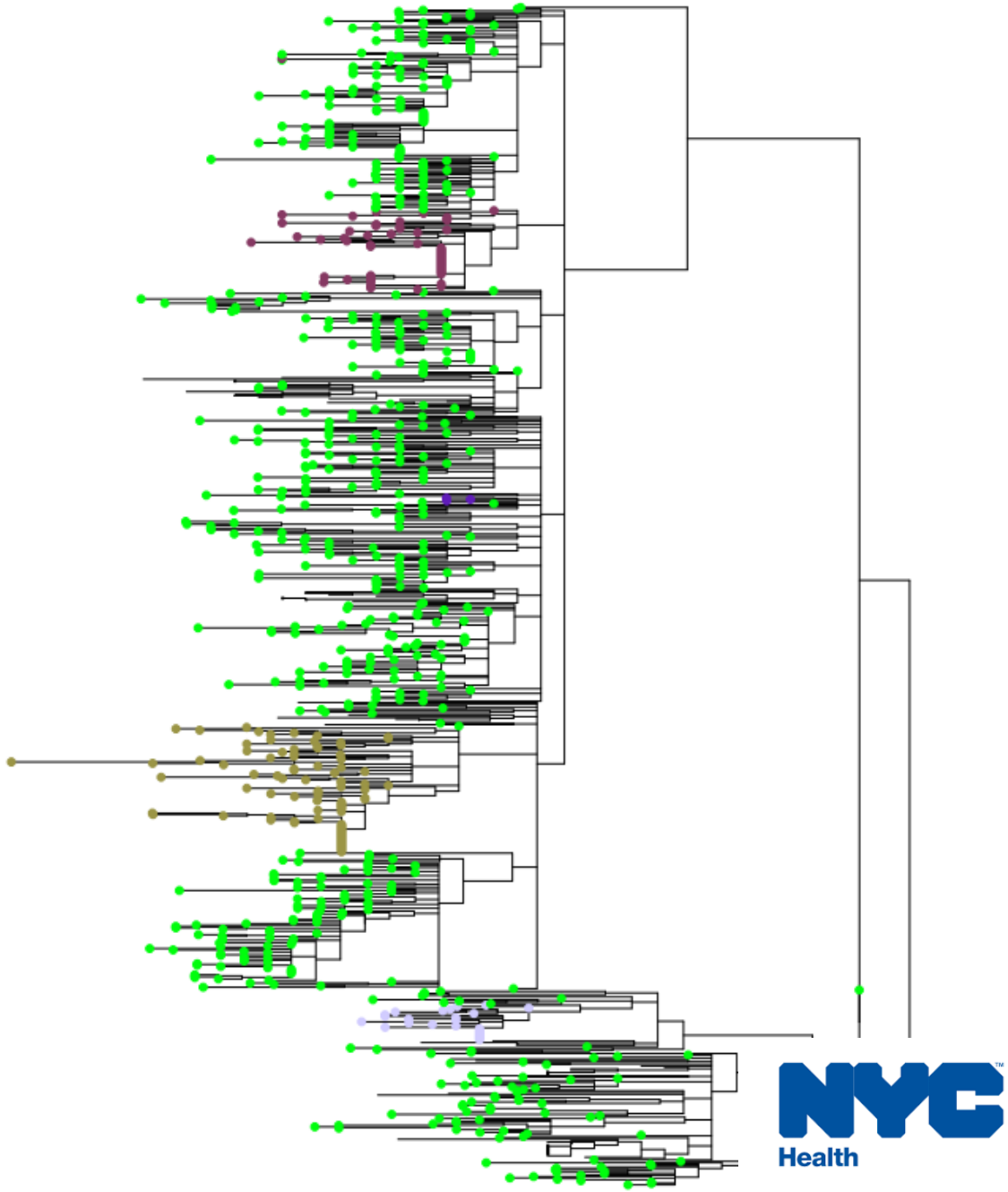


Comparison of pangoLEARN vs UShER

Delta and sublineages – pangoLEARN 09.22



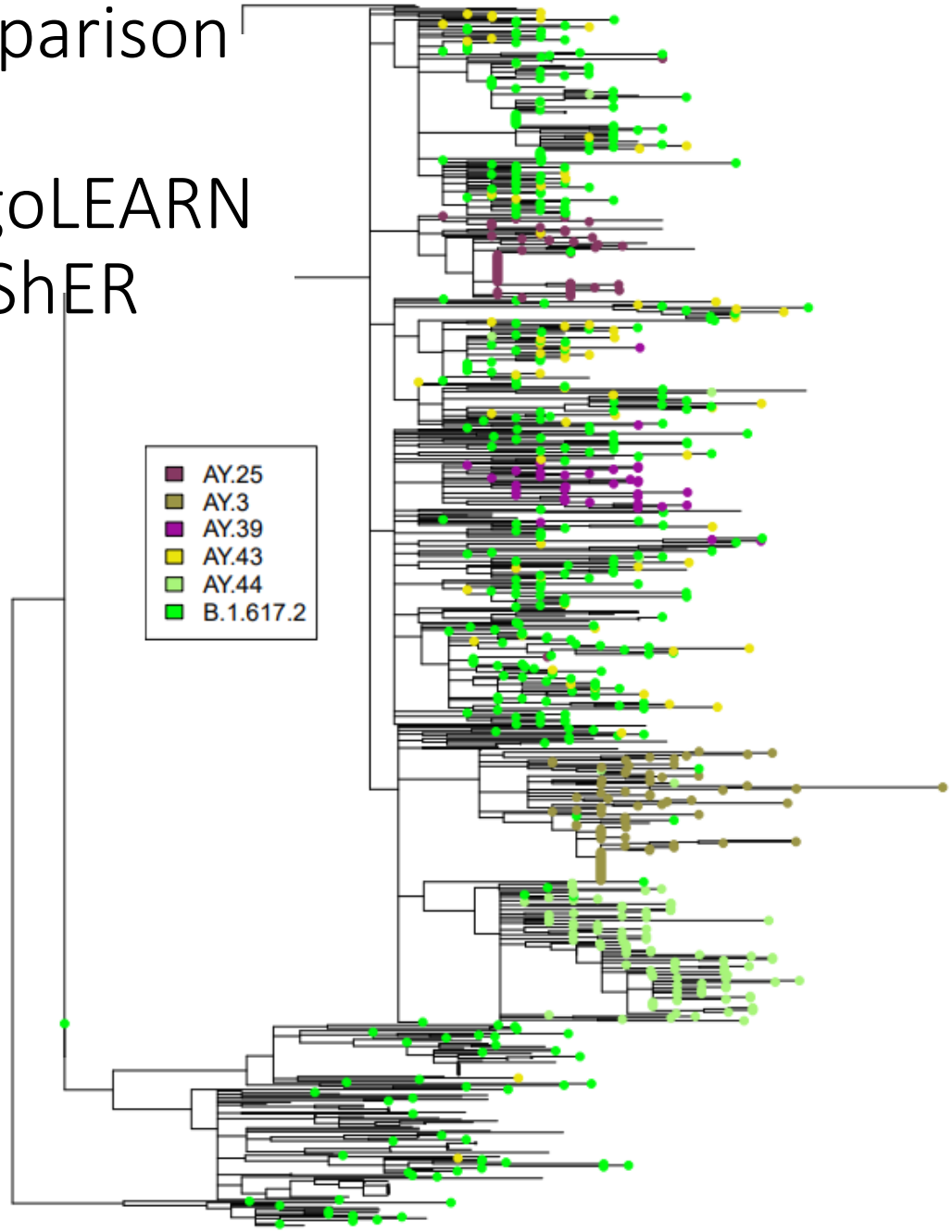
Delta and sublineages – UShER 09.22



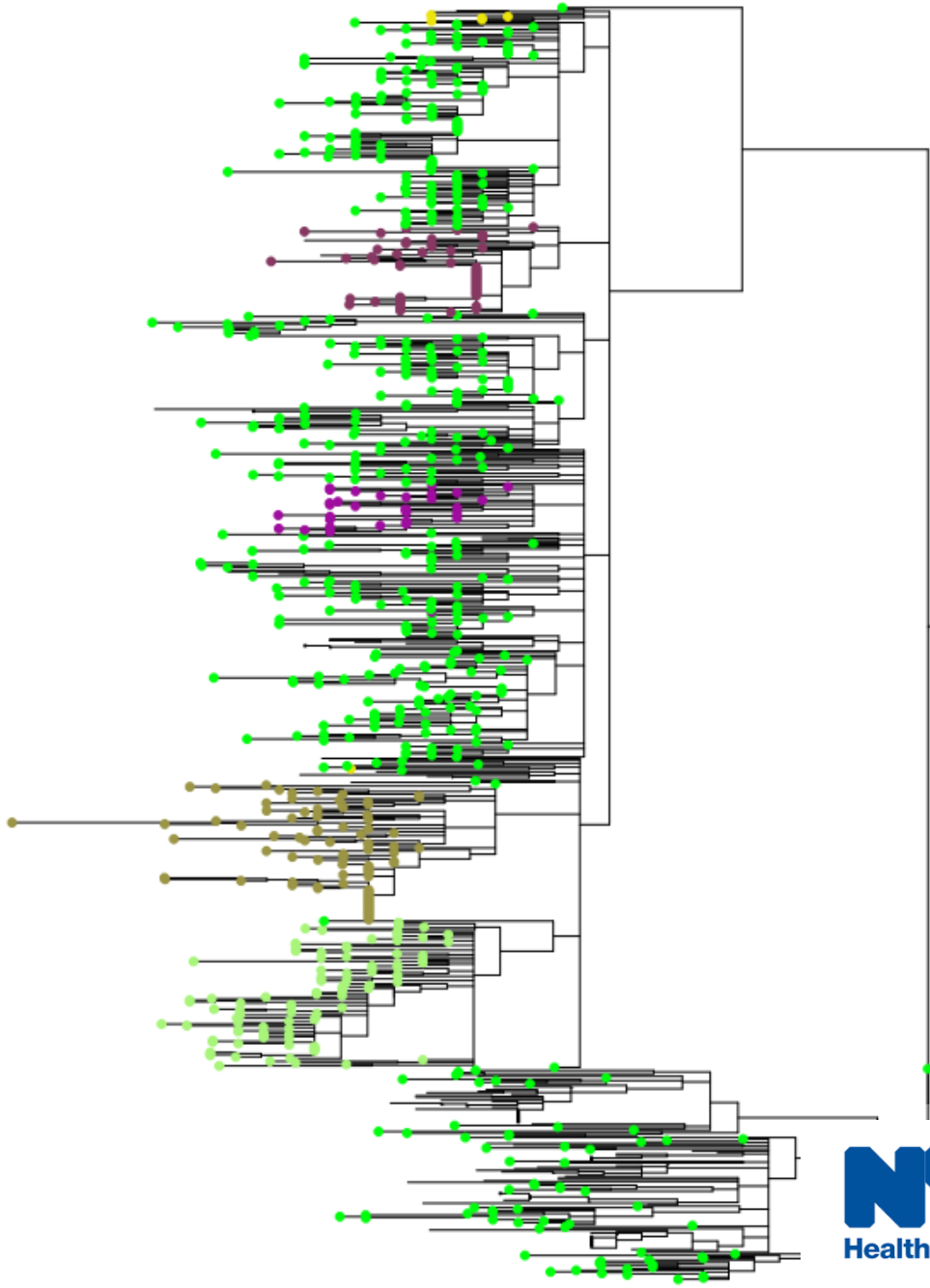
NYC Health

Comparison of pangoLEARN vs UShER

Delta and sublineages – pangoLEARN 11.04

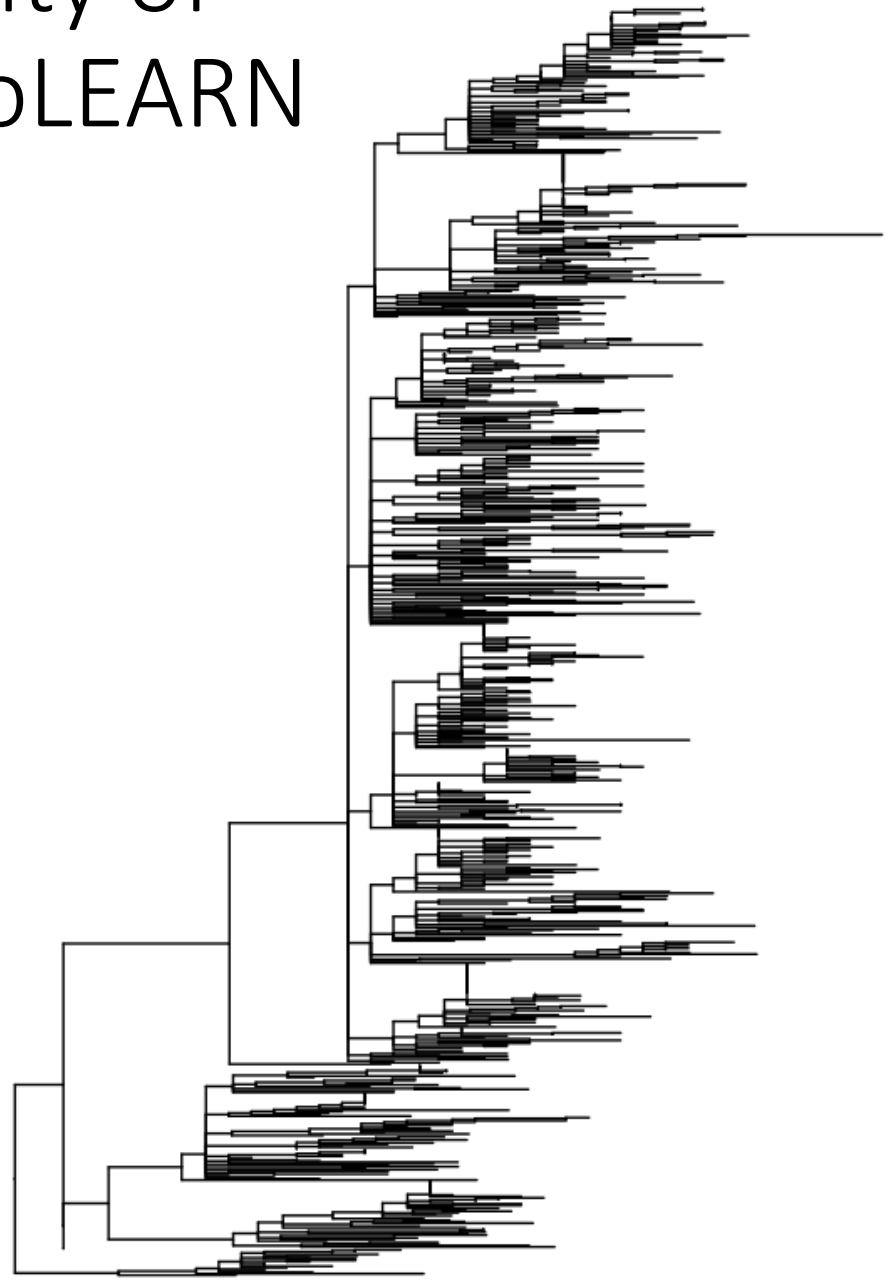


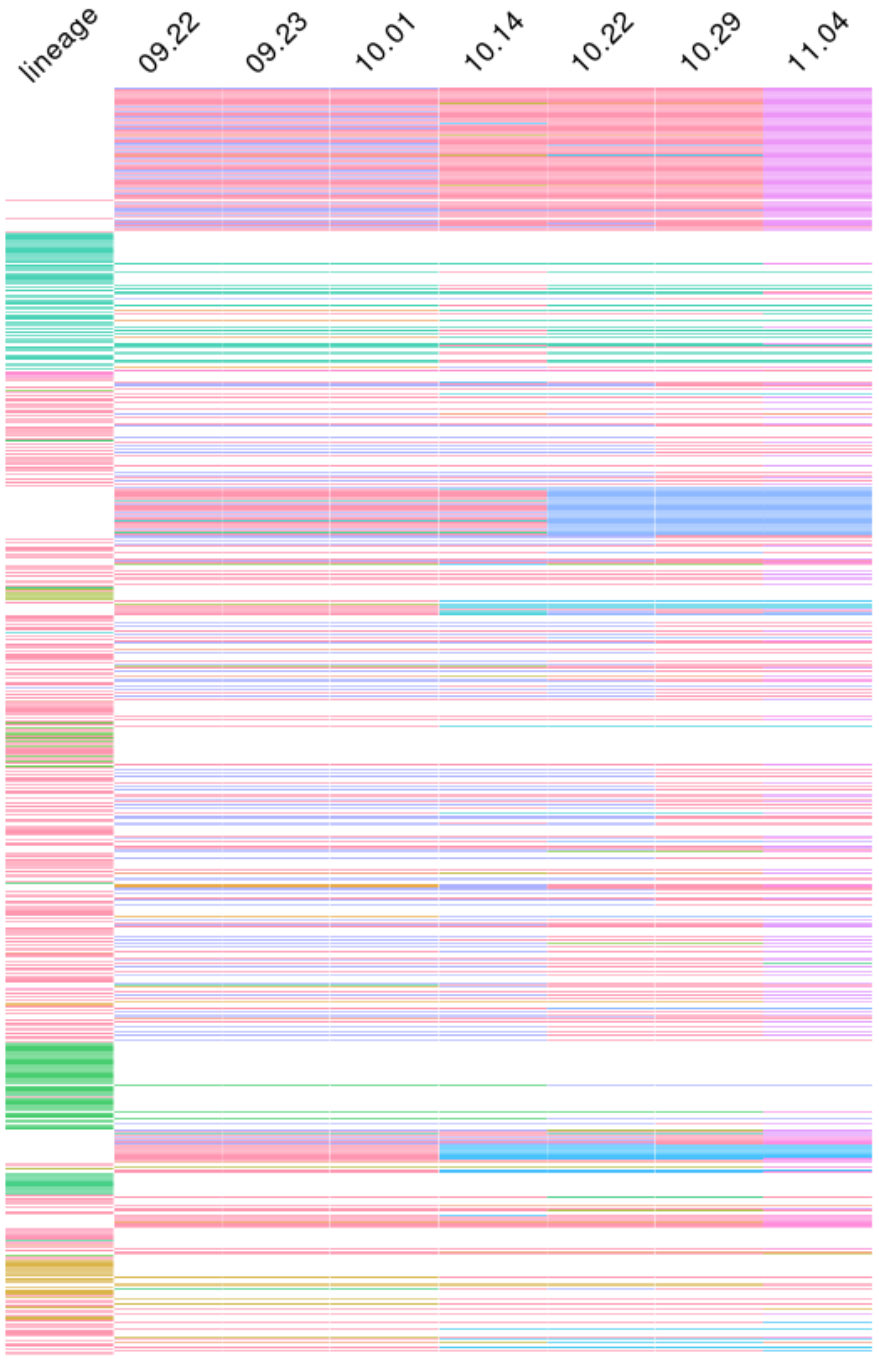
Delta and sublineages – UShER 11.04



NYC Health

Stability of pangoLEARN calls



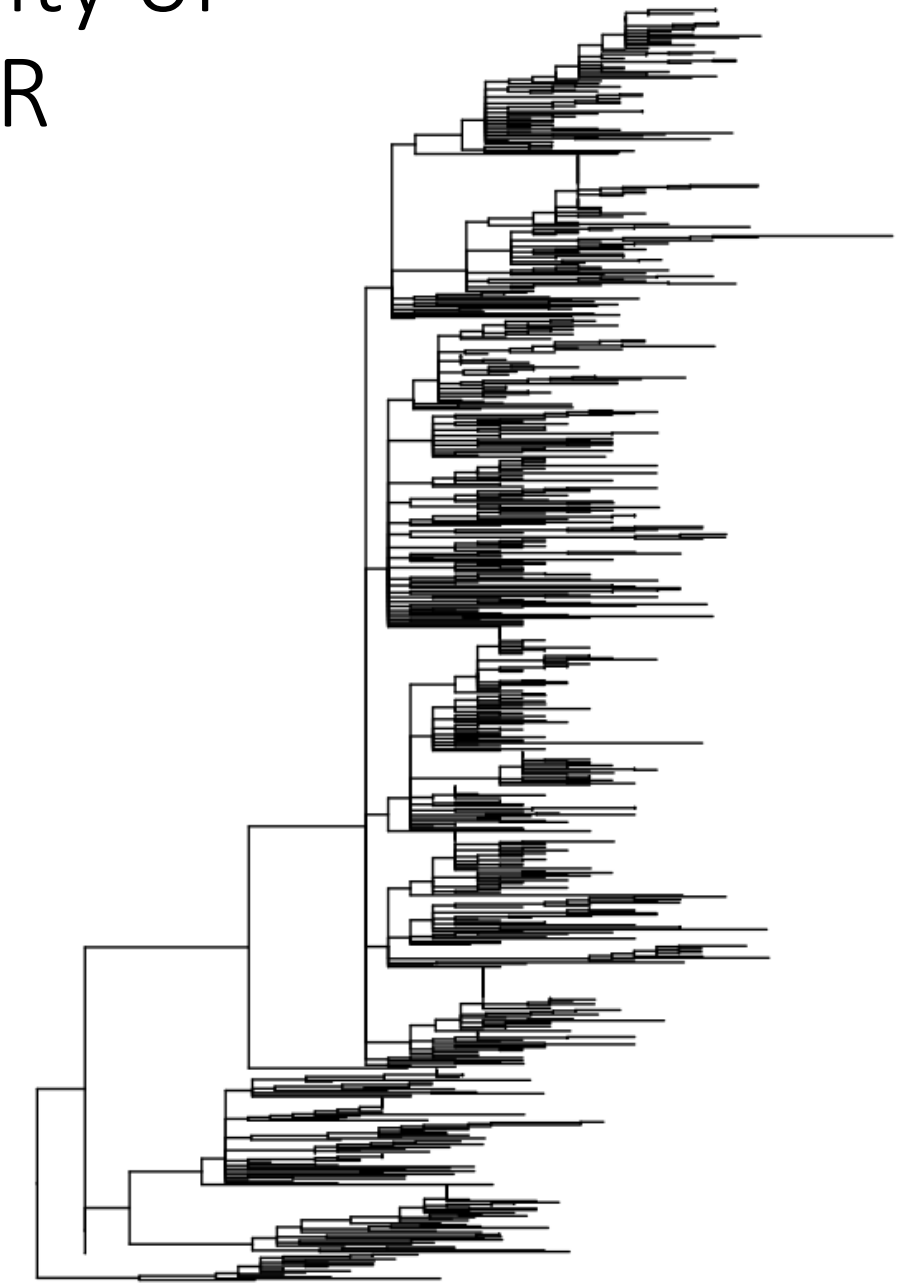


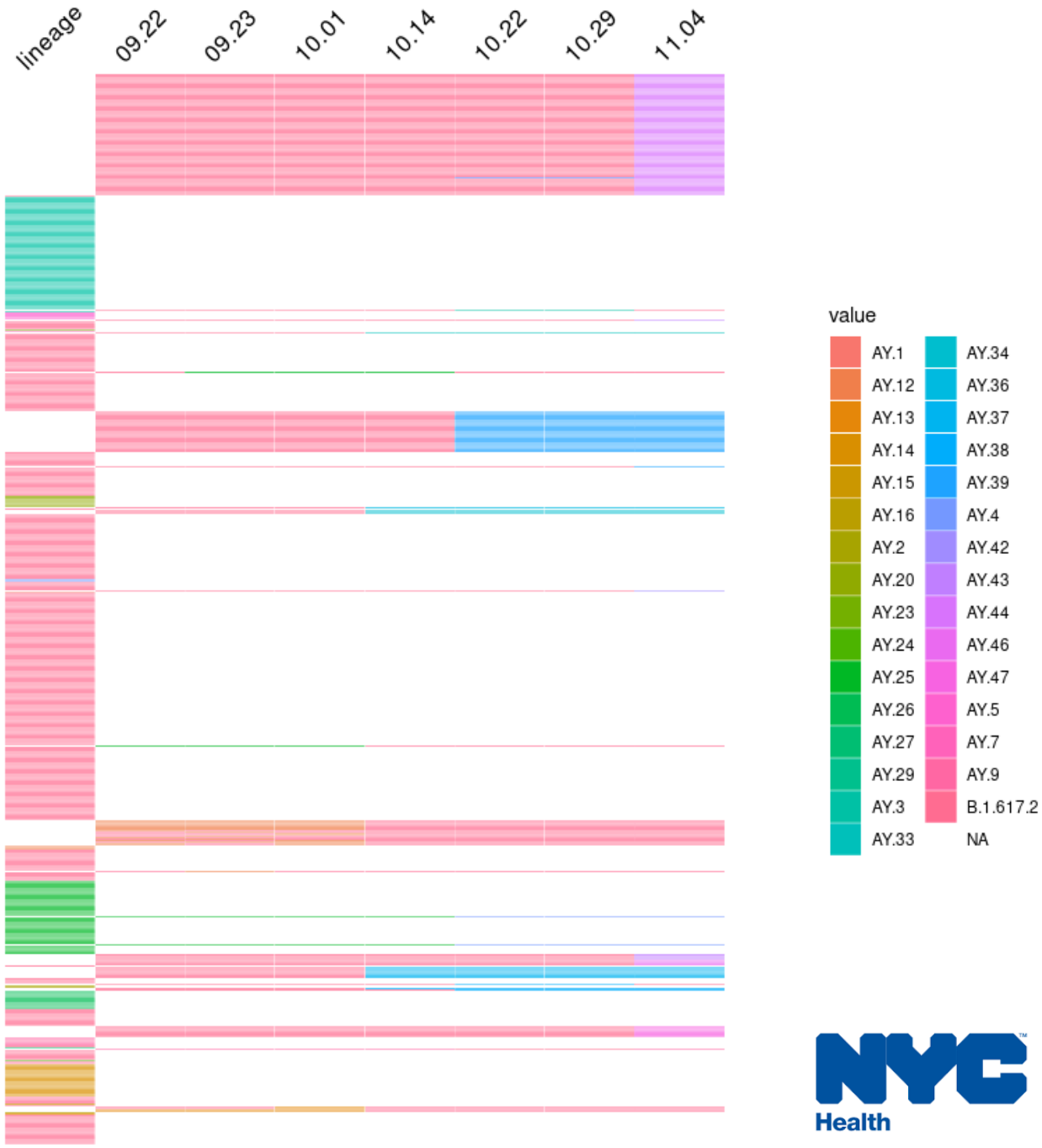
value

AY.1	AY.33
AY.10	AY.34
AY.11	AY.35
AY.12	AY.36
AY.13	AY.37
AY.14	AY.38
AY.15	AY.39
AY.16	AY.4
AY.2	AY.41
AY.20	AY.42
AY.21	AY.43
AY.23	AY.44
AY.24	AY.45
AY.25	AY.46
AY.26	AY.47
AY.27	AY.5
AY.29	AY.7
AY.3	AY.9
AY.30	B.1.617.2
AY.32	NA

NYC Health

Stability of UShER calls





NYC Health

Scorpio influence on lineage calls

- 207 samples called by pangoLEARN had lineage calls overwritten by scorpio at least once compared to 36 samples called by UShER
  - Most common lineages replaced for pangoLEARN: AY.4, P.1, AY.25
  - Most common lineages replaced for UShER: B.1, AY.25

Conclusions

- pangoLEARN has become more consistent over the time period studied
  - There is a learning period for new emerging lineages (e.g AY.4, AY.43)
- UShER is more stable over time than pangoLEARN
- Scorpio is used more often to “hotfix” pangoLEARN calls than UShER calls