

Session 4

Alignment and phylogenetic tree building

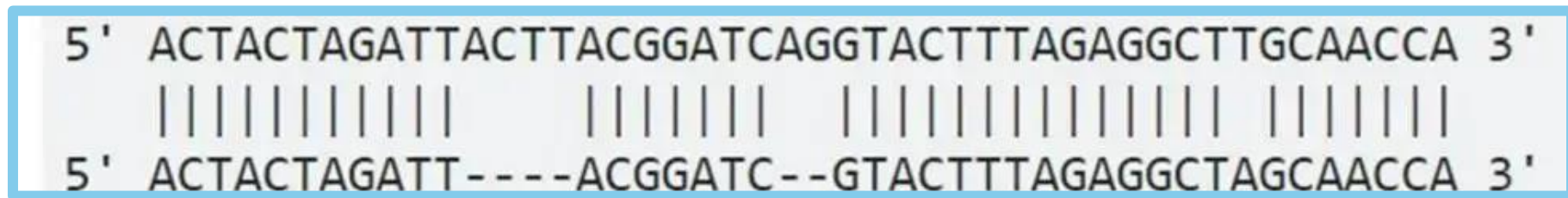
Alignment

What is sequence alignment?

- The process of lining up DNA, RNA, or protein sequences to identify similarities
- Similarities between sequences indicate likely evolutionary relationships between sequences
- After aligning our sequence, we can apply statistical models to investigate evolutionary relationships

Pair-wise Sequence Alignment

- Insert gaps in one of the two sequences to allow the two sequences to line up with one another.
- Works by maximizing matching characters while minimizing mismatches and indels (insertions and deletions)
- Several algorithms exist, focusing on either global alignment (the full length of the sequence) or local alignment (centered on similar regions)



```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
  |||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

The diagram illustrates a pair-wise sequence alignment between two DNA sequences. The top sequence is 5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'. The bottom sequence is 5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'. Vertical bars (|) indicate matching characters between the two sequences. The bottom sequence contains four gaps (represented by dashes) to align with the top sequence: a 4-base gap after 'ACTACTAGATT', a 2-base gap after 'ACGGATC', and a 1-base gap after 'GTACTTTAGAGGCT'.

Multiple sequence alignment (MSA)

- Aligning three or more sequences simultaneously
- A couple of strategies are common:
 - Progressive alignment: initially align pairs of sequences and add on additional sequences afterwards one-by-one
 - Iterative alignment: iteratively build the alignment with subsets of sequences, building small alignments then revising them as you merge the subsets
- MSA is much slower than pairwise alignment, and its speed scales with the length of sequences as well as the number of sequences involved

Alignment can be easy or difficult

GCGGCCCA	TCAGGTAGTT	GGTGG
GCGGCCCA	TCAGGTAGTT	GGTGG
GCGTTCCA	TCAGCTGGTT	GGTGG
GCGTCCCA	TCAGCTAGTT	GGTGG
GCGGCGCA	TTAGCTAGTT	GGTGA
*****	*****	*****

Easy

TTGACATG	CCGGGG---A	AACCG
TTGACATG	CCGGTG--GT	AAGCC
TTGACATG	-CTAGG---A	ACGCG
TTGACATG	-CTAGGGAAC	ACGCG
TTGACATC	-CTCTG---A	ACGCG
*****	??????????	*****

Difficult due
to insertions
or deletions
(indels)

Key terms for MSA

- **Consensus sequence:** the most representative sequence of the alignment, based on the most frequent bases at each site
- **Site position:** the position number associated with a nucleotide with respect to the alignment
- **Gaps:** the sites with a “-”, these indicate gaps between aligned regions, inserted to help improve the overall alignment
- **Indel:** insertions and deletions, these cause gaps in the overall alignment and represent larger regions of poor alignment
- **% identity:** the percentage of bases at a site that match the consensus sequence

What MSA tools to use and when?

- MUSCLE: progressive MSA suitable for medium-large datasets up to 1000 sequences
- MAFFT: progressive MSA suitable for large datasets, up to 30,000 sequences or a several long sequences (if closely related)
 - This is the one I use the most!
- Clustal Omega: a Hidden Markov Model alignment tool designed for very large datasets

Genetic Distance

Distance Methods

1. Calculate the distance matrix

1	ATGACTTAGATTGGCTCGTACATCCGGTA
2	ATCACTTATATCAACTCGAACAGCCGTTA
3	ATGGGTTCGAGGAGCTCGTAGAGCCGTTA
4	ATGACTTCGATTAGCTCGTAGAGCCGGTA

Distance Matrix

	1	2	3	4
1	-			
2		-		
3			-	
4				-

Distance Methods

1. Calculate the distance matrix

1	at G actta G at TGG ctcg T aca T ccg G ta
2	at C actta T at CA actcg A aca G ccg T ta

Distance Matrix

	1	2	3	4
1	-	8		
2		-		
3			-	
4				-

Distance Methods

1. Calculate the distance matrix

1	atg A CttAga TTG gctcgtaCa T ccg G ta
3	atg G GttCga GGA gctcgta GaG ccg T ta

	1	2	3	4
1	-	8	9	
2		-		
3			-	
4				-

Distance Methods

1. Calculate the distance matrix

1	atgactt A gatt G gctcgta CaT ccggta
4	atgactt C gatt A gctcgta GaG ccggta

Distance Matrix

	1	2	3	4
1	-	8	9	4
2		-		
3			-	
4				-

Distance Methods

1. Calculate the distance matrix

2	at CAC ttATaTCaActcgAaCagccgtta
3	at GGG ttCGa GGa GctcgTaGagccgtta

Distance Matrix

	1	2	3	4
1	-	8	9	4
2		-	10	
3			-	
4				-

Distance Methods

1. Calculate the distance matrix

2	atCactt A TatCaActcgAaCagccgTta
4	atGactt C GatTaGctcgTaGagccgGta

Distance Matrix

	1	2	3	4
1	-	8	9	4
2		-	10	
3			-	
4				-

Distance Methods

1. Calculate the distance matrix

2	atCacttATatCaActcgAaCagccgTta
4	atGacttCGatTaGctcgTaGagccgGta

Distance Matrix

	1	2	3	4
1	-	8	9	4
2		-	10	8
3			-	
4				-

Distance Methods

1. Calculate the distance matrix

1	ATGACTTAGATTGGCTCGTACATCCGGTA
2	ATCACTTATATCAACTCGAACAGCCGTTA
3	ATGGGTTCGAGGAGCTCGTAGAGCCGGTTA
4	ATGACTTCGATTAGCTCGTAGAGCCGGTA

Distance Matrix

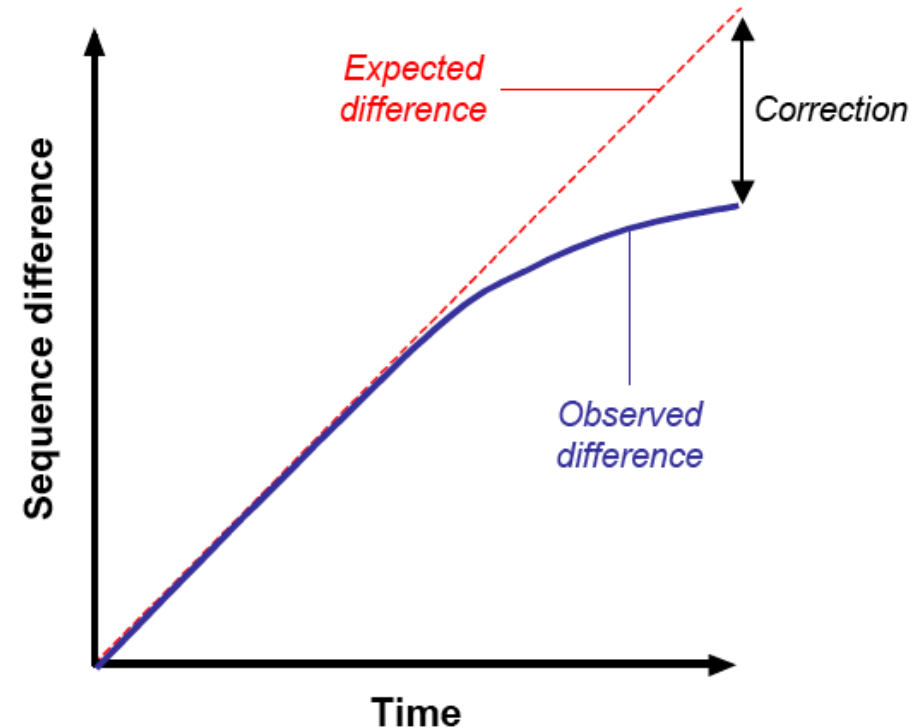
	1	2	3	4
1	-	8	9	4
2		-	10	8
3			-	5
4				-

Saturation and distance correction

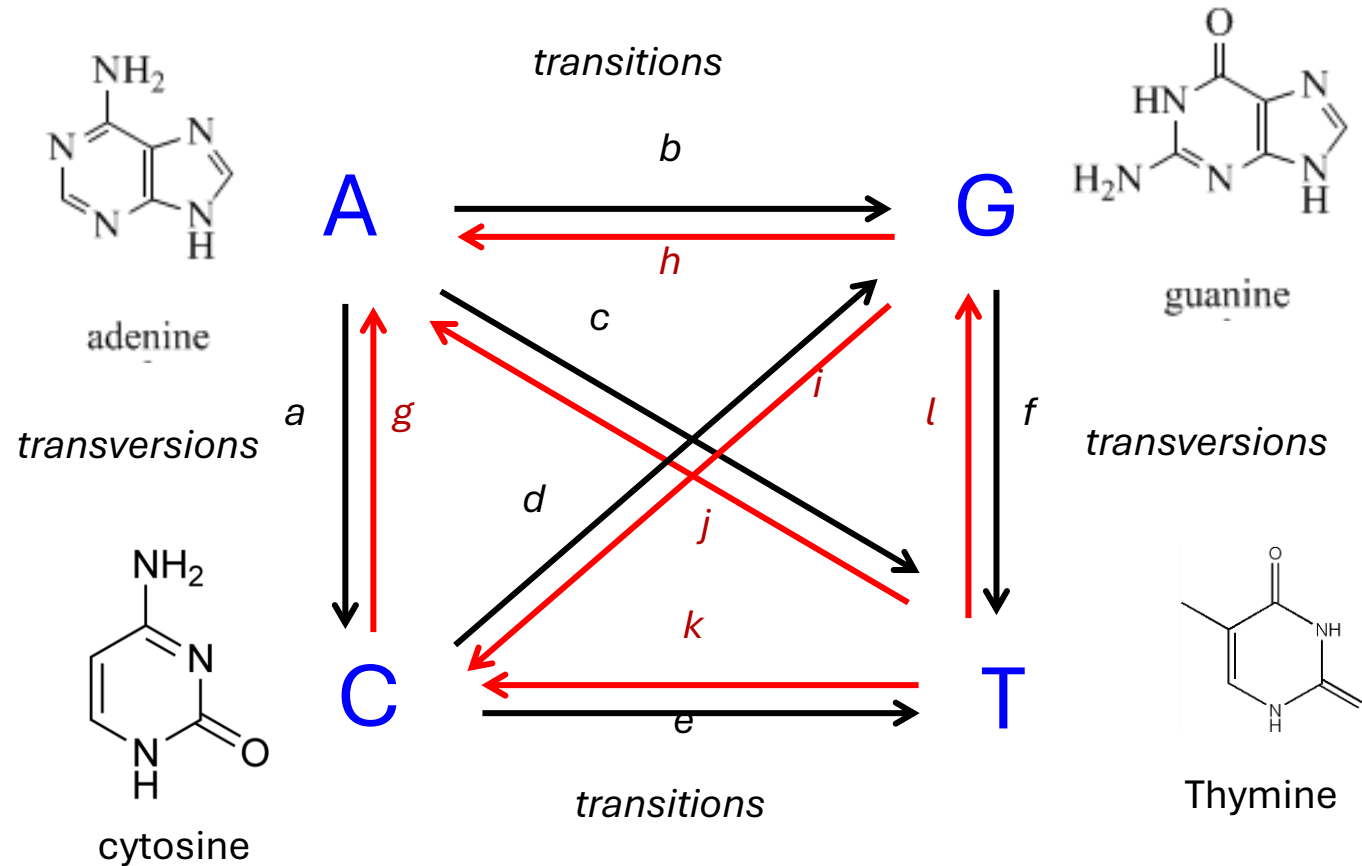
As substitutions are accumulated between two sequences, they become **saturated**: most of the changing sites have changed before

The observed differences between the two sequences is not linear with time, due to **multiple substitutions at the same location**

Observed genetic differences need to be '**corrected**' to recover the changes overprinted by multiple hits



Not all substitutions happen with the same likelihood



Models of evolution

Several probabilistic **models of evolution** have been developed to convert observed distances into measures of actual evolutionary distances

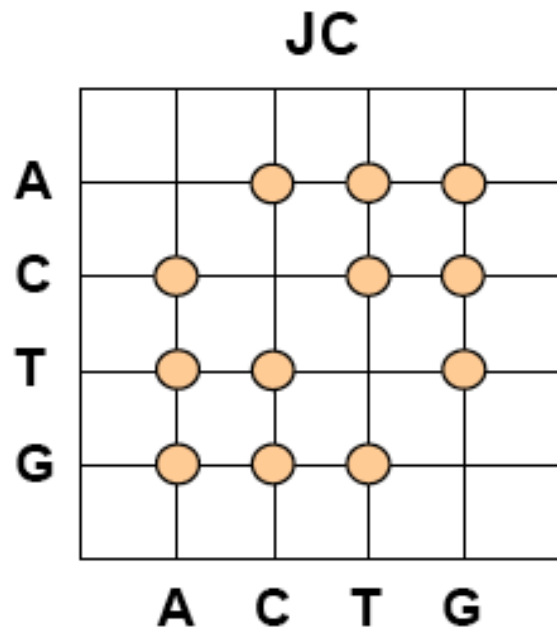
The relative complexity of these models is a function of the extent of the biological, biochemical and evolutionary assumptions (i.e. parameters) they incorporate

Substitutions are usually described as probabilities of mutational events, mathematically modelled by matrices of relative rates:

$$\mathbf{P}_t = \begin{pmatrix} P_{AA} & P_{AC} & P_{AT} & P_{AG} \\ P_{CA} & P_{CC} & P_{CT} & P_{CG} \\ P_{TA} & P_{TC} & P_{TT} & P_{TG} \\ P_{GA} & P_{GC} & P_{GT} & P_{GG} \end{pmatrix}$$

Jukes-Cantor (JC)

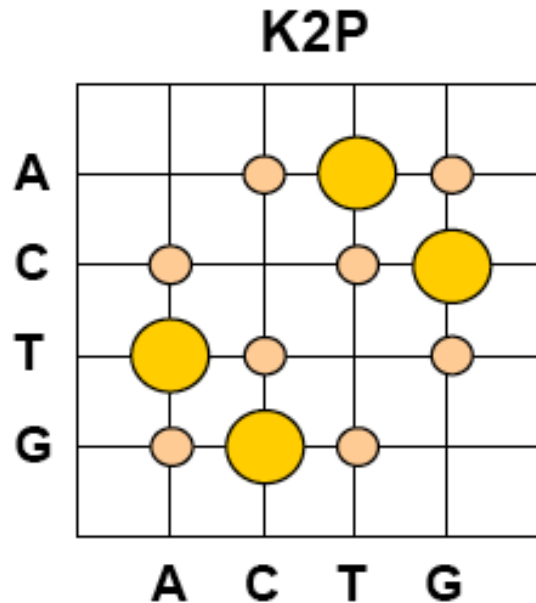
- First proposed model
- It assumes that the four bases have equal frequencies and all substitutions are equally likely



$$\mathbf{P}_t = \left\{ \begin{array}{cccc} - & \alpha & \alpha & \alpha \\ \alpha & - & \alpha & \alpha \\ \alpha & \alpha & - & \alpha \\ \alpha & \alpha & \alpha & - \end{array} \right\}$$

Kimura's 2 parameter

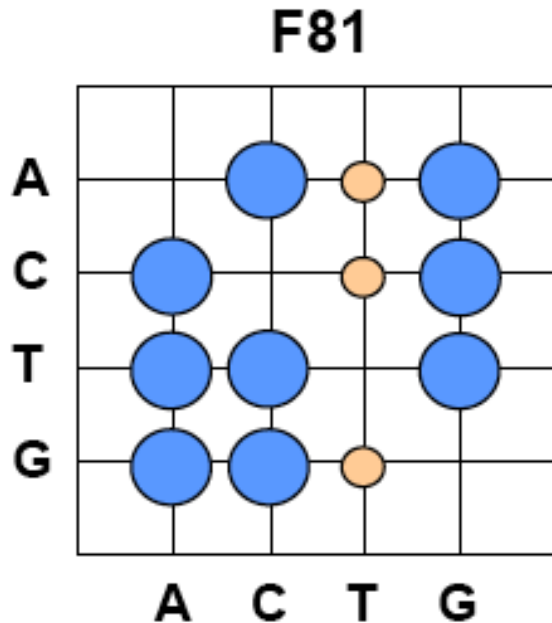
- Transitions are generally more frequent than transversions
- K2P model assumes that the rate of transitions per site (α) differs from the rate of transversions per site (β)



$$\mathbf{P}_t = \begin{Bmatrix} - & \beta & \alpha & \beta \\ \beta & - & \beta & \alpha \\ \alpha & \beta & - & \beta \\ \beta & \alpha & \beta & - \end{Bmatrix}$$

Felsenstein (1981)

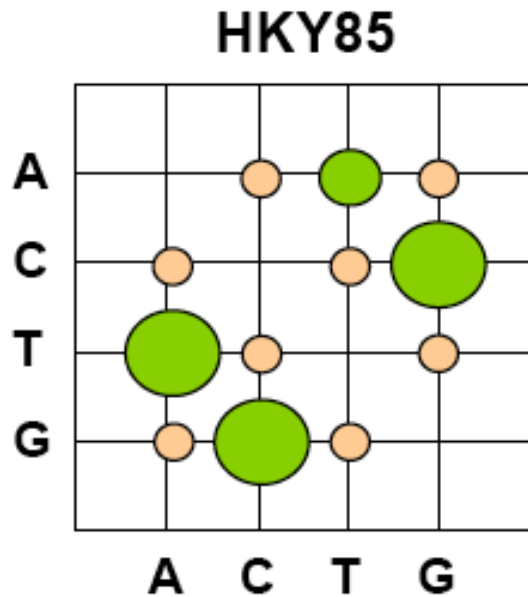
- If some substitutions are more common in one sequence than others, some substitutions may be more frequent than others
- F81 model allows the frequency (π) of the four nucleotides to be different



$$\mathbf{P}_t = \begin{Bmatrix} - & \pi_C \alpha & \pi_T \alpha & \pi_G \alpha \\ \pi_A \alpha & - & \pi_T \alpha & \pi_G \alpha \\ \pi_A \alpha & \pi_C \alpha & - & \pi_G \alpha \\ \pi_A \alpha & \pi_C \alpha & \pi_T \alpha & - \end{Bmatrix}$$

Hasegawa, Kishino and Yano

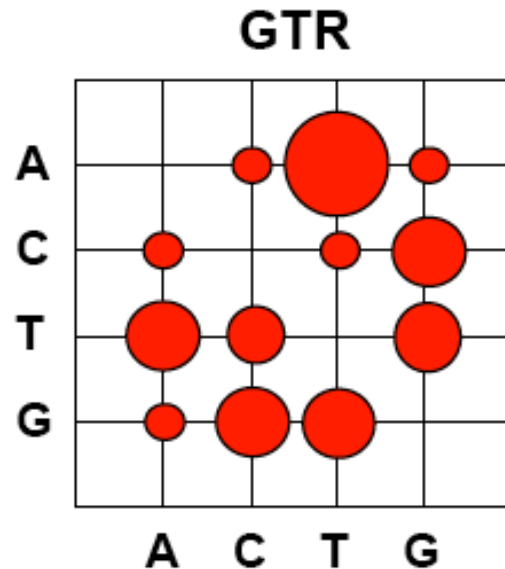
- The HKY85 model allows rates of transitions and transversions to differ and base frequencies to vary



$$\mathbf{P}_t = \begin{Bmatrix} - & \pi_C \beta & \pi_T \alpha & \pi_G \beta \\ \pi_A \beta & - & \pi_T \beta & \pi_G \alpha \\ \pi_A \alpha & \pi_C \beta & - & \pi_G \beta \\ \pi_A \alpha & \pi_C \alpha & \pi_T \beta & - \end{Bmatrix}$$

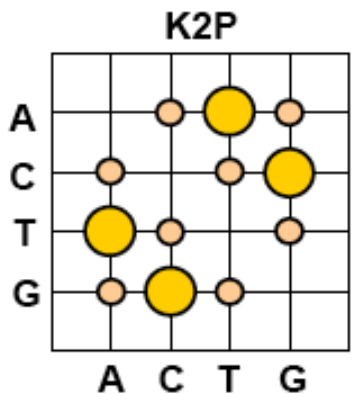
General Time Reversible

- The GTR/REV model allows each possible substitution to have its own probability
- Substitutions are reversible (i.e. substitutions from i to j has the same probability as a substitution from j to i)

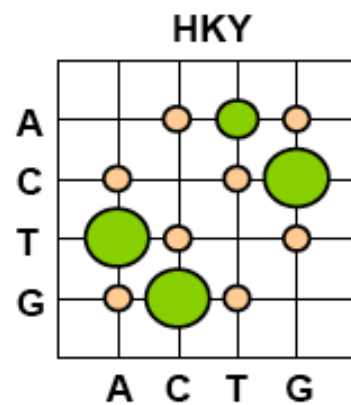


$$\mathbf{P}_t = \begin{Bmatrix} - & \pi_C \mathbf{a} & \pi_T \mathbf{b} & \pi_G \mathbf{c} \\ \pi_A \mathbf{a} & - & \pi_T \mathbf{d} & \pi_G \mathbf{e} \\ \pi_A \mathbf{b} & \pi_C \mathbf{c} & - & \pi_G \mathbf{f} \\ \pi_A \mathbf{d} & \pi_C \mathbf{e} & \pi_T \mathbf{f} & - \end{Bmatrix}$$

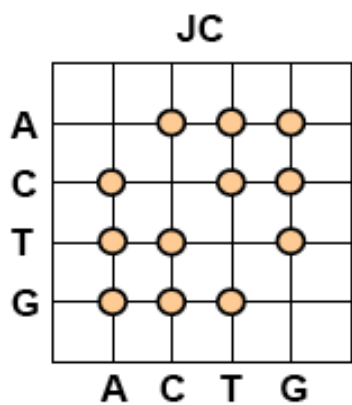
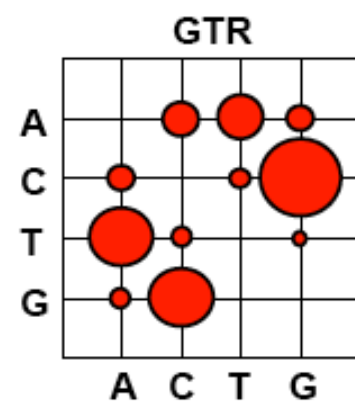
Allow for unequal
base exchangeabilities



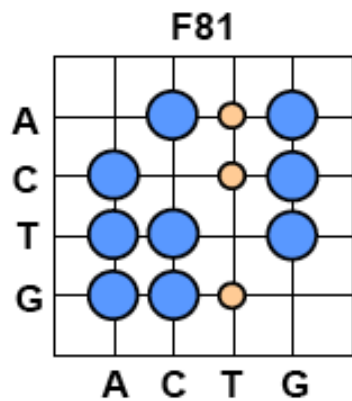
Allow for unequal
base frequencies



Allow for all six base
substitutions to vary



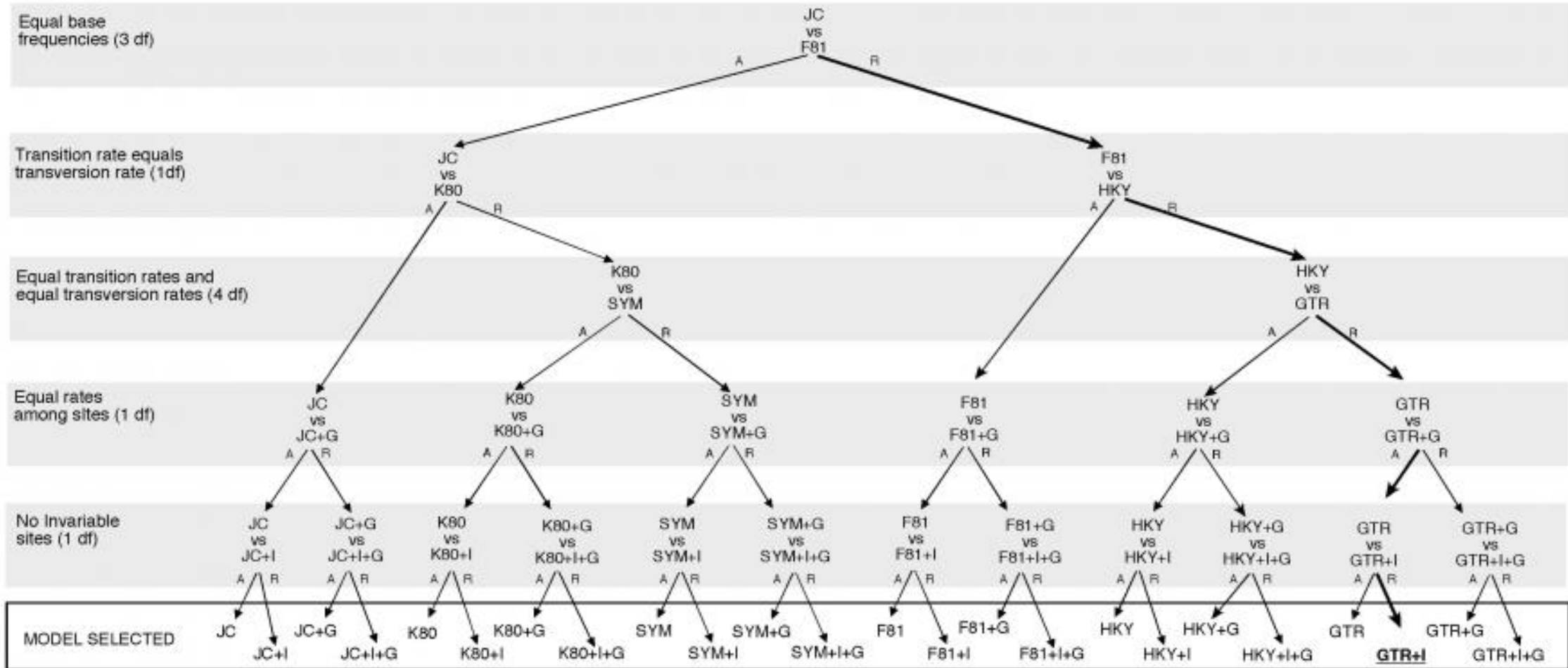
Allow for unequal
base frequencies



Allow for unequal
base exchangeabilities

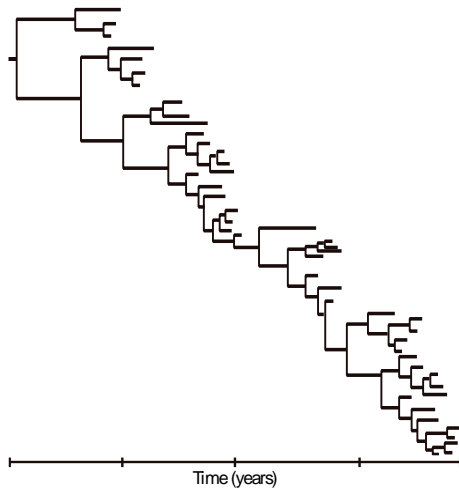


After Whelan et al. 2001

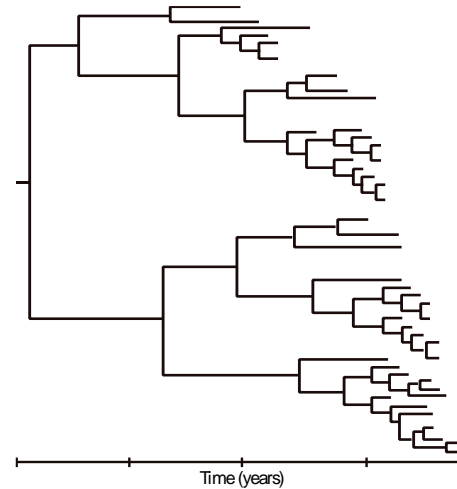


Building Phylogenetic Trees

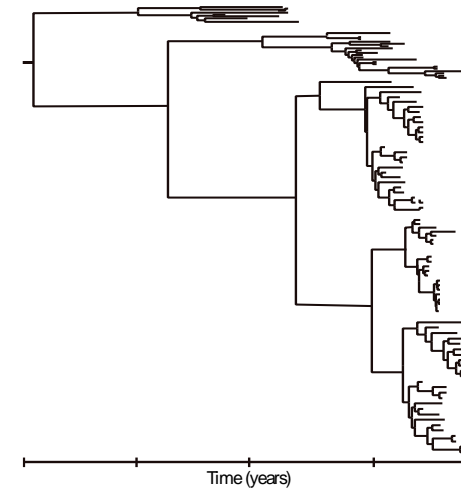
Phylogeny is shaped by host ecology and epidemiology



**Human
Influenza
(Seasonal)**



**Avian
Influenza
(Wild bird)**



**Avian Influenza
(Domestic
poultry)**

Methods to infer evolutionary trees

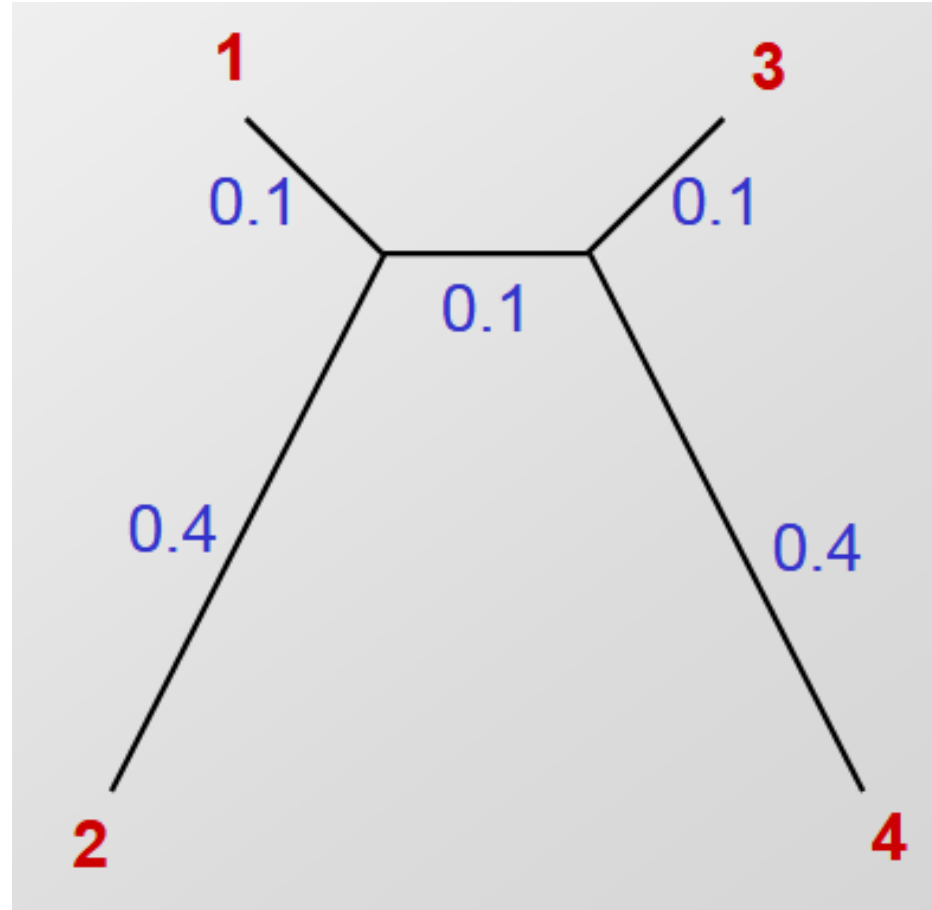
	Character based methods	Noncharacter based methods
Explicit model of evolution	Maximum-likelihood (including Bayesian) methods	Pairwise-distance methods
No explicit model of evolution	Maximum-parsimony Methods	

1. **Nucleotides are characters**
2. **Substitution model describes the process of character change**

Distance methods

- These methods use genetic distance to build a tree without any additional funny business
- UPGMA and Neighbor-joining methods are the most common
- Super-fast and not computationally demanding
- Treats all genetic changes equally, making it a poorly-fitting model for many problems
- Only one possible tree for any dataset

Neighbor joining



Maximum likelihood-based methods

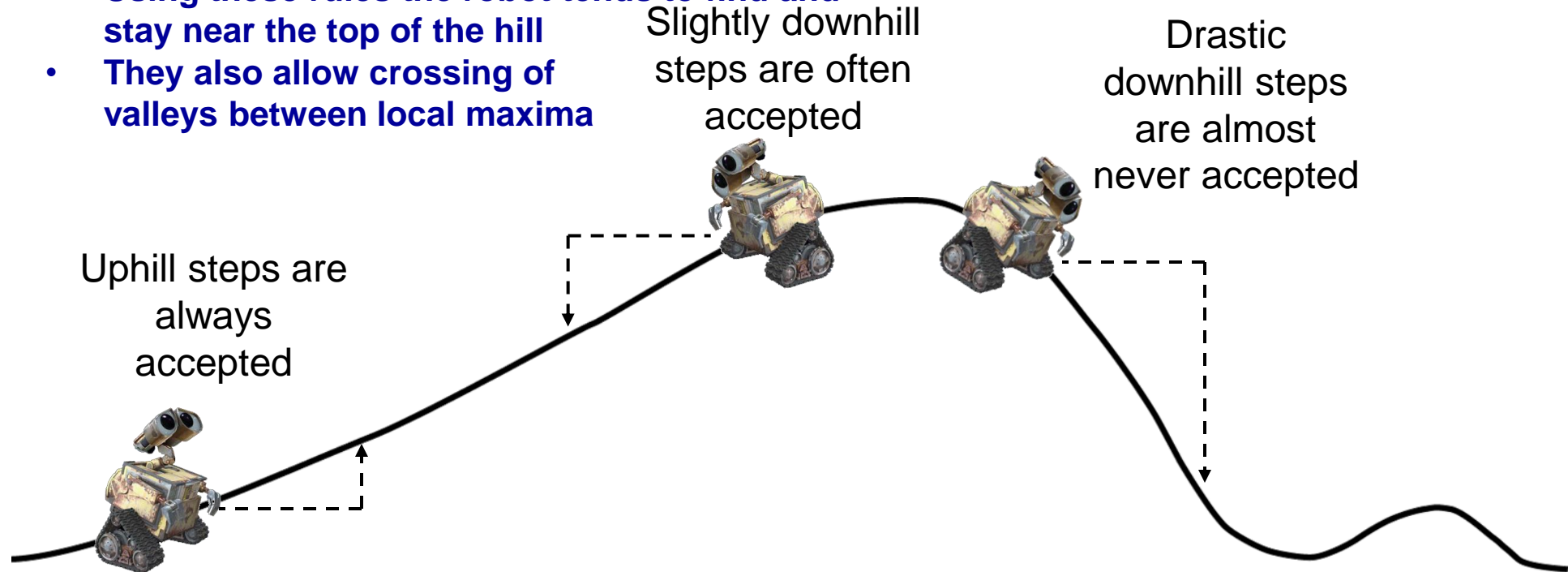
- Considers each nucleotide one at a time, incorporating the evolutionary models we learned about in the previous section.
- Allows us to identify events like convergent evolution which may be covered up by the “saturation” problem we described in the previous section.
- Generates many different tree topologies and identifies the tree with the highest likelihood given the data and model.
- Offers a nice balance in speed.
- IQ-Tree and FastTree are two of my favorites!

Bayesian Methods

- Based around the idea of Bayesian statistics, that is, we estimate the probability given “prior information”
- Bayesian methods allow us to construct complex models to describe the evolution of our pathogen
- This tends to be the slowest method, but the most flexible and precise
- Done using RevBayes or BEAST

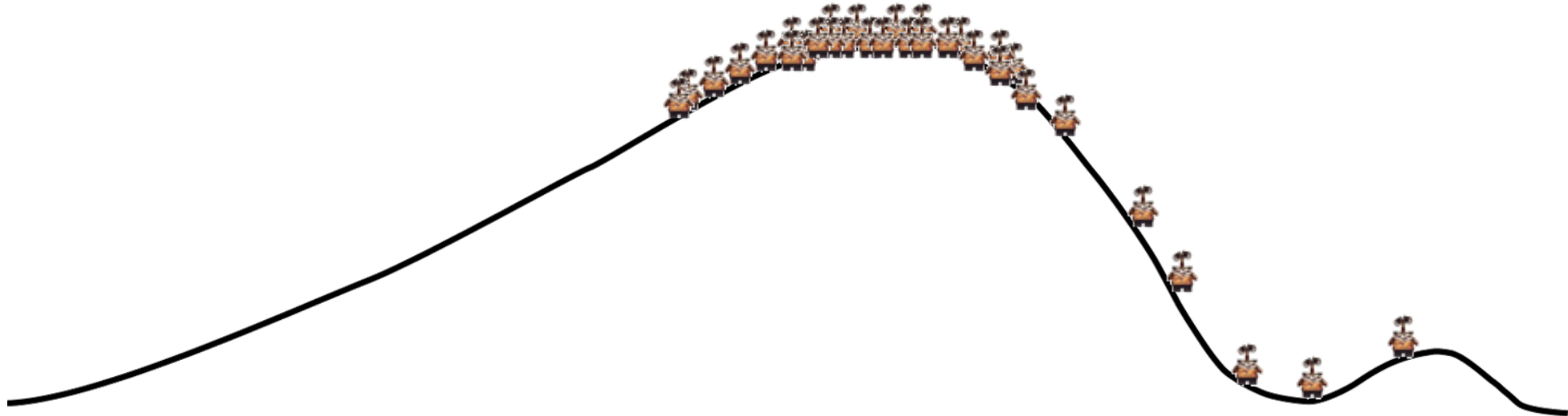
MCMC (Markov Chain Monte Carlo)

- MCMC searches allow both uphill and downhill moves
- It has a few simple rules
- Using these rules the robot tends to find and stay near the top of the hill
- They also allow crossing of valleys between local maxima

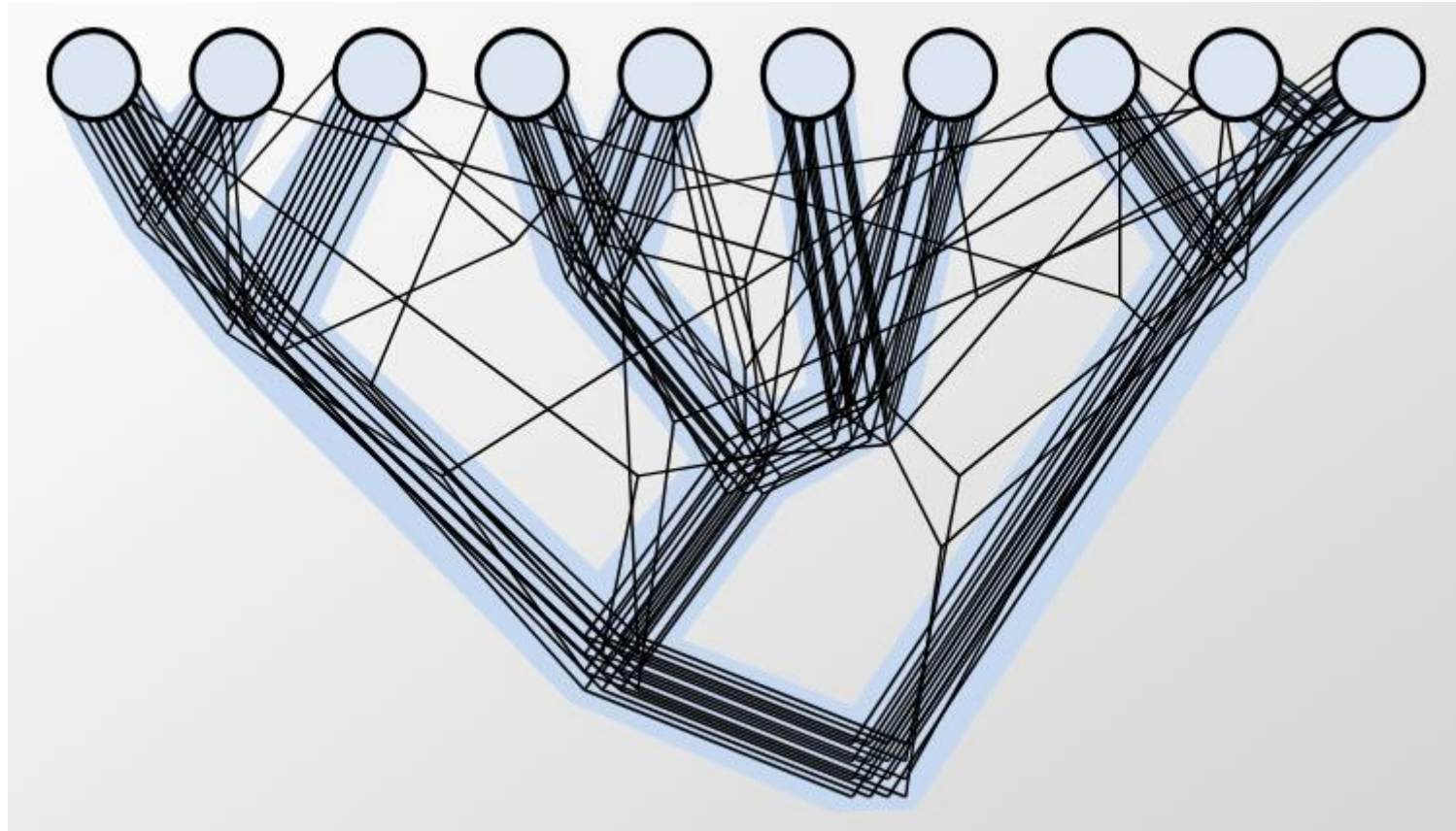


MCMC

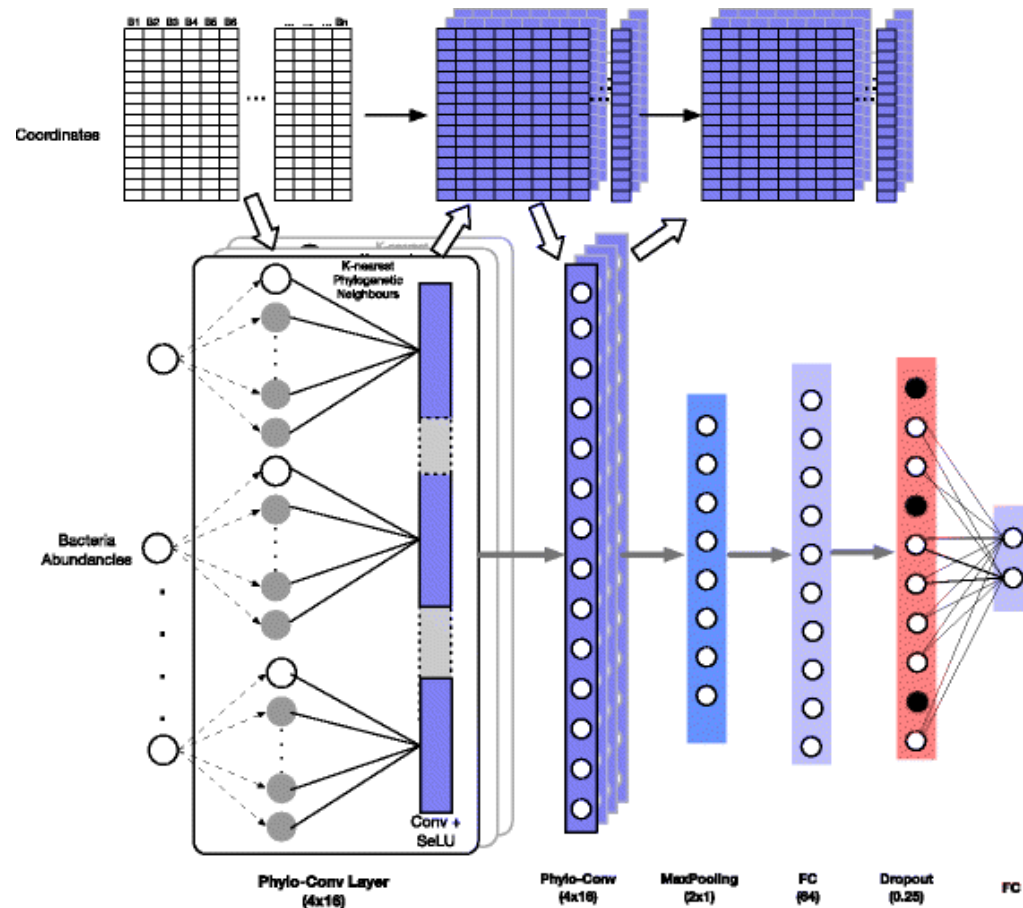
- An MCMC has no end point (it does not search for the 'best' tree like ML)
- Instead it explores tree space
- The rules mean it spends most of its time exploring trees that fit the data well
- Because it has no ultimate goal we must tell it when to stop



We end up with a collection of trees which we can summarize!



Deep learning and other machine learning methods



- Much newer methods
- Leverage neural networks or large language models to infer evolutionary relationships
- Lack a formal statistical model
- Provide excellent predictive performance at the cost of a “black box”
- Slow to train, but fast to run!

Bootstrapping for tree reliability

- While Bayesian methods provide a 95% highest posterior density, for Maximum Likelihood methods, we use a bootstrap to determine if a given node is reliable
- A subset of the alignment is used to generate multiple trees, and by counting the proportion of trees containing each clade, we have an estimate for the statistical confidence for each branch
- Each bootstrap replicate takes as much time to compute as the original tree, and we typically do this 100-1000 times to get our confidence measures



Nextstrain

Real-time tracking of pathogen evolution

About us

An open-source project to harness the scientific and public health potential of pathogen genome data

Core pathogens

Continually updated views of a range of pathogens maintained by the Nextstrain team

SARS-CoV-2

Up-to-date analyses and a range of resources for SARS-CoV-2, the virus responsible for COVID-19 disease

Open source tooling

Bioinformatic workflows, analysis tools and visualization apps for use by the community

Nextclade

In-browser phylogenetic placement, clade assignment, mutation calling and sequence quality checks

Nextstrain Groups

Datasets and narratives shared by research labs, public health entities and others

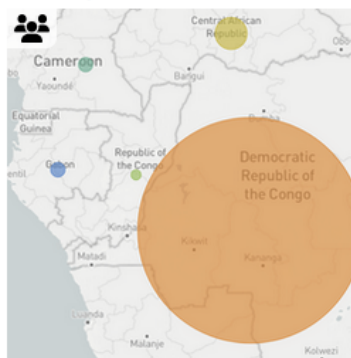
Featured analyses

SARS-CoV-2



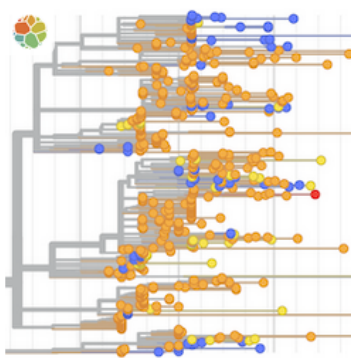
Ongoing evolution and spread of SARS-CoV-2

Mpox in the DRC



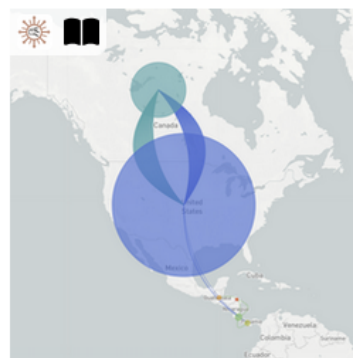
INRB analysis of ongoing mpox clade I outbreak in the DRC

H5N1 cattle outbreak



Influenza H5N1 cattle outbreak in the USA (genotype B3.13)

HPAI outbreak



Highly pathogenic avian influenza in North America

Philosophy

Pathogen Phylogenies

In the course of an infection and over an epidemic, pathogens naturally accumulate random mutations to their genomes. This is an inevitable consequence of error-prone genome replication. Since different genomes typically pick up different mutations, mutations can be used as a marker of transmission in which closely related genomes indicate closely related infections. By reconstructing a *phylogeny* we can learn about important epidemiological phenomena such as spatial spread, introduction timings and epidemic growth rate.

Actionable Inferences

However, if pathogen genome sequences are going to inform public health interventions, then analyses have to be rapidly conducted and results widely disseminated. Current scientific publishing practices hinder the rapid dissemination of epidemiologically relevant results. We thought an open online system that implements robust bioinformatic pipelines to synthesize data from across research groups has the best capacity to make epidemiologically actionable inferences.

This Website

This website aims to provide a *real-time* snapshot of evolving pathogen populations and to provide interactive data visualizations to virologists, epidemiologists, public health officials and citizen scientists. Through interactive data visualizations, we aim to allow exploration of continually up-to-date datasets, providing a novel surveillance tool to the scientific and public health communities.

Future Directions

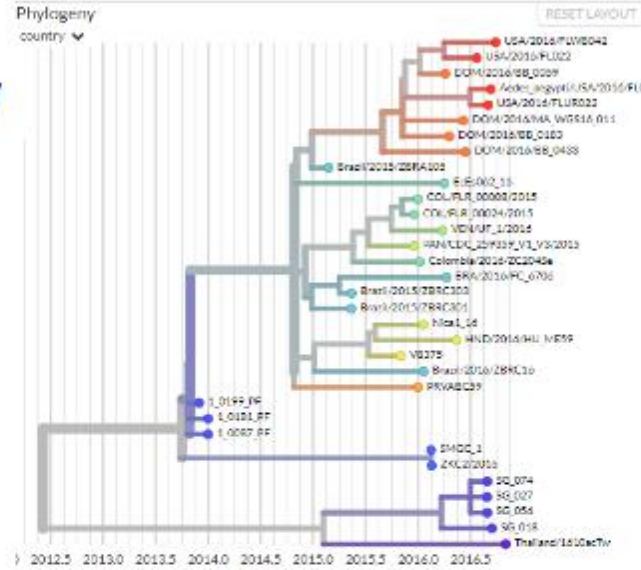
Nextstrain is under active development and we have big plans for its future, including visualization, bioinformatics analysis and an increasing number and variety of datasets. If you have any questions or ideas, please [contact us](#).

A bioinformatics and data viz toolkit

Nextstrain provides an open-source toolkit enabling the bioinformatics and visualization you see on this site. Tweak our analyses and create your own using the same tools we do. We aim to empower the wider genomic epidemiology and public health communities.

Nextstrain: Default view

Phylogeny



Map



Genome

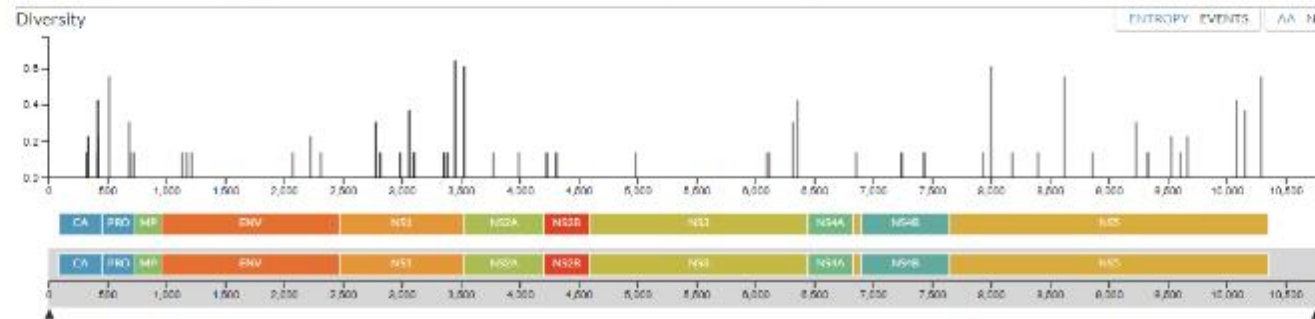


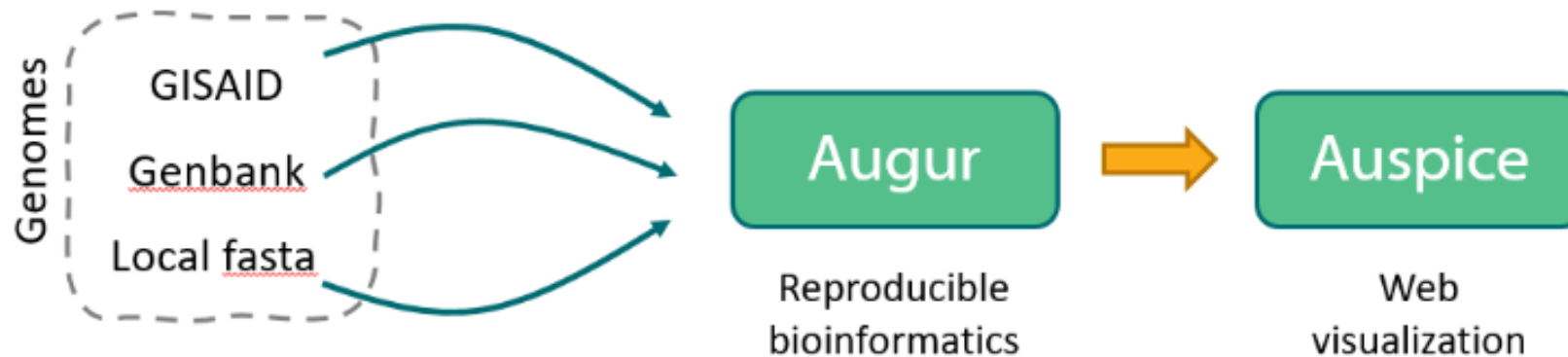
Image from Trevor Bedford Group: nextstrain.org

Slides adapted from CDC AMD Genomic Epi Toolkit:
https://www.cdc.gov/advanced-molecular-detection/media/pdfs/ToolkitModule_3.1-508C.pdf

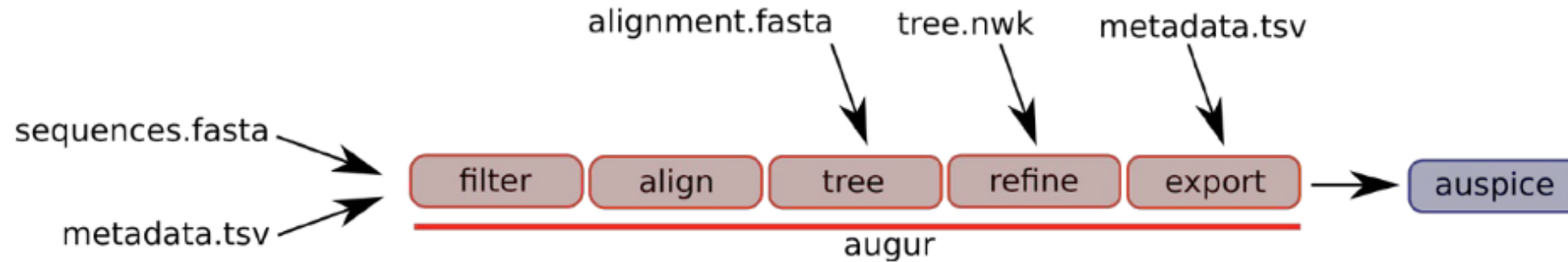
Nextstrain architecture

Two goals, two components

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



Augur: What does it do?



- Input data

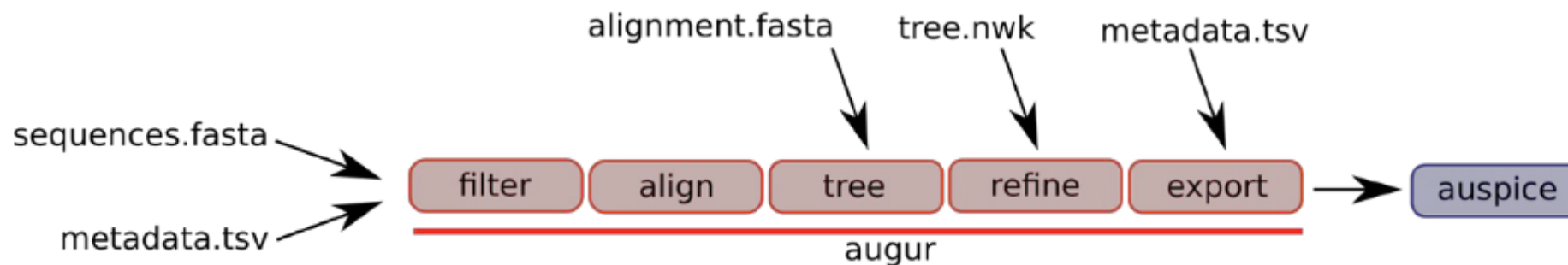
- `sequences.fasta`
- `metadata.tsv`
- *Can also import a tree if already constructed (like a Bayesian tree)*

- Visualization data for Auspice

- Colors, `lat_longs`, reference genome

Adapted from Nathan Grubaugh nextstrain.org

Augur: What does it do?



■ What augur does

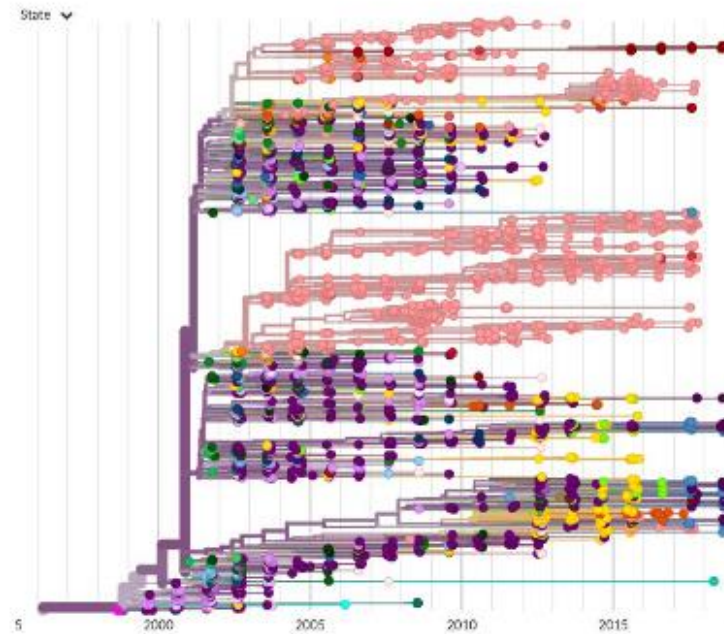
- Prepare pathogen sequences and metadata
- Align sequences
- Construct a phylogeny from aligned sequences
- Annotate the phylogeny with inferred ancestral pathogen dates, sequences, and traits
- Export the annotated phylogeny and corresponding metadata into auspice-readable text file (JSON)

Adapted from Nathan Grubaugh nextstrain.org

Auspice: What does it do?

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

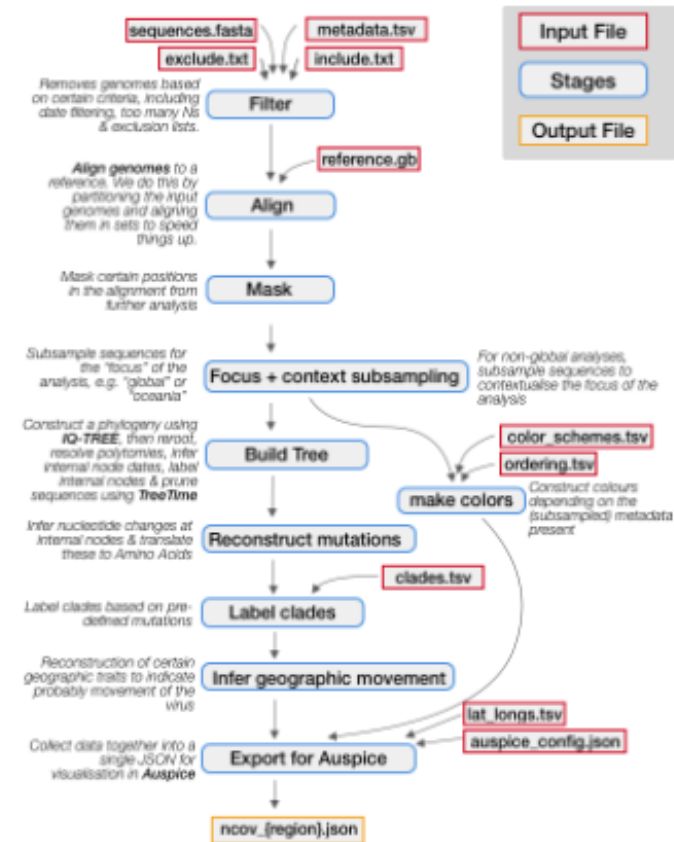
```
{
  "aa_muts": {},
  "attr": {
    "clock_length": 0.001,
    "date": "1996-03-10",
    "div": 0,
    "lineage": "pre-NY",
    "lineage_confidence": {
      "NY99": 0.47670023438528425,
      "pre-NY": 0.5231279601042924
    },
    "lineage_entropy": 0.6937274360549726,
    "num_date": 1996.1897774953388,
    "num_date_confidence": [
      1995.4406449972648,
      1997.0887781968595
    ],
    "state": "Israel",
    "state_confidence": {
      "CT": 0.006183160743180048,
      "Israel": 0.5217402934773154,
      "NY": 0.4024864049221443,
      "TX": 0.0043324152837630955
    },
    "state_entropy": 1.1783556353611715
  },
  "branch_length": 0.001,
  "children": [
    {
      "aa_muts": {},
      "attr": {
        "authors": "Malkinson et al",
```



Adapted from Nathan Grubaugh nextstrain.org

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



<https://docs.nextstrain.org/en/latest/tutorials/SARS-CoV-2/steps/orientation-workflow.html>

Slides adapted from CDC AMD Genomic Epi Toolkit:
https://www.cdc.gov/advanced-molecular-detection/media/pdfs/ToolkitModule_3.1-508C.pdf

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:

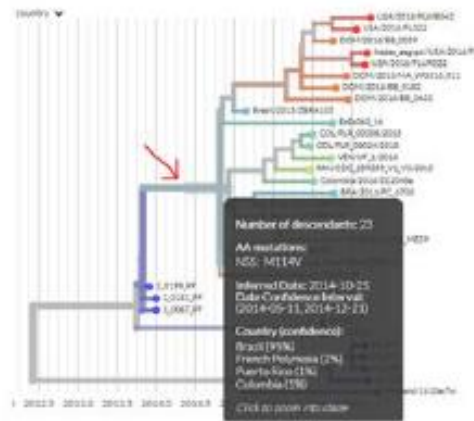
1. Options for visualizing and sharing results
2. Interpreting your results
3. Writing a narrative to highlight key findings

Slides adapted from CDC AMD Genomic Epi Toolkit:
https://www.cdc.gov/advanced-molecular-detection/media/pdfs/ToolkitModule_3.1-508C.pdf

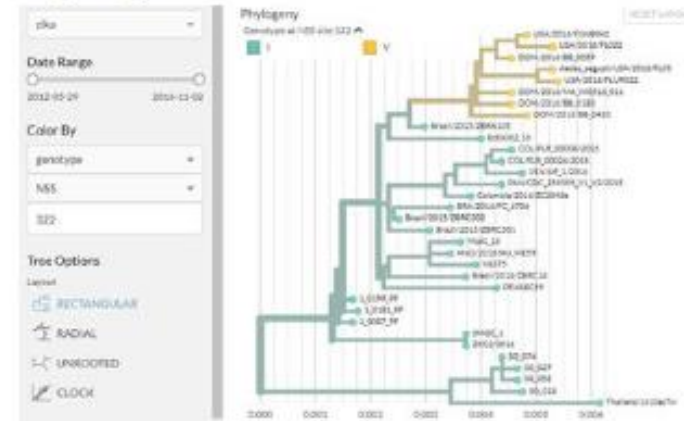
Nextstrain documentation

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

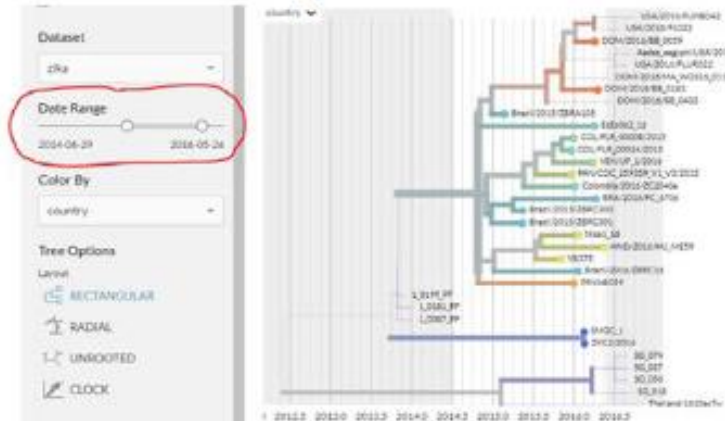
Node details



Highlight variants



Select date ranges



Slides adapted from CDC AMD Genomic Epi Toolkit:
https://www.cdc.gov/advanced-molecular-detection/media/pdfs/ToolkitModule_3.1-508C.pdf

Images from nextstrain.org

Genomic epidemiology of SARS-CoV-2 with subsampling focused globally over the past 6 months



Built with [nextstrain/ncov](#). Maintained by the [Nextstrain team](#). Data updated 2025-04-20. Enabled by data from [GISAID](#).

Showing 3958 of 3958 genomes sampled between Dec 2019 and Apr 2025.

Phylogeny

Location ^

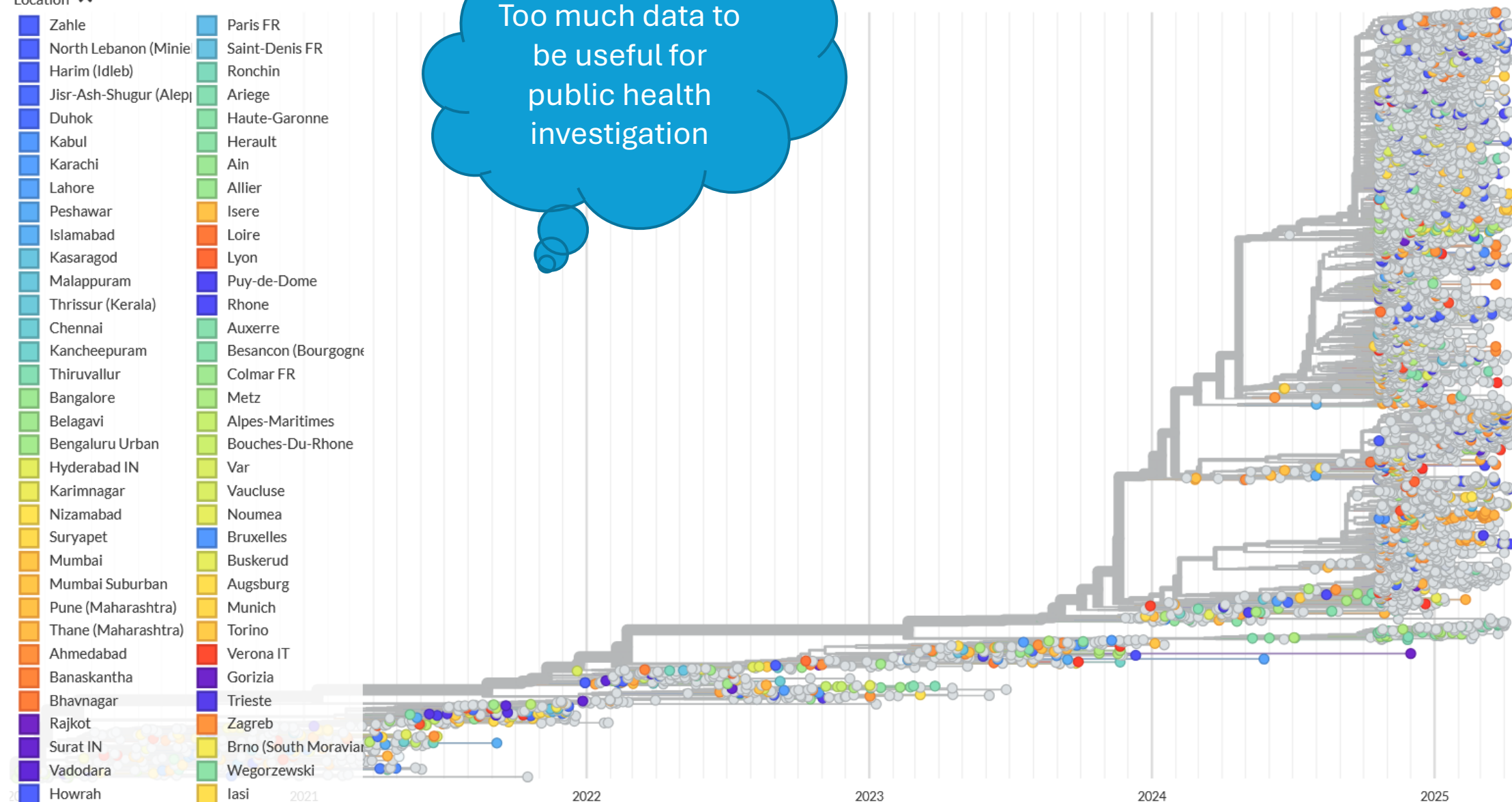
Zahle	Paris FR
North Lebanon (Minie)	Saint-Denis FR
Harim (Idleb)	Ronchin
Jisr-Ash-Shugur (Alepx)	Ariege
Duhok	Haute-Garonne
Kabul	Herault
Karachi	Ain
Lahore	Allier
Peshawar	Iserre
Islamabad	Loire
Kasaragod	Lyon
Malappuram	Puy-de-Dome
Thrissur (Kerala)	Rhone
Chennai	Auxerre
Kancheepuram	Besancon (Bourgogne)
Thiruvallur	Colmar FR
Bangalore	Metz
Belagavi	Alpes-Maritimes
Bengaluru Urban	Bouches-Du-Rhone
Hyderabad IN	Var
Karimnagar	Vaucluse
Nizamabad	Noumea
Suryapet	Bruxelles
Mumbai	Buskerud
Mumbai Suburban	Augsburg
Pune (Maharashtra)	Munich
Thane (Maharashtra)	Torino
Ahmedabad	Verona IT
Banaskantha	Gorizia
Bhavnagar	Trieste
Rajkot	Zagreb
Surat IN	Brno (South Moravia)
Vadodara	Wegorzewski
Howrah	Iasi

Too much data to
be useful for
public health
investigation



ZOOM TO SELECTED

RESET LAYOUT



Caveats

- If there are no samples (genomes) from a certain location in the dataset being visualized, then there is no data to display on our map!

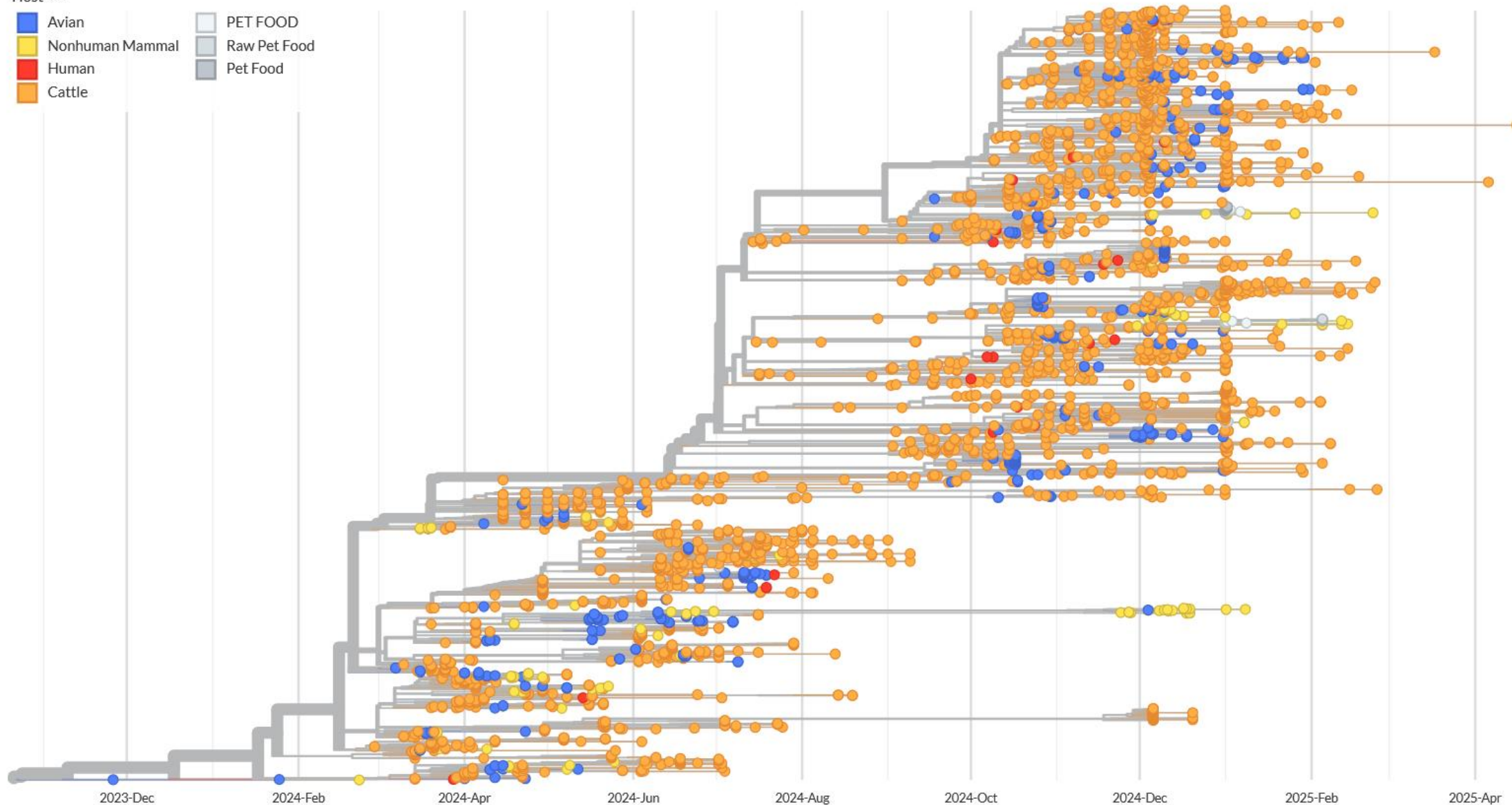
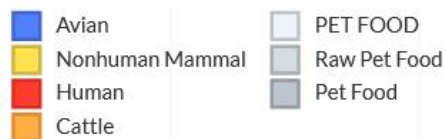
Full genome analysis of the ongoing influenza A/H5N1 cattle outbreak in North America

Built with [nextstrain/avian-flu](#). Maintained by [Louise Moncla](#) and [the Nextstrain team](#). Data updated 2025-04-16. Enabled by data from [USDA](#), [Andersen Lab](#) and [GenBank](#).

Showing 4043 of 4043 genomes sampled between Nov 2023 and Apr 2025.

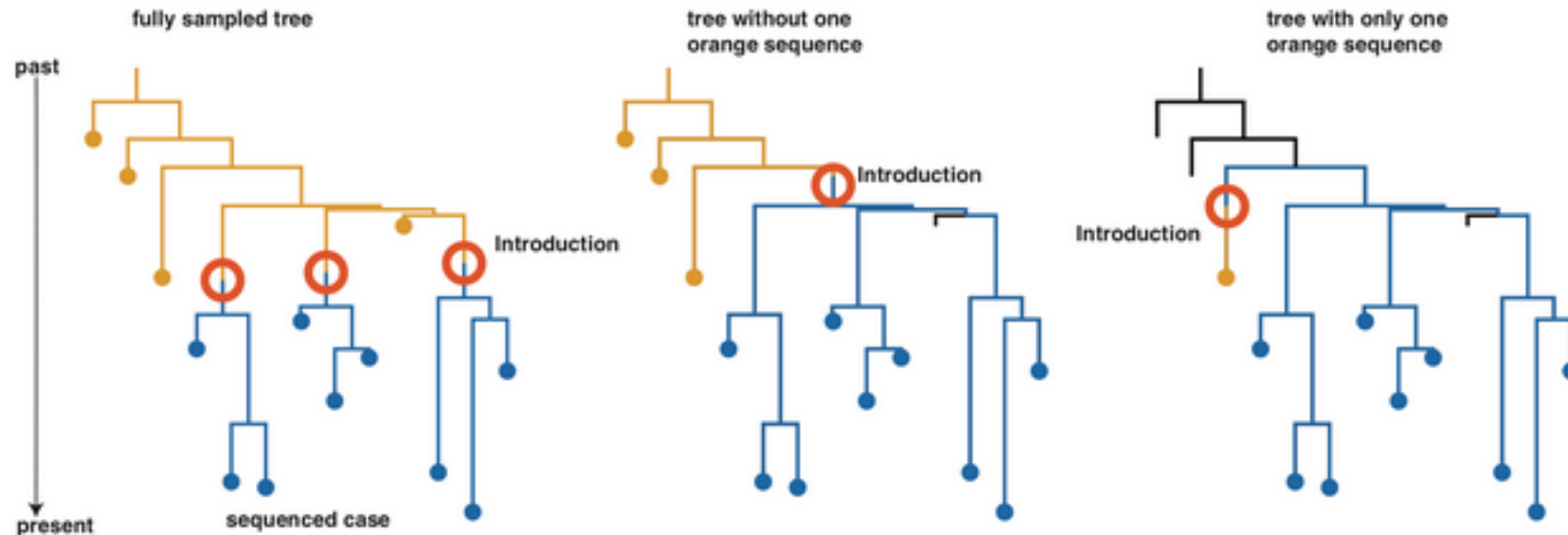
Phylogeny

Host ^



Caveats

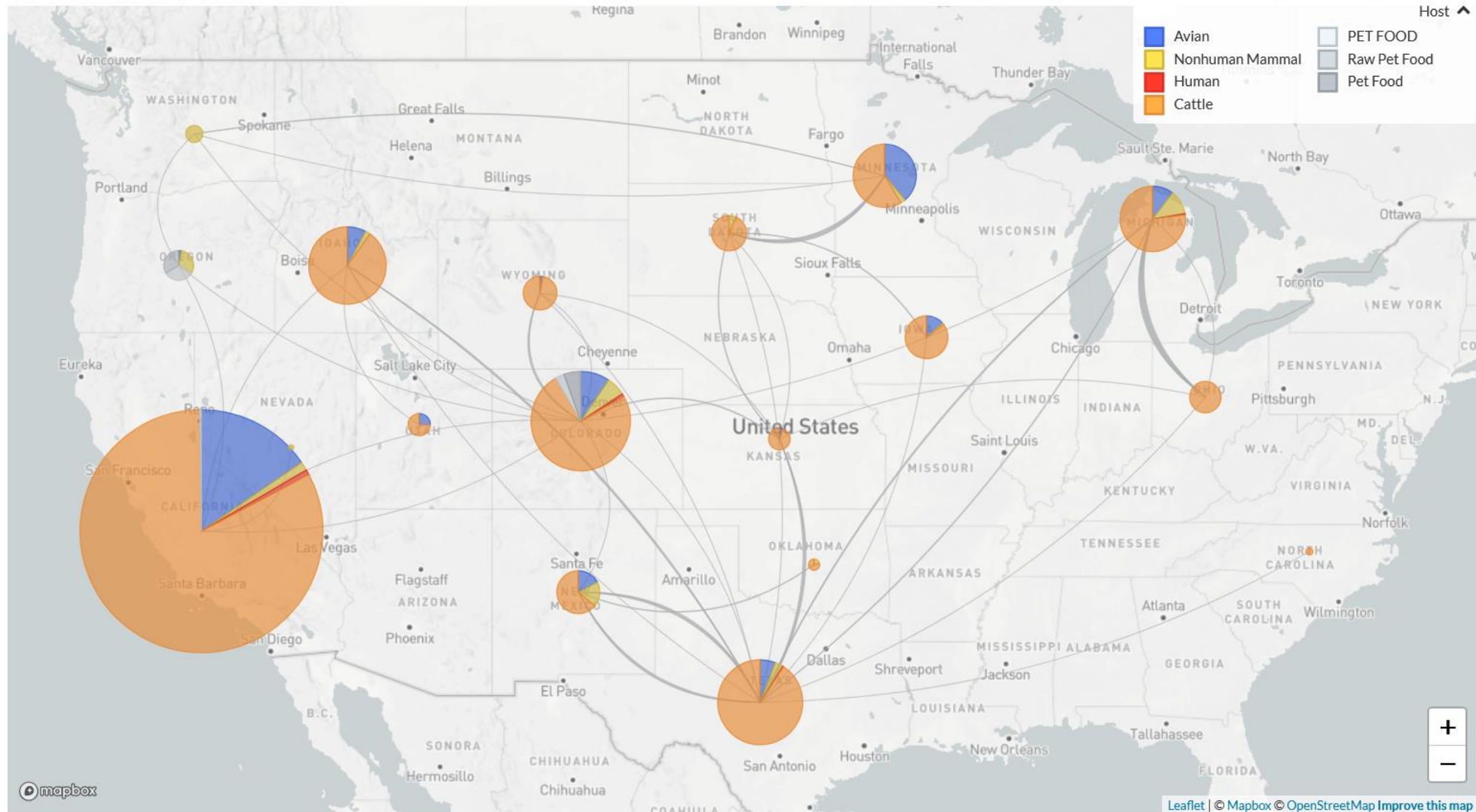
- If we are missing cases, we may underestimate the number of introductions
- Overall, the inferred locations of where a lineage has been in the past should be considered as highly uncertain



Transmissions

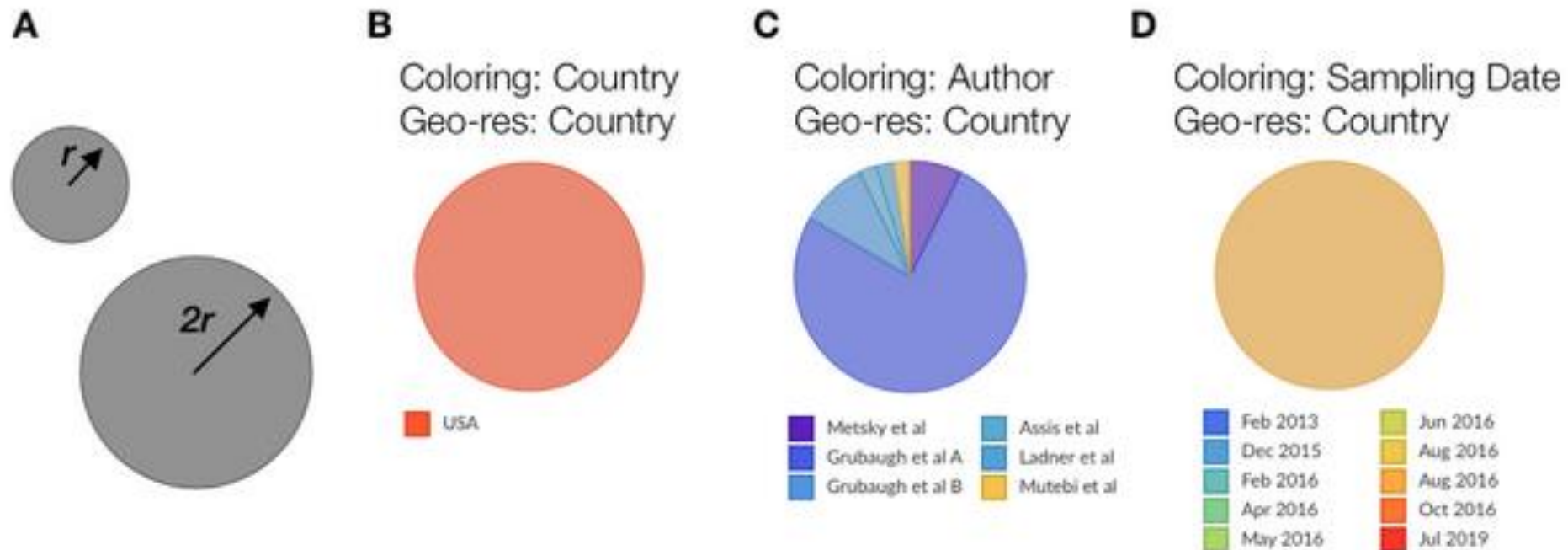
RESET ZOOM

Host ^



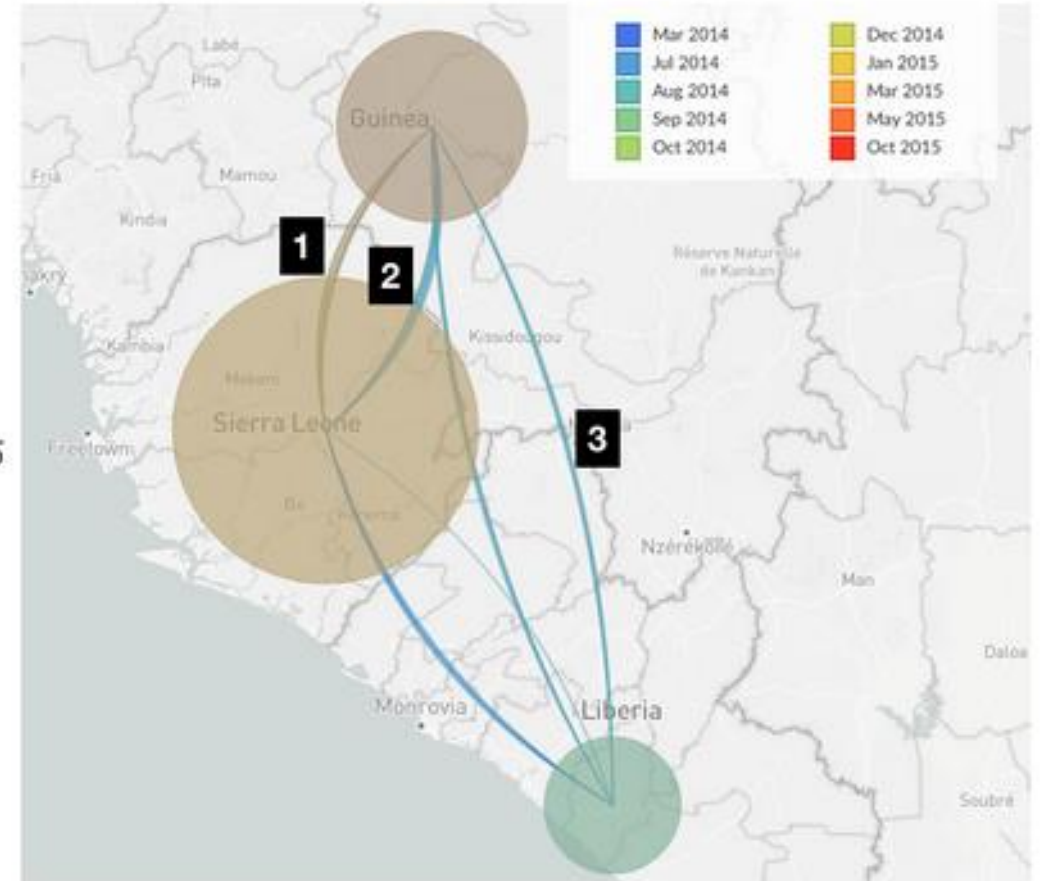
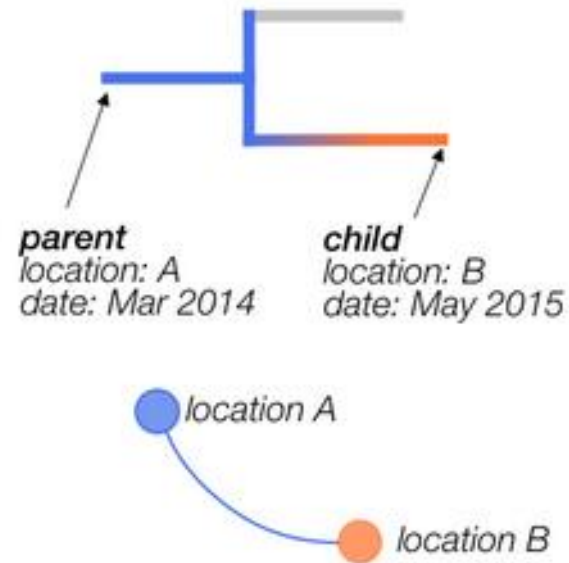
How do we read the transmission map?

- The size of the circles corresponds to the number of samples
- The color of the circles corresponds to the metadata (location, host, etc.)



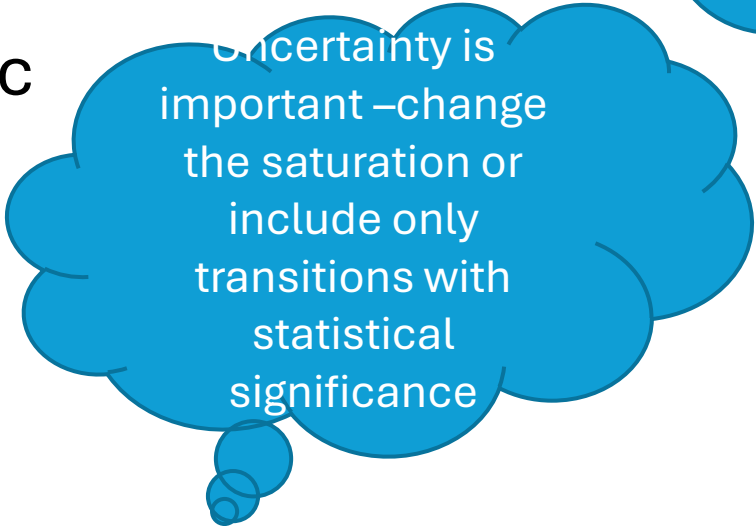
How do we read the transmission map?

- Transmissions (lines) represent our best estimate of the movement of the pathogen over the selected time slice
- Each individual line represents a parent-child branch in the phylogenetic tree

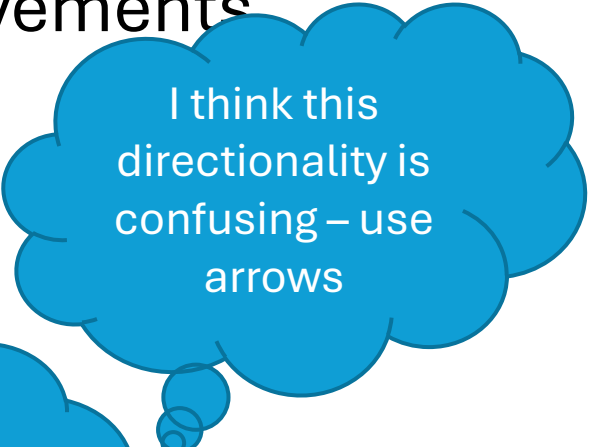


How do we read the transmission map?

- Each arc is actually a collection of individual lines of identical thickness. Lines that look thicker represent more movements between locations in the tree
- Line colored by metadata
- Uncertainty is not included
- Directionality: orientation of the arc
- Animation (“play” button)

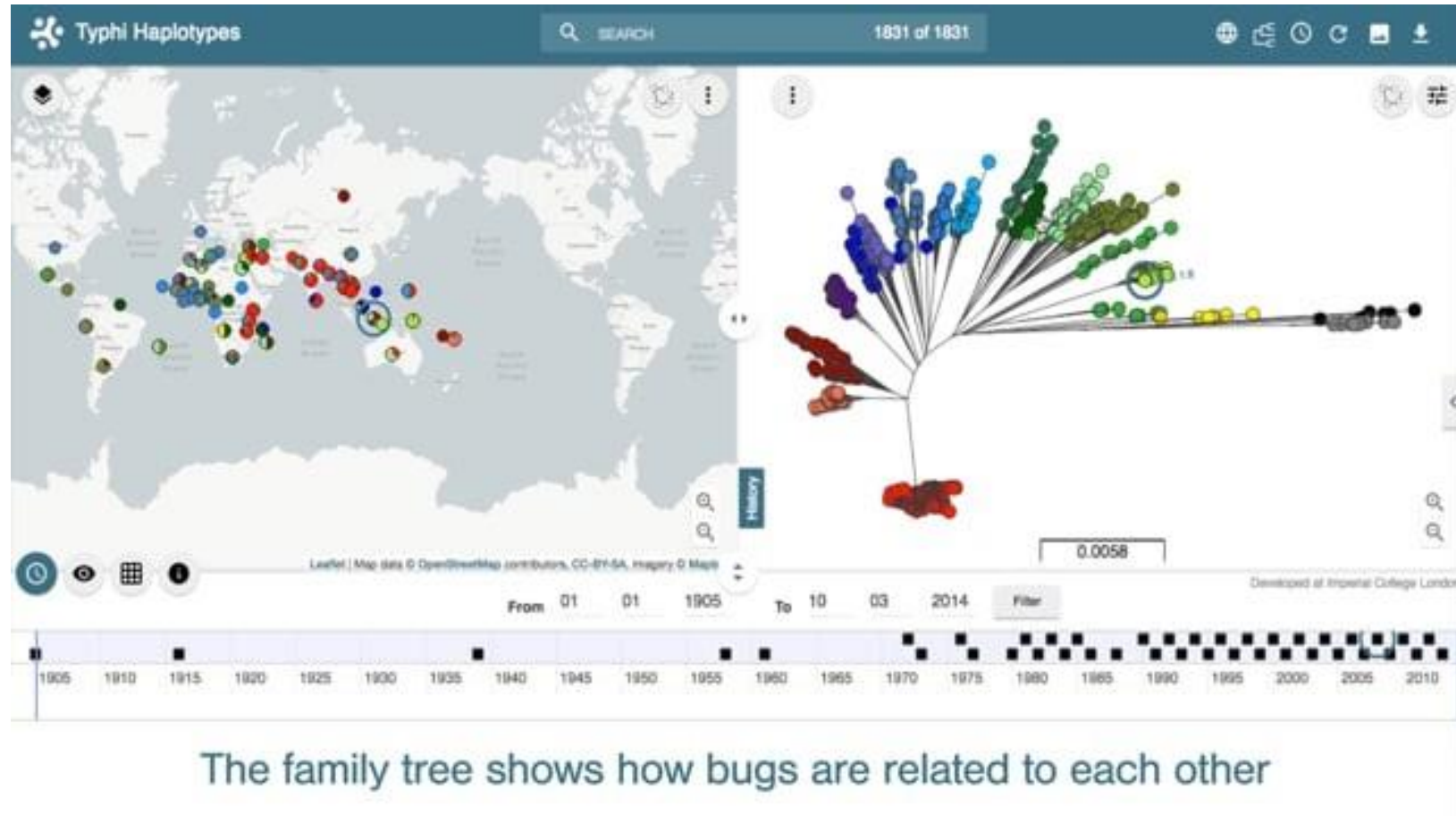


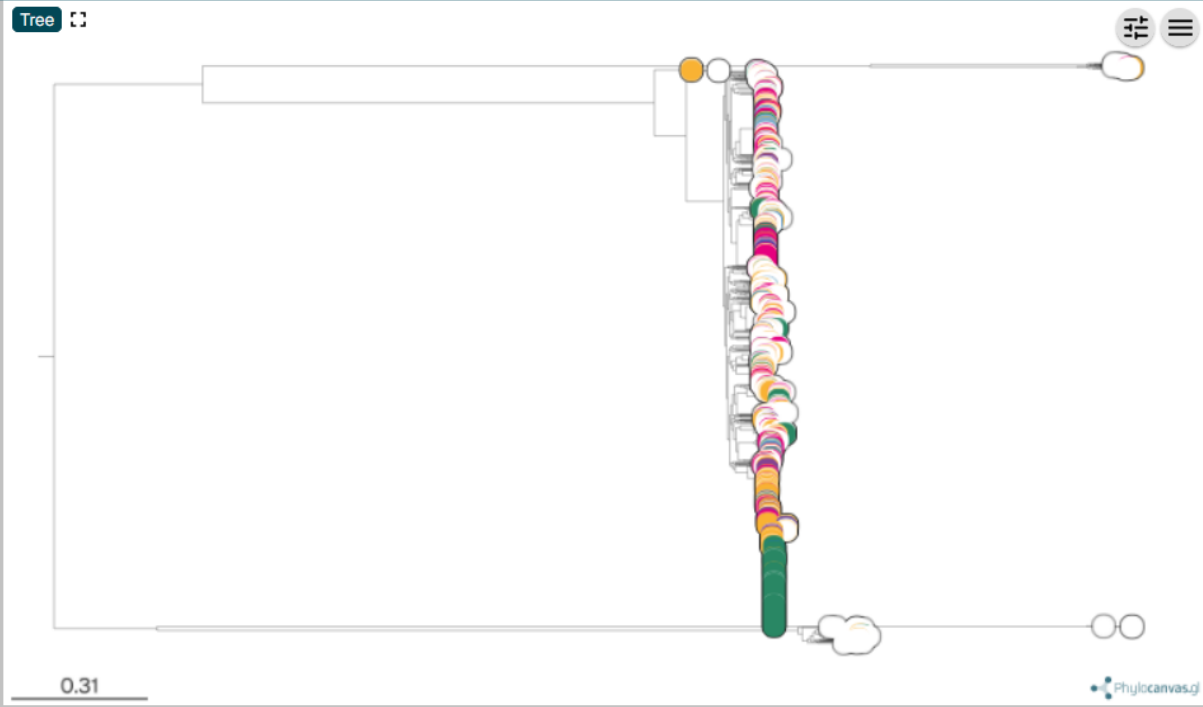
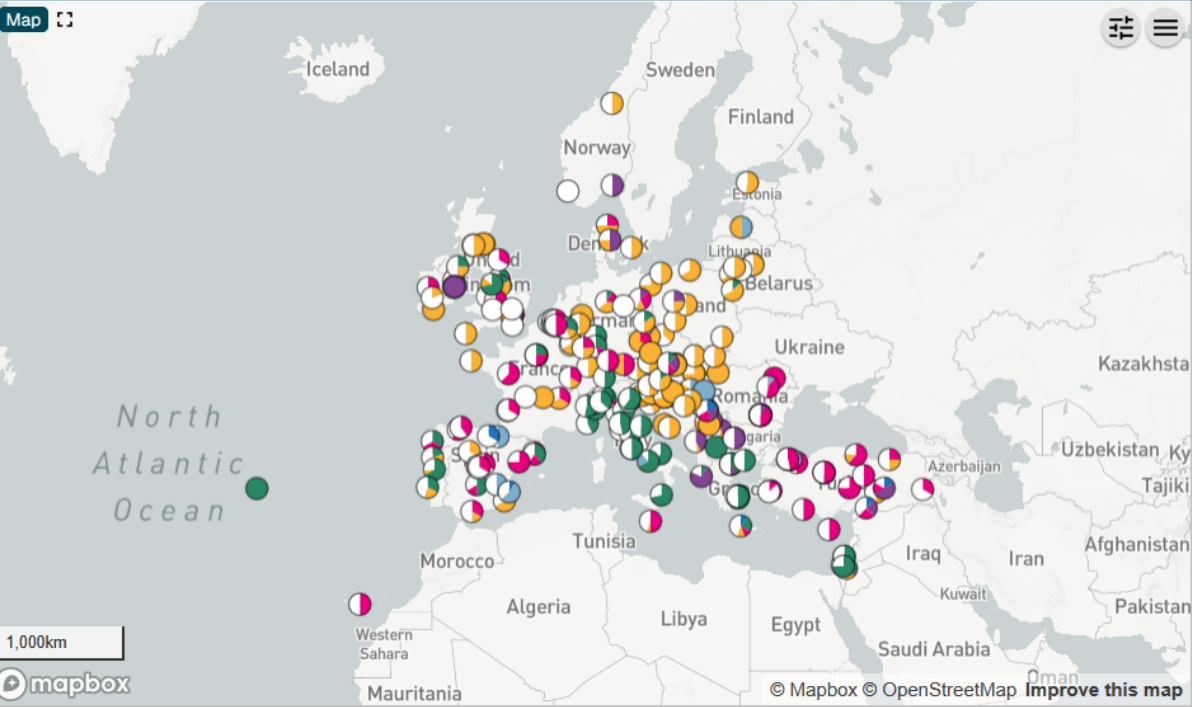
Uncertainty is important – change the saturation or include only transitions with statistical significance



I think this directionality is confusing – use arrows

Microreact





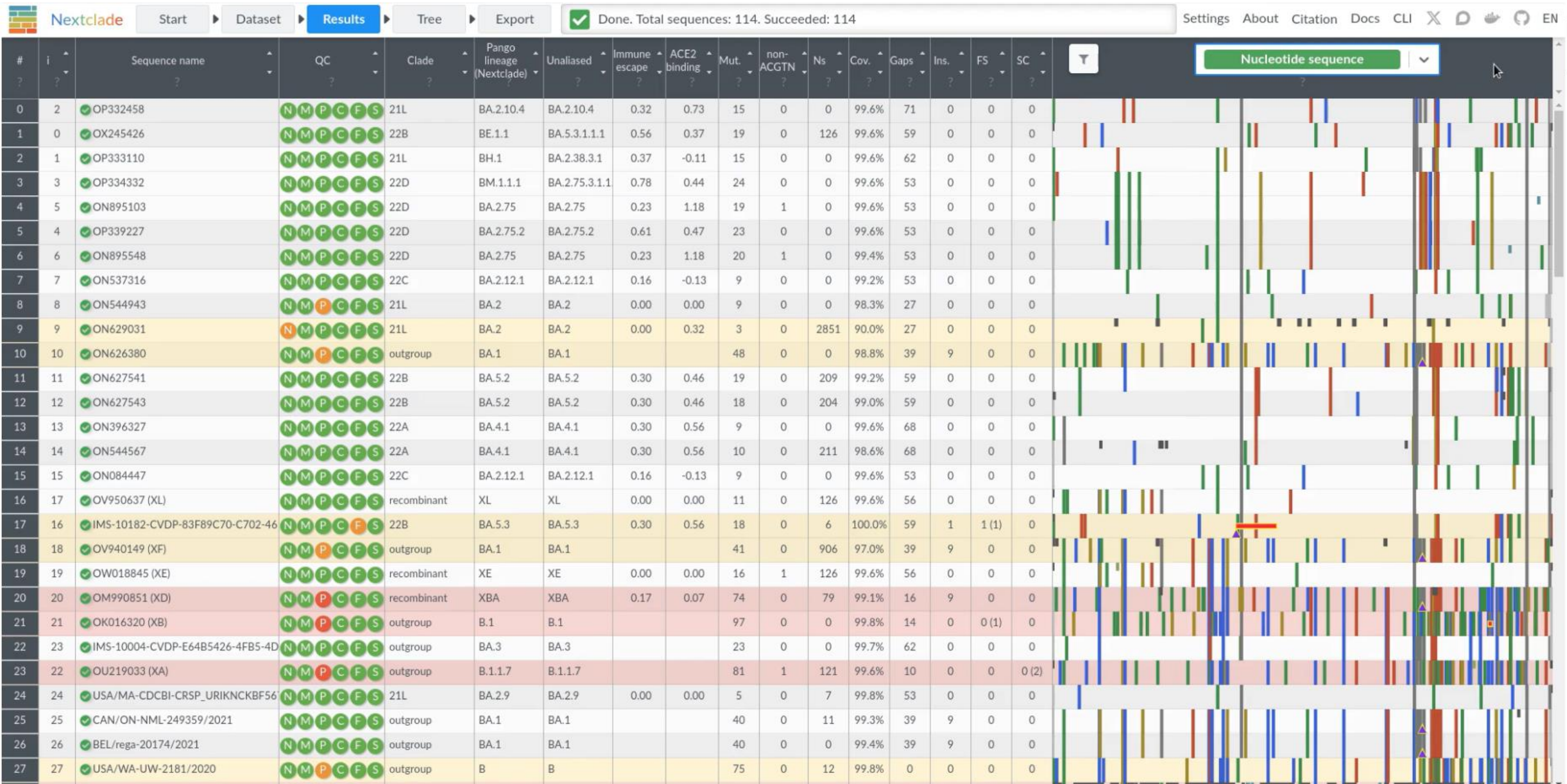
Metadata

	Carbapenem	id	EuSCAPE ID	Sample name	Run accession	Assembly accession	Isolate type	Meropenem MIC	Country	Sample type	Clinical impact	Epidemiology	Species
<input type="checkbox"/>	Other	17870_2#2	EuSCAPE_A...	202556	ERR1204815	GCA_9005...	Suspected n...	0.12	Austria	Urine	Infection	Hospital Ac...	K. pr
<input type="checkbox"/>	Other	17870_2#34	EuSCAPE_P...	209/14	ERR1204847	GCA_90051...	Suspected n...	0.12	Poland	Urine	Infection	Community ...	K. pr
<input type="checkbox"/>	Other	17870_2#39	EuSCAPE_P...	5786/13	ERR1204852	GCA_90051...	Suspected n...	0.12	Poland	Urine	Infection	Hospital Ac...	K. pr
<input type="checkbox"/>	Other	17870_2#41	EuSCAPE_P...	622/14	ERR1204854	GCA_90051...	Suspected n...	0.12	Poland	Blood	Infection	Unknown	K. pr

The screenshot shows the Nextclade web application interface. At the top, the browser address bar displays 'clades.nextstrain.org'. The Nextclade logo is in the top left, and navigation links like 'What's new' and a language dropdown are in the top right. The main heading 'Nextclade beta v0.9.0' is centered, followed by the tagline 'Clade assignment, mutation calling, and sequence quality checks'. Below this, two primary modes are offered: 'Simple' (no installation, drop file) and 'Private' (no remote processing). A row of four analysis features is listed: Mutation Calling, Clade Assignment, Phylogenetic Placement, and Quality Control. On the right, a large panel for sequence upload is shown. It has a dropdown for 'SARS-CoV-2' and a toggle for 'Simple mode' (which is active). Below this is a 'Sequences' section with a 'required' label and three input methods: 'From file', 'From URL', and 'Paste'. A dashed box contains a 'Drag & Drop a file here' instruction, a 'FASTA' file icon, and a 'Select a file' button. A 'Show me an Example' link is at the bottom of this panel.

<https://clades.nextstrain.org/>

Slides adapted from CDC AMD Genomic Epi Toolkit:
https://www.cdc.gov/advanced-molecular-detection/media/pdfs/ToolkitModule_3.1-508C.pdf



QC metrics

Nextclade implements a variety of quality control metrics to quickly spot problems in your sequencing/assembly pipeline. You can get a quick idea which of your sequences are having problems by sorting the results table from bad to good (click on the upper arrow in the column "QC"). Bad sequences are colored red, mediocre ones yellow and good ones white. You can view detailed results of the QC metrics by hovering your mouse over a sequences QC entry:

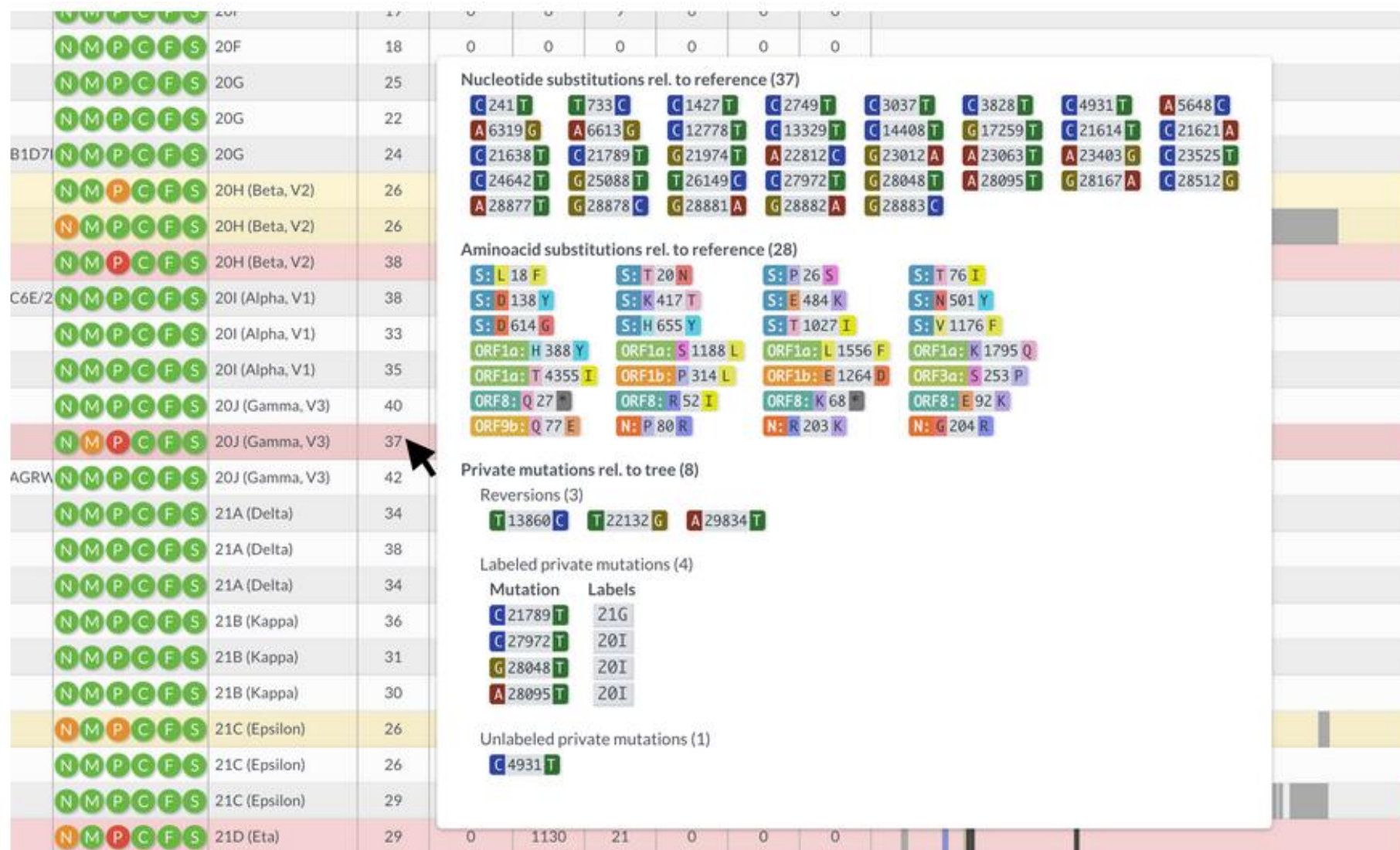
ID	Sequence name	QC	Clade	Mut.	non-ACGTN	Ns	Gaps
23	✓ Switzerland/AG-ETHZ-430474/2020	N M P C F S					
42	▲ USA/CA-CDC-LC0027675/2021	N M P C F S					
29	✓ BHR/30005866/2021	N M P C F S					
35	✓ USA/CA-CDC-LC0024684/2021	N M P C F S					
34	▲ USA/NJ-CDC-LC0019972/2021	N M P C F S					
22	✓ Switzerland/un-ETHZ-420516/2020	N M P C F S					
9	✓ HongKong/VM20006541/2020	N M P C F S					
67	✓ USA/FL-CDC-ASC210057011/2021	N M P C F S					
28	▲ USA/VA-DCLS-3814/2021	N M P C F S					
84	✓ Switzerland/100064/2020	N M P C F S					
30	▲ USA/TX-CDC-STM-000011246/2021	N M P C F S					
50	✓ USA/GA-CDC-ASC210019884/2021	N M P C F S					
73	✓ USA/IL-CDC-LC0051439/2021	N M P C F S					
57	✓ USA/VA-DCLS-5091/2021	N M P C F S					
21	✓ USA/JGGF/2020	N M P C F S					

Overall QC score: 158
Overall QC status: bad
Detailed QC assessment:

- N Missing Data:** mediocre
Missing data found. Total Ns: 2942 (3000 allowed). QC score: 98
- M Mixed Sites:** good
No issues
- P Private Mutations:** good
No issues
- C Mutation Clusters:** good
No issues
- F Frame shifts:** good
No issues
- S Stop codons:** mediocre
1 misplaced stop codon(s) detected. Affected gene(s): ORF8. QC score: 75

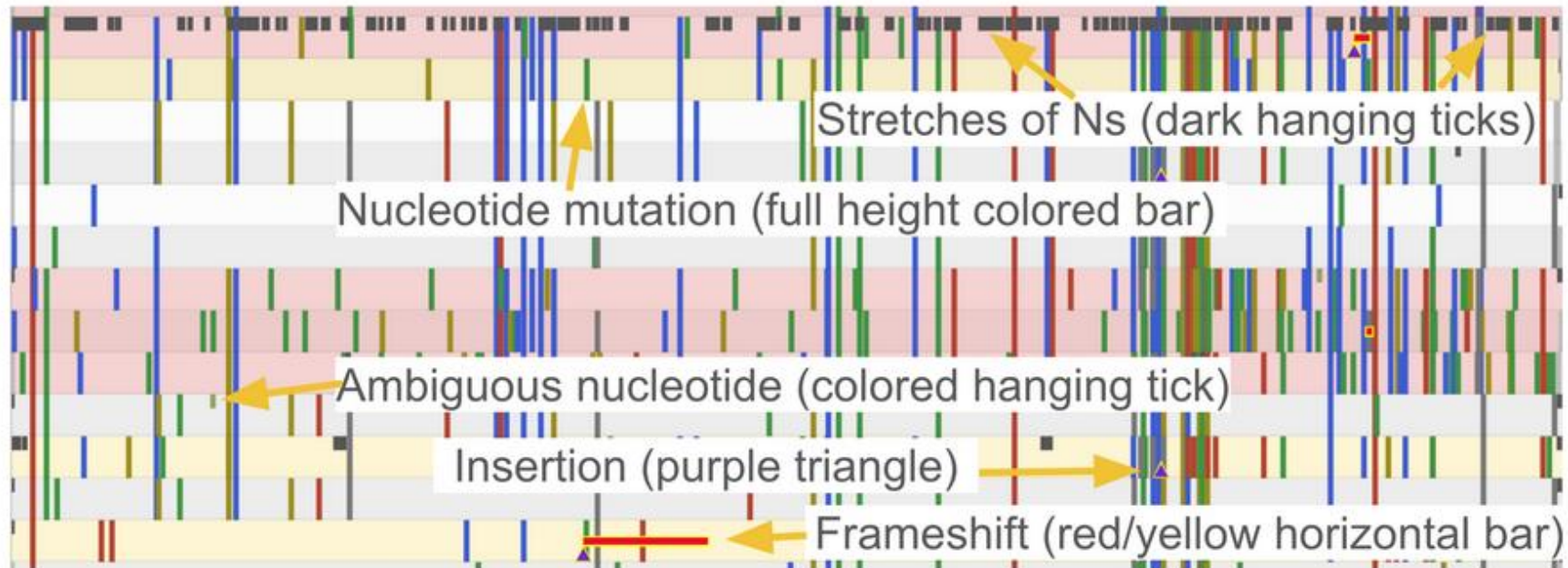
<https://docs.nextstrain.org/projects/nextclade/en/stable/index.html>

Every icon corresponds to a different metric. See [Quality control](#) section for the detailed explanation of QC metrics.

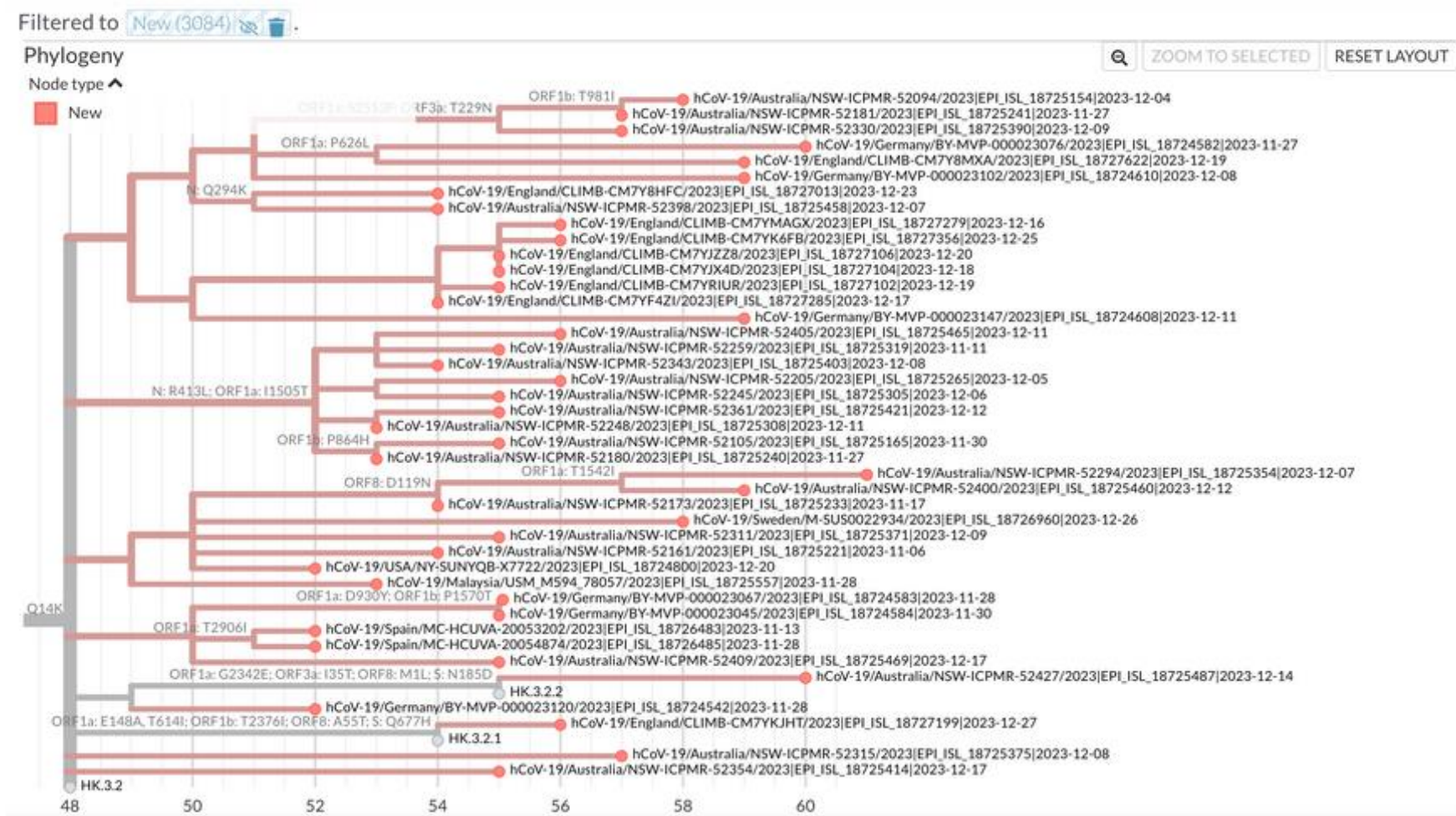


Alignment viewer

To the right of the table you can see the alignment with mutations and regions with missing data highlighted in grey. You can quickly check how segments of missing data are distributed on the genome - whether it's a few big chunks clustering in one area or many small missing segments.



If you hover over a feature, you can see its name and coordinates.

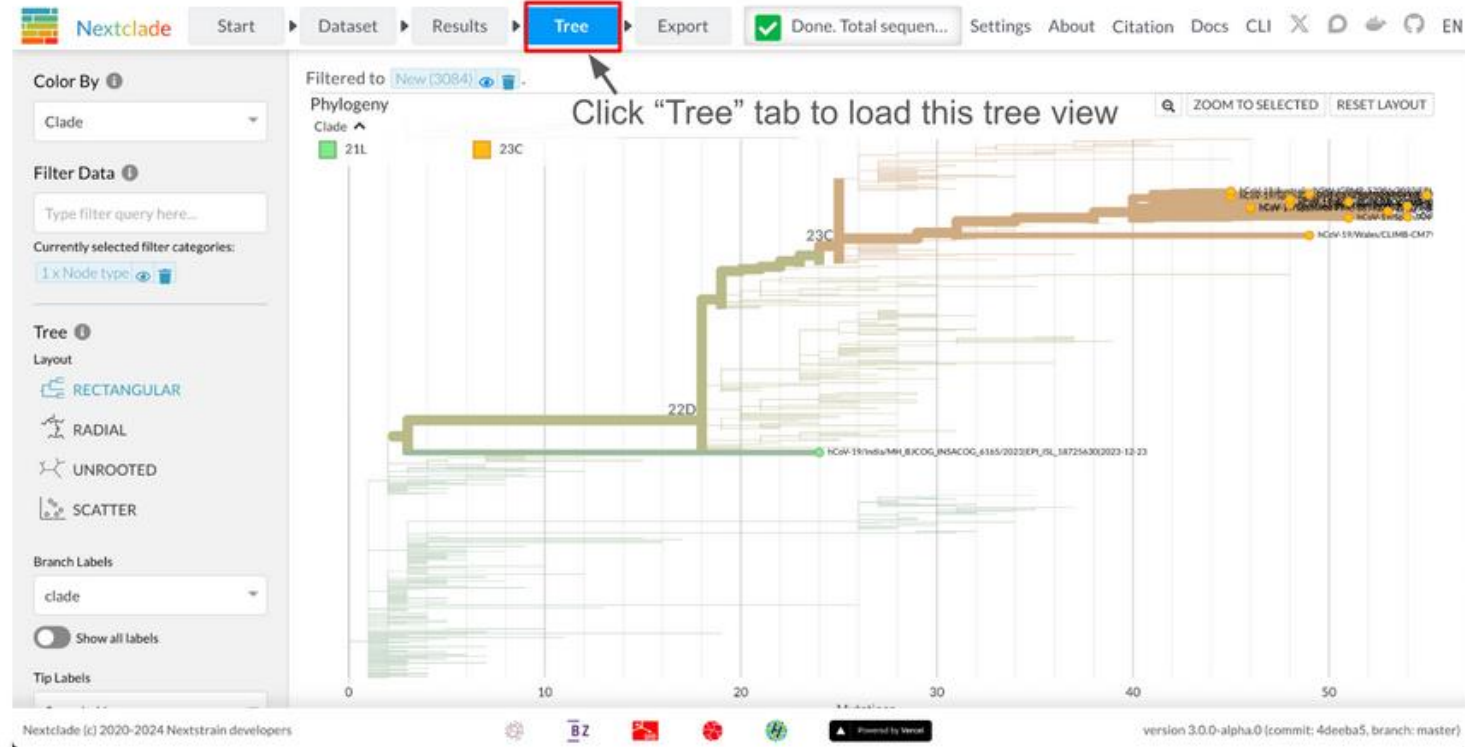


For a more accurate tree including your sequences, you can use [Usher](#), which works out of the box with SARS-CoV-2, hMPXV, RSV-A, and RSV-B (as of January 2024).

<https://docs.nextstrain.org/projects/nextclade/en/stable/index.htm>

In order to assign clades to sequences, Nextclade **places** all new sequences on a reference tree. You can view the resulting tree by clicking on the tree tab at the top left.

The tree is visualized by **Nextstrain Auspice**. By default, only your uploaded sequences are highlighted.



Nextclade runs a greedy parsimony tree builder on user provided sequences. This means that approximate ancestral relationships between your sequences are visible on the tree. Given the simplicity of the tree builder, the tree is not guaranteed to be optimal. In the screenshot below, all but the 3 grey sequences are user provided. Nextclade has grouped related user provided sequences into clusters, based on shared mutations.

UShER: Real-time phylogenetic placement



- Ultrafast Sample placement on Existing tRees
- Designed to take user sequences and
 1. Accurately place them onto global phylogeny
 2. Construct new subtrees
 3. Enable easy visualization
- Runs quickly (<1 second) to facilitate genomic epidemiology

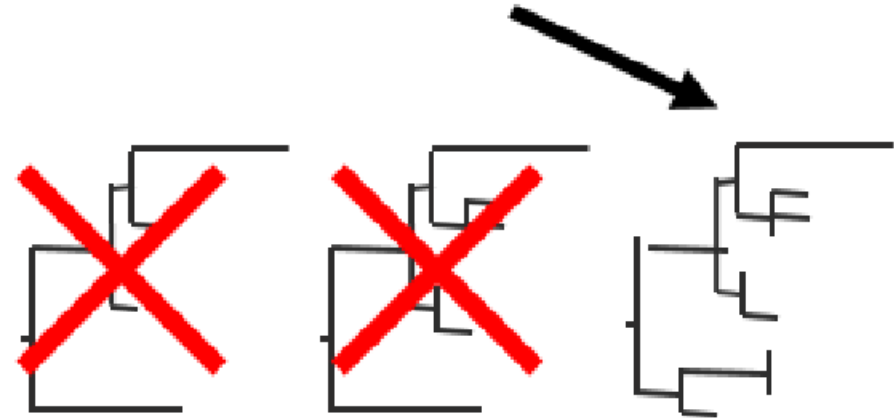
Constant flow and huge datasets overwhelm typical phylogenetics approaches

- Typical phylogenetic workflow:

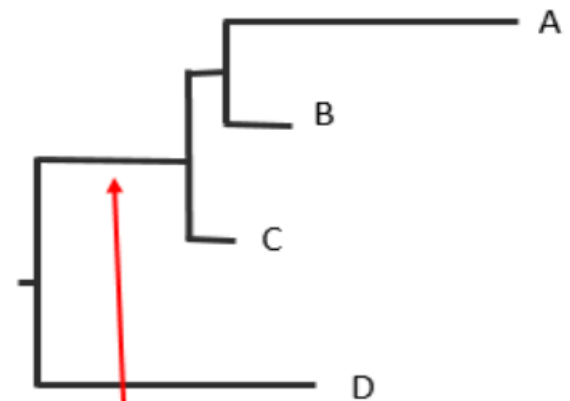
1. Gather data
2. Calculate tree
3. More data
4. Recalculate tree
5. **More data!**
6. **Recalculate tree?**

Repeat... forever

A	G	C	T	T	A	C	T	A	A	T	C	C	G	G	C	C	G	A	A	T	T	A	G	G	T	C	
A	G	T	T	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	C	T	A	G	G	T	C
A	G	T	C	T	A	T	T	A	A	T	T	C	G	A	G	C	A	G	A	A	C	T	T	G	G	T	C
A	G	T	T	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	T	C
A	G	A	T	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	C	T	T	G	G	T	C
A	G	A	T	T	G	C	T	A	A	T	T	C	G	A	G	C	C	G	A	A	T	T	A	G	G	T	C
A	G	A	T	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	T	T	A	G	G	T	C
A	G	T	C	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	T	T	A	G	G	A	C
A	G	C	T	T	A	T	T	A	A	T	T	C	G	T	G	C	T	G	A	A	C	T	C	G	G	A	C
A	G	C	T	T	A	T	T	A	A	T	T	C	G	A	G	C	T	G	A	A	C	T	C	G	G	A	C

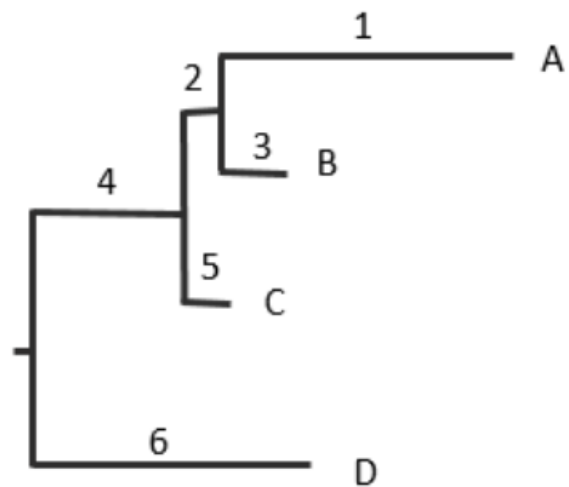


Using parsimony, UShER maps mutations onto the existing tree.



A. 2U, 4U, 6G
B. 2U, 4U, 7U
C. 2U, 9C
D. 10U

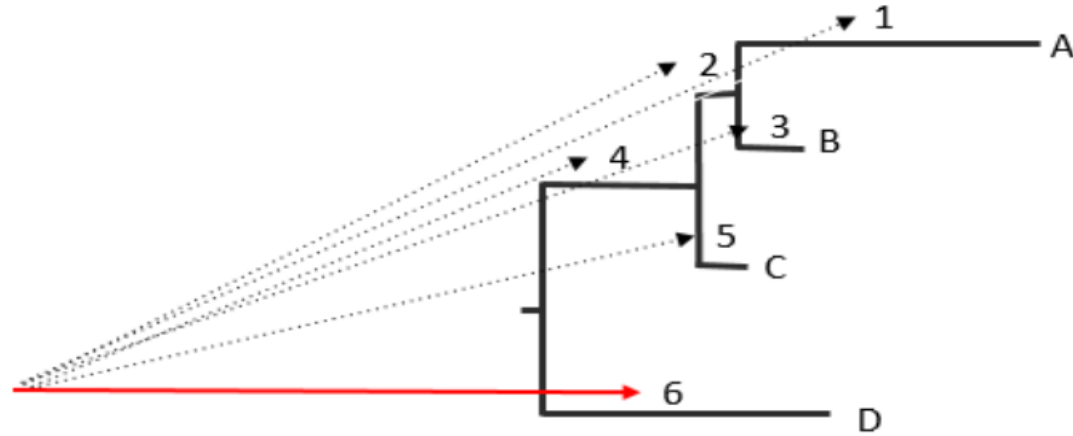
UShER stores this **mutation annotated tree**.



1. 6G
2. 4U
3. 7U
4. 2U
5. 9C
6. 10U

New samples are added using maximum parsimony by checking every possible placement.

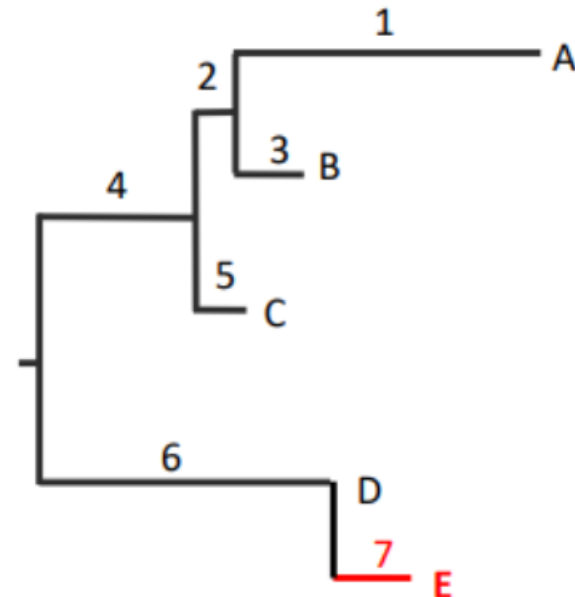
Sample E: 10U, 14G



1. 6G
2. 4U
3. 7U
4. 2U
5. 9C
6. 10U
7. 14G

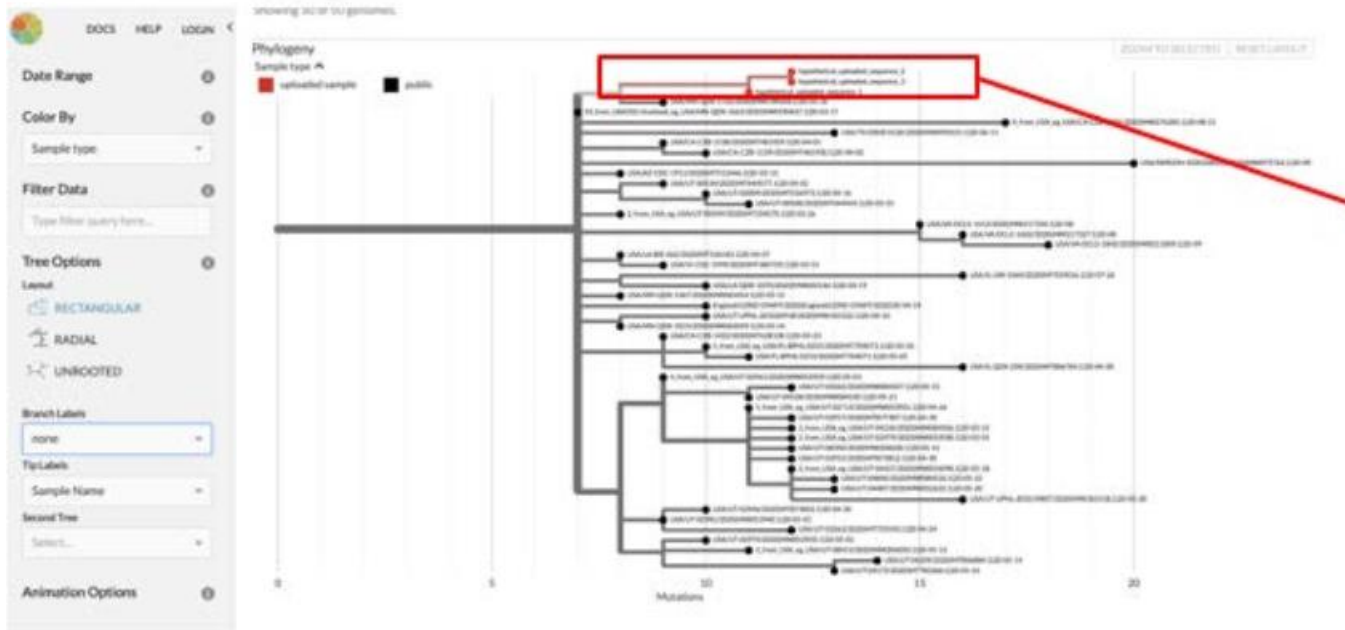
USHER finds the correct placement in 97% of the cases.

When incorrect, placements are still usually very close to the true site.



1. 6G
2. 4U
3. 7U
4. 2U
5. 9C
6. 10U
7. 14G

UShER output



UShER outputs a subtree of 50 most closely related samples to a user's sample.

User's sample in red

This subtree can be visualized and explored using the Nextstrain platform.