UShER and autolin: Identifying viral lineages

Angie Hinrichs, UCSC

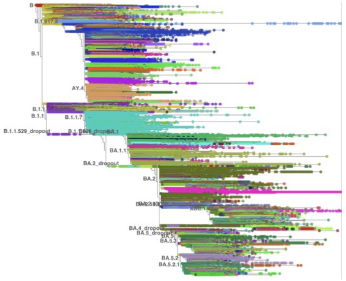
Viral Sub-Species Classification Workshop

April 8-10, 2024

UC SANTA CRUZ | Genomics Institute

Overview

- UShER: Ultrafast Sample placement in Existing tRee
- UCSC's daily updated UShER tree of 16 million SARS-CoV-2 genomes
- UShER in the Pango lineage ecosystem
- **`autolin`**: automating discovery of new lineages



Pandemic phylogenetics is different

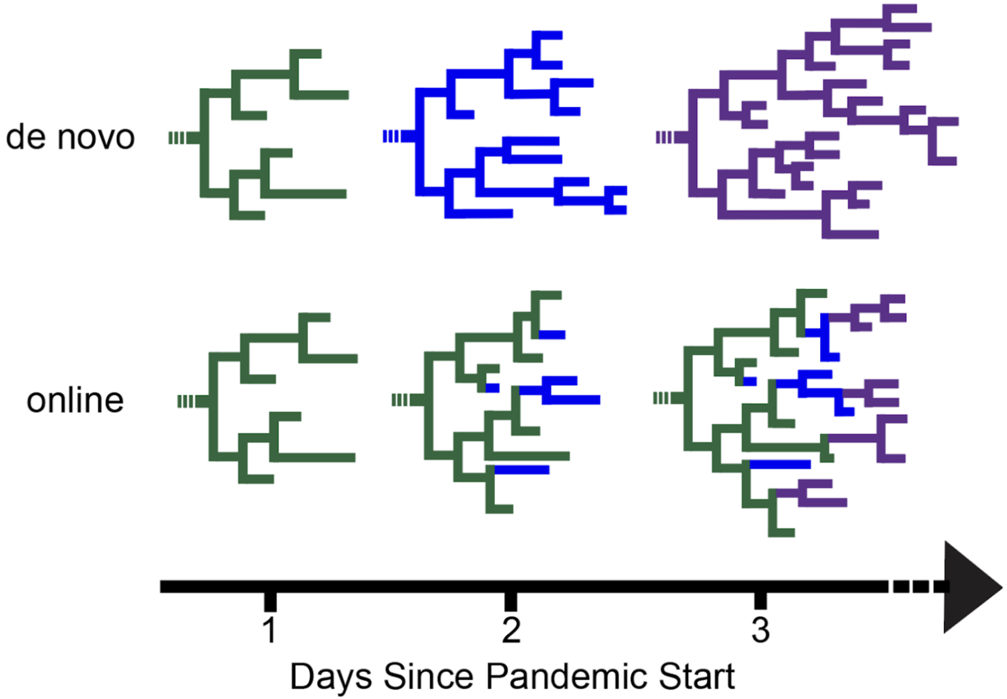
Traditional phylogenetics:

- Thousands of genomes over decades
- Highly diverged genomes
- Maximum likelihood estimation…

Pandemic phylogenetics:

- Tens of thousands of genomes per day
- Many similar sequences
- … would be too slow

UShER is an Online Phylogenetics Application



Yatish Turakhia,
UCSC → UCSD

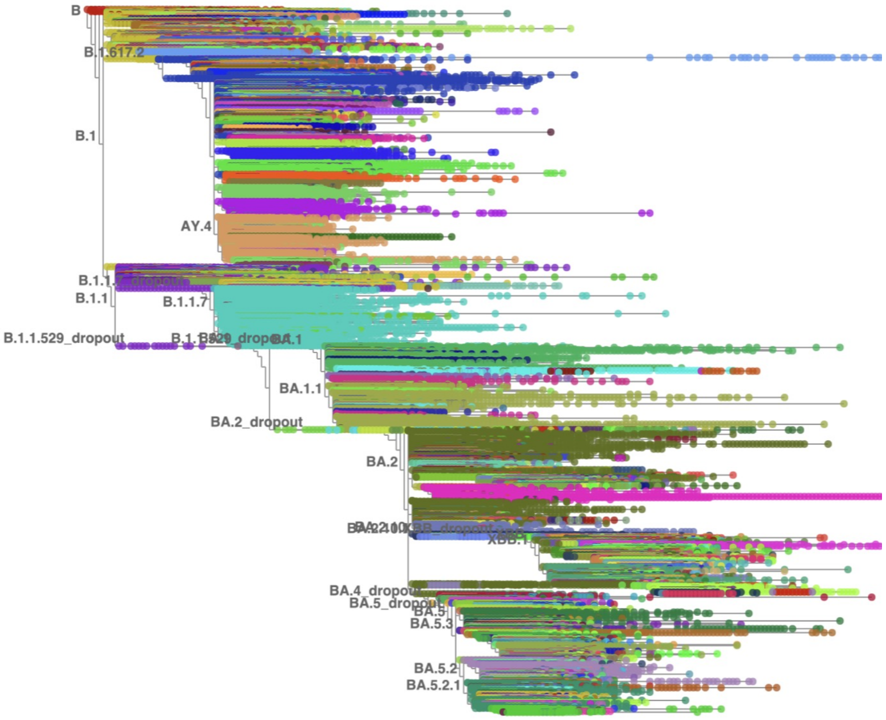
Cheng Ye, UCSD

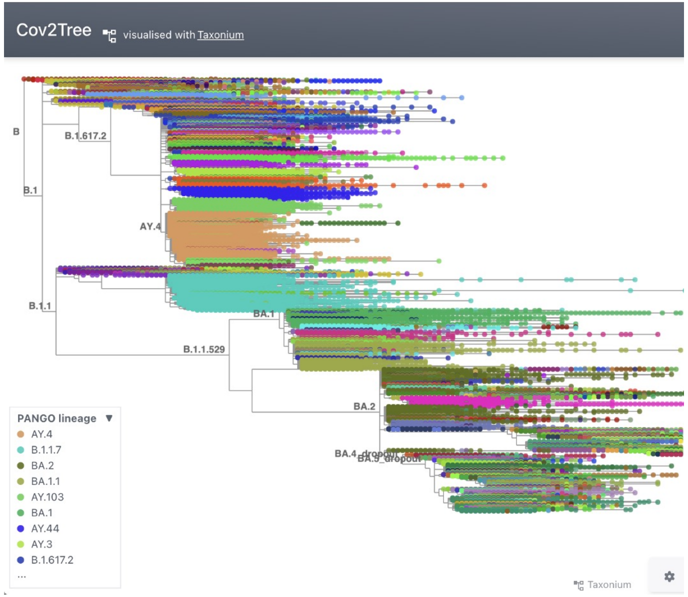
Jakob McBroome,
UCSC (graduated)

Russ Corbett-Detig, UCSC

UC SANTA CRUZ | Genomics Institute

UCSC UShER tree: 16 million genomes and counting





UC SANTA CRUZ | Genomics Institute

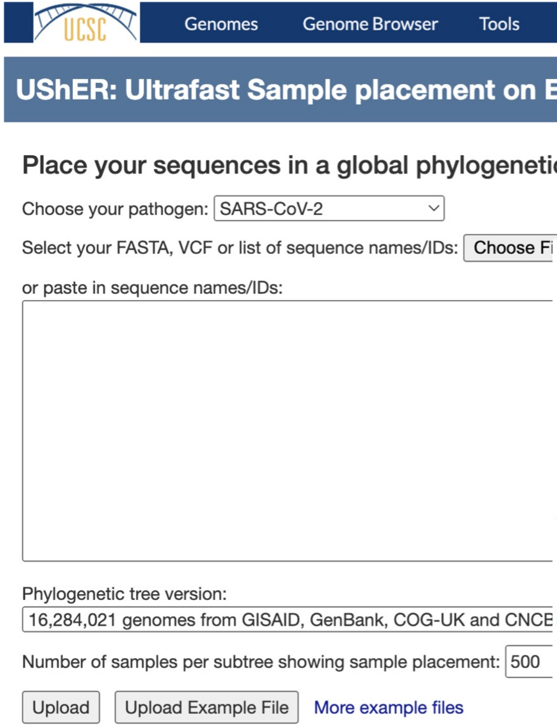
McBroome *et al. Mol Biol Evol.* 2021
https://doi.org/10.1093/molbev/msab264

UShER web interface ~ https://usher.bio/

Upload sequences →

or

Paste in names or IDs →



UShER: Ultrafast Sample placement on Existing tRee

Place your sequences in a global phylogenetic tree

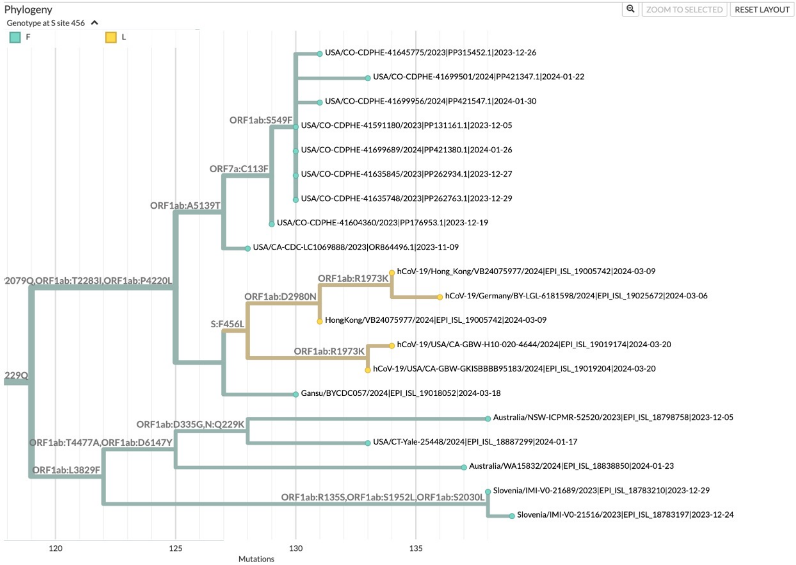
Choose your pathogen: SARS-CoV-2

Select your FASTA, VCF or list of sequence names/IDs: Choose Fi

or paste in sequence names/IDs:

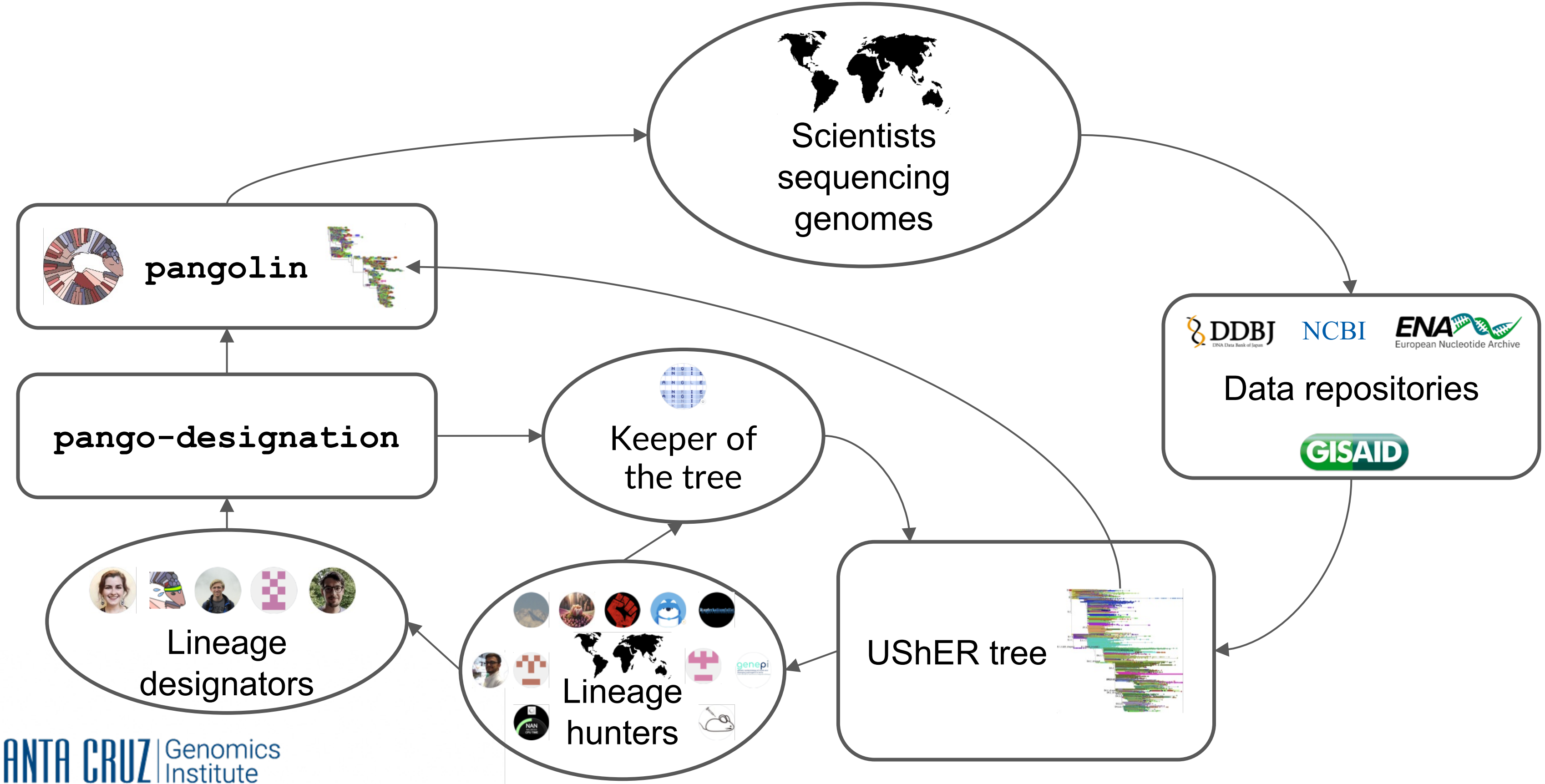
Phylogenetic tree version: 16,284,021 genomes from GISAID, GenBank, COG-UK and CNCE

Number of samples per subtree showing sample placement: 500



UC SANTA CRUZ Genomics Institute

UShER's roles in the Pango lineage ecosystem



Scientists sequencing genomes

pangolin

pango-designation

Keeper of the tree

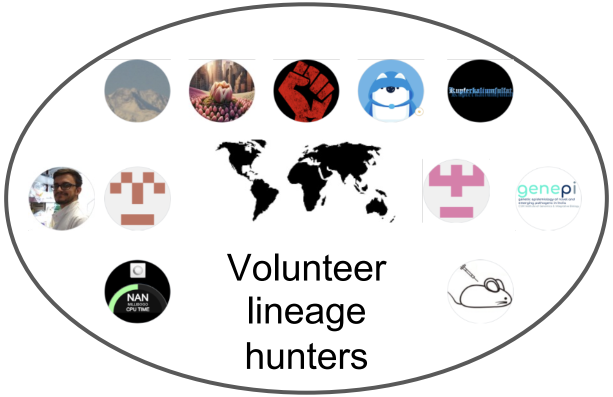
Data repositories

UShER tree

Lineage designators

Lineage hunters

Sustainable? Broadly applicable?



Volunteer lineage hunters

Can we automate (some of) the work?

What are lineage hunters looking for?

→Genetically distinct branches with epidemiological events:

- ✓ Rapid growth
- ✓ Introductions into new geographic regions
- ✓ Interesting mutations (SARS-CoV-2: Spike, immune evasion)
- … Recombination
- − Change in phenotype

`autolin`: automate search for new lineages

- Input: mutation-annotated tree (e.g. UShER, Nextstrain Augur output)
- Identifies branches comparable to what a human would pick out by eye
- Ranks candidate lineages by growth, highly configurable weighting
- Can extend a pre-existing lineage system
- Scalable to SARS-CoV-2 volumes of data

McBroome *et al. Nat Microbiol* 2024 https://doi.org/10.1038/s41564-023-01587-5

Jakob McBroome
UCSC (graduated)

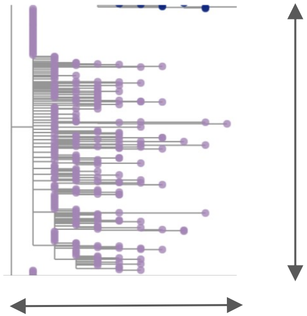
UC SANTA CRUZ | Genomics Institute

`autolin`: finding “lineage-y” branches

Information-theoretic Genotype Representation Index (GRI)

$$\mathrm{GRI} = \frac{N \times D}{\frac{S}{N} + D}$$

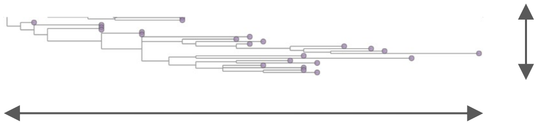
Quickly computed for all nodes in tree



More descendants

Less divergence

→ Higher GRI



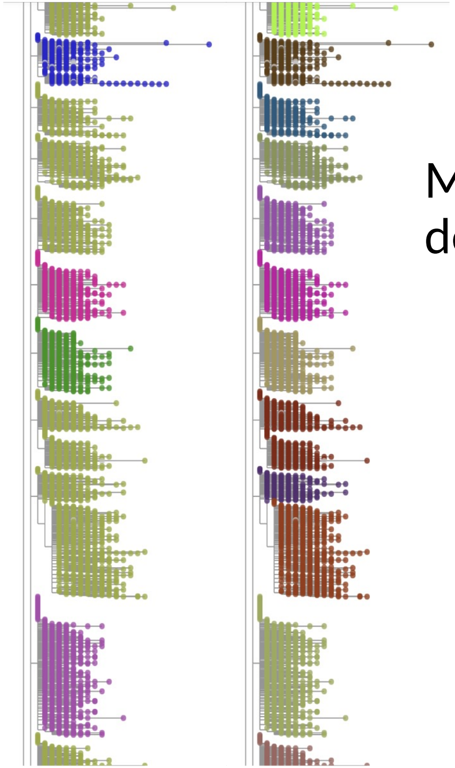
Fewer descendants

More divergence

→ Lower GRI

`autolin`: comparable to human designations

Fewer branches designated



Pango

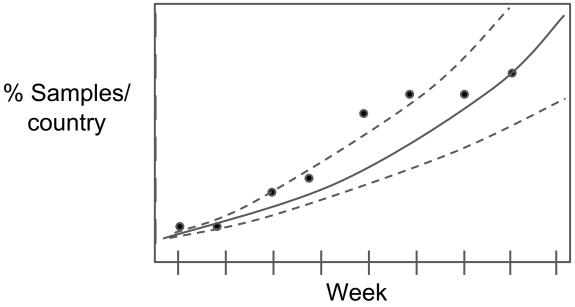
Autolin

More branches designated

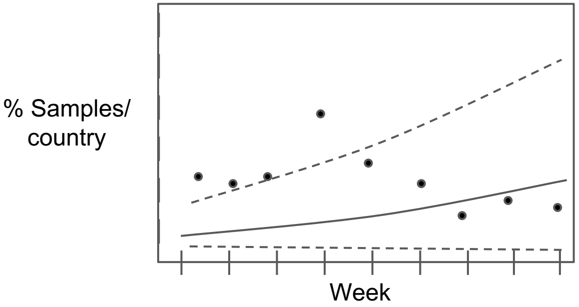
UC SANTA CRUZ | Genomics Institute

autolin: ranking candidate lineages: growth

Lower bound of 95% CI exponential growth fit



Good fit:
Lower bound pretty good



Bad fit:
Lower bound very low

UC SANTA CRUZ | Genomics Institute

`autolin`: ranking candidate lineages: sample weights

USA, UK, Europe are overrepresented

→ proportionally increase weights of other countries' samples

Sample weights can be completely user-defined

`autolin`: ranking candidate lineages: mutations

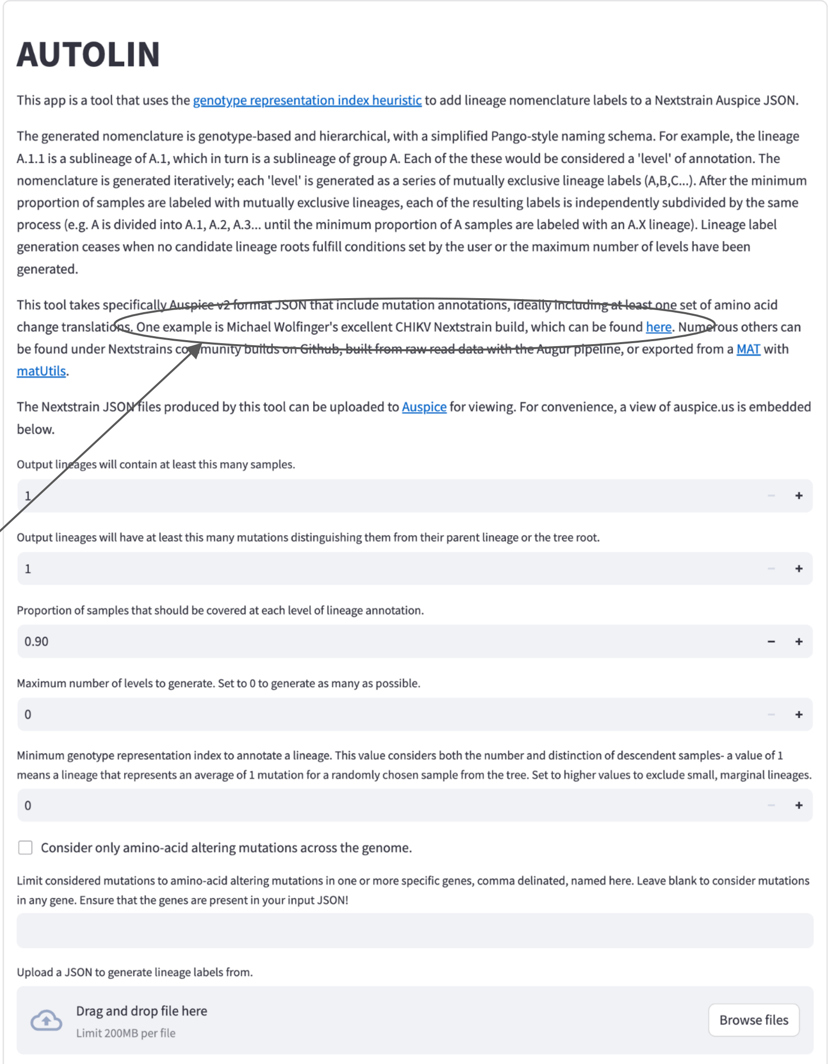
Options:

- Restrict to gene of interest
- Consider only amino-acid changing mutations
- User-defined mutation weights

`autolin` web app

autolin.bio

Just drag & drop the .json from a Nextstrain Augur build!



AUTOLIN

This app is a tool that uses the genotype representation index heuristic to add lineage nomenclature labels to a Nextstrain Auspice JSON.

The generated nomenclature is genotype-based and hierarchical, with a simplified Pango-style naming schema. For example, the lineage A.1.1 is a sublineage of A.1, which in turn is a sublineage of group A. Each of the these would be considered a 'level' of annotation. The nomenclature is generated iteratively; each 'level' is generated as a series of mutually exclusive lineage labels (A,B,C...). After the minimum proportion of samples are labeled with mutually exclusive lineages, each of the resulting labels is independently subdivided by the same process (e.g. A is divided into A.1, A.2, A.3... until the minimum proportion of A samples are labeled with an A.X lineage). Lineage label generation ceases when no candidate lineage roots fulfill conditions set by the user or the maximum number of levels have been generated.

This tool takes specifically Auspice v2 format JSON that include mutation annotations, ideally including at least one set of amino acid change translations. One example is Michael Wolfinger's excellent CHIKV Nextstrain build, which can be found here. Numerous others can be found under Nextstrains community builds on Github, built from raw read data with the Augur pipeline, or exported from a MAT with matUtils.

The Nextstrain JSON files produced by this tool can be uploaded to Auspice for viewing. For convenience, a view of auspice.us is embedded below.

Output lineages will contain at least this many samples.

1 − +

Output lineages will have at least this many mutations distinguishing them from their parent lineage or the tree root.

1 − +

Proportion of samples that should be covered at each level of lineage annotation.

0.90 − +

Maximum number of levels to generate. Set to 0 to generate as many as possible.

0 − +

Minimum genotype representation index to annotate a lineage. This value considers both the number and distinction of descendent samples- a value of 1 means a lineage that represents an average of 1 mutation for a randomly chosen sample from the tree. Set to higher values to exclude small, marginal lineages.

0 − +

☐ Consider only amino-acid altering mutations across the genome.

Limit considered mutations to amino-acid altering mutations in one or more specific genes, comma delinated, named here. Leave blank to consider mutations in any gene. Ensure that the genes are present in your input JSON!

Upload a JSON to generate lineage labels from.

UC SANTA CRUZ | Genomics Institute

Online Analysis is Great for…

1. **Collaboration** - findings and analyses are comparable.
2. **Scalability** - only add a fraction of the data to an ever-growing analysis object.
3. **Reproducibility** - archived, curated, and documented analysis.
4. **Equity** - resource-limited researchers can obtain comprehensive analysis at a fraction of the cost.

https://rdcu.be/duEFP

Russ Corbett-Detig, UCSC

Conclusions

We have great tools for going from genomes to lineages

→ Support open sharing of pathogen genomes

→ Support lineage system maintenance

Acknowledgements

Team UShER (UC Santa Cruz):

- Russ Corbett-Detig
- Jakob McBroome
- Alex Kramer
- Bryan Thornlow
- Nicolas Ayala
- Adriano de Bernardi Schneider
- Lily Karim
- Koorous Vargha
- Jeltje van Baren
- Jen Martin
- Marc Perry
- David Haussler

Team UShER (UC San Diego): UCSD

- Yatish Turakhia
- Cheng Ye
- Kyle Smith
- Sumit Walia
- Devika Torvi
- Shoh Mollenkamp

Australian National University Rob Lanfear

EMBL-EBI Nick Goldman, Nicola de Maio

Crick Inst. Theo Sanderson

Funding: CDC NIH SCHMIDT FUTURES **Pat & Rowland Rebele**

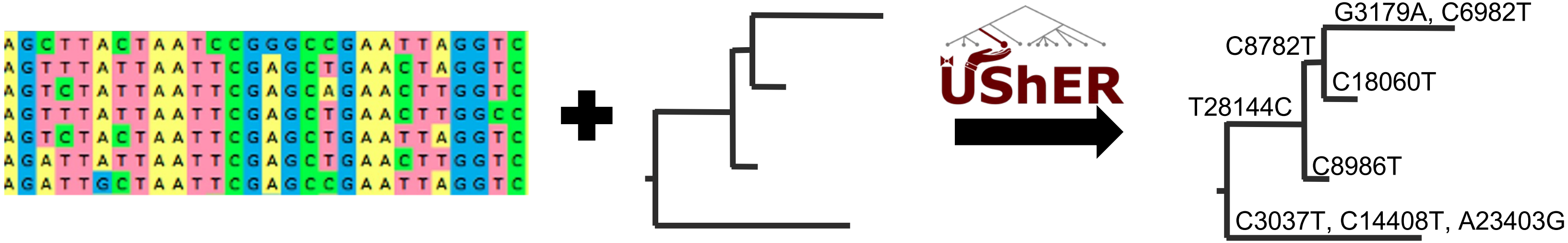
Genomes: GISAID DDBJ INSDC GenBank NCBI ENA EMBL-EBI COG-UK CNCB

UC SANTA CRUZ Genomics Institute

UShER: what makes it Ultrafast?

Yatish Turakhia
UCSC → UCSD

~~1. Full MSA~~ → compact binary-encoded Mutation-Annotated Tree



```
AGCTTACTAATCCGGGCCGAATTAGGTC
AGTTTATTAATTCGAGCTGAACTAGGTC
AGTCTATTAATTCGAGCAGAACTTGGTC
AGTTTATTAATTCGAGCTGAACTTGGCC
AGTCTACTAATTCGAGCTGAATTAGGTC
AGATTATTAATTCGAGCTGAACTTGGTC
AGATTGCTAATTCGAGCCGAATTAGGTC
```

~~1. Maximum likelihood estimation~~ → parsimony

2. Utilize all the CPUs

Cheng Ye
UCSD

Turakhia *et al*. *Nature Genetics* 2021. https://doi.org/10.1038/s41588-021-00862-7

UC SANTA CRUZ Genomics Institute