Better lives through livestock

Phylogenetics

John Juma
Animal and Human Health

International Livestock Research Institute
Nairobi, Kenya

ILRI
INTERNATIONAL
LIVESTOCK RESEARCH
INSTITUTE

CGIAR

Phylogenetic inference

Phylogenetics is the study of the evolutionary history and relationships among individuals, groups of organisms (e.g., species) or other biological entities with evolutionary histories (e.g., genes).

Phylogenetic tree reconstruction: basic concepts

1. Alignment (both building the data model and extracting a phylogenetic dataset).

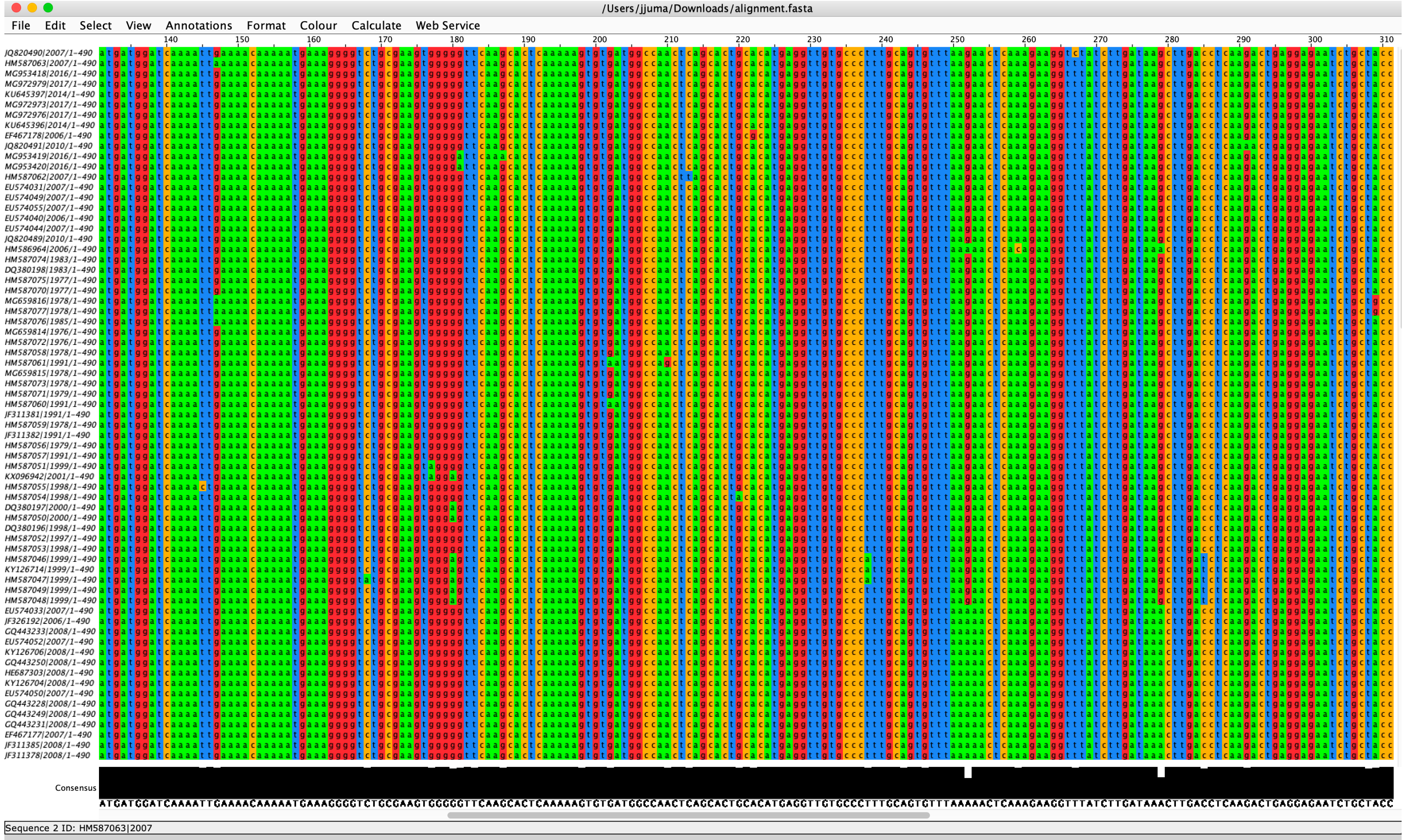
2. Determining the substitution model

3. Tree building

4. Tree evaluation

Alignment: Building the model

Phylogenetic sequence data usually consist of multiple sequence alignments; the individual, aligned-base positions are commonly referred to as "sites".



JQ820490|2007/1-490
HM587063|2007/1-490
MG953418|2016/1-490
MG972979|2017/1-490
KU645397|2014/1-490
MG972973|2017/1-490
MG972976|2017/1-490
KU645396|2014/1-490
EF467178|2006/1-490
JQ820491|2010/1-490
MG953419|2016/1-490
MG953420|2016/1-490
HM587062|2007/1-490
EU574031|2007/1-490
EU574049|2007/1-490
EU574055|2007/1-490
EU574040|2006/1-490
EU574044|2007/1-490
JQ820489|2010/1-490
HM586964|2006/1-490
HM587074|1983/1-490
DQ380198|1983/1-490
HM587075|1977/1-490
HM587070|1977/1-490
MG659816|1978/1-490
HM587077|1978/1-490
HM587076|1985/1-490
MG659814|1976/1-490
HM587072|1976/1-490
HM587058|1978/1-490
HM587061|1991/1-490
MG659815|1978/1-490
HM587073|1978/1-490
HM587071|1979/1-490
HM587060|1991/1-490
JF311381|1991/1-490
HM587059|1978/1-490
JF311382|1991/1-490
HM587056|1979/1-490
HM587057|1991/1-490
HM587051|1999/1-490
KX096942|2001/1-490
HM587055|1998/1-490
HM587054|1998/1-490
DQ380197|2000/1-490
HM587050|2000/1-490
DQ380196|1998/1-490
HM587052|1997/1-490
HM587053|1998/1-490
HM587046|1999/1-490
KY126714|1999/1-490
HM587047|1999/1-490
HM587049|1999/1-490
HM587048|1999/1-490
EU574033|2007/1-490
JF326192|2006/1-490
GQ443233|2008/1-490
EU574052|2007/1-490
KY126706|2008/1-490
GQ443250|2008/1-490
HE687303|2008/1-490
KY126704|2008/1-490
EU574050|2007/1-490
GQ443228|2008/1-490
GQ443249|2008/1-490
GQ443231|2008/1-490
EF467177|2007/1-490
JF311385|2008/1-490
JF311378|2008/1-490

Consensus

ATGATGGATCAAAATTGAAAACAAAAATGAAAGGGGTCTGCGAAGTGGGGGTTCAAGCACTCAAAAAGTGTGATGGCCAACTCAGCACTGCACATGAGGTTGTGCCCTTTGCAGTGTTTAAAAACTCAAAGAAGGTTTATCTTGATAAACTTGACCTCAAGACTGAGGAGAATCTGCTACC

Sequence 2 ID: HM587063|2007

Alignment: Building the model

Multiple Sequence Alignment (MSA) Tools / Software:

1. Clustal W
2. MUSCLE
3. MAFFT
4. Clustal Omega
5. T-Coffee
6. MULTALIN

……

Alignment: Extraction of a phylogenetic data set

For length-variable sequences with insertions deletions (indels):

Deleting unambiguously aligned regions and inserting or deleting gaps to more accurately reflect probable evolutionary processes that led to the divergence between sequences.

Determining the substitution model

- Pairwise sequence distances are calculated assuming a Markov chain model of nucleotide substitution.

- In general, substitutions are more frequent between bases that are biochemically more similar. In the case of DNA, the four types of transition (A → G, G → A, C → T, T → C) are usually more frequent than the eight types of transversion (A → C, A → T, C → G, G → T, and the reverse).

Model	Assumption
JC69	Equal rate of substitution between any two nucleotides
K80	Different rates for transitions and transversions
TN93	Different rates of transitions and transversions, heterogeneous base frequencies, and between-site variation of the substitution rate
HKY85	Variable base frequencies, one transition rate and one transversion rate
GTR	Variable base frequencies, symmetrical substitution matrix

Tree building

- Distance-based methods: use the amount of dissimilarity (the distance) between two aligned sequences to derive trees.
  - Unweighted Pair Group Method with Arithmetic Mean (UPGMA)
  - Neighbor Joining (NJ)
  - Minimum Evolution (ME)
  - Fitch-Margoliash (FM)
- Character-based methods
  - Maximum Parsimony (MP)
  - Maximum Likelihood (ML)
- Bayesian methods

Tree evalution

- Bootstrapping: Bootstrapping can be considered a two-step process comprising the generation of (many) new data sets from the original set and the computation of a number that gives the proportion of times that a particular branch (e.g., a taxon) appeared in the tree. That number is commonly referred to as the bootstrap value or support value.

- Likelihood ratio tests

Phylogenetic tree structure

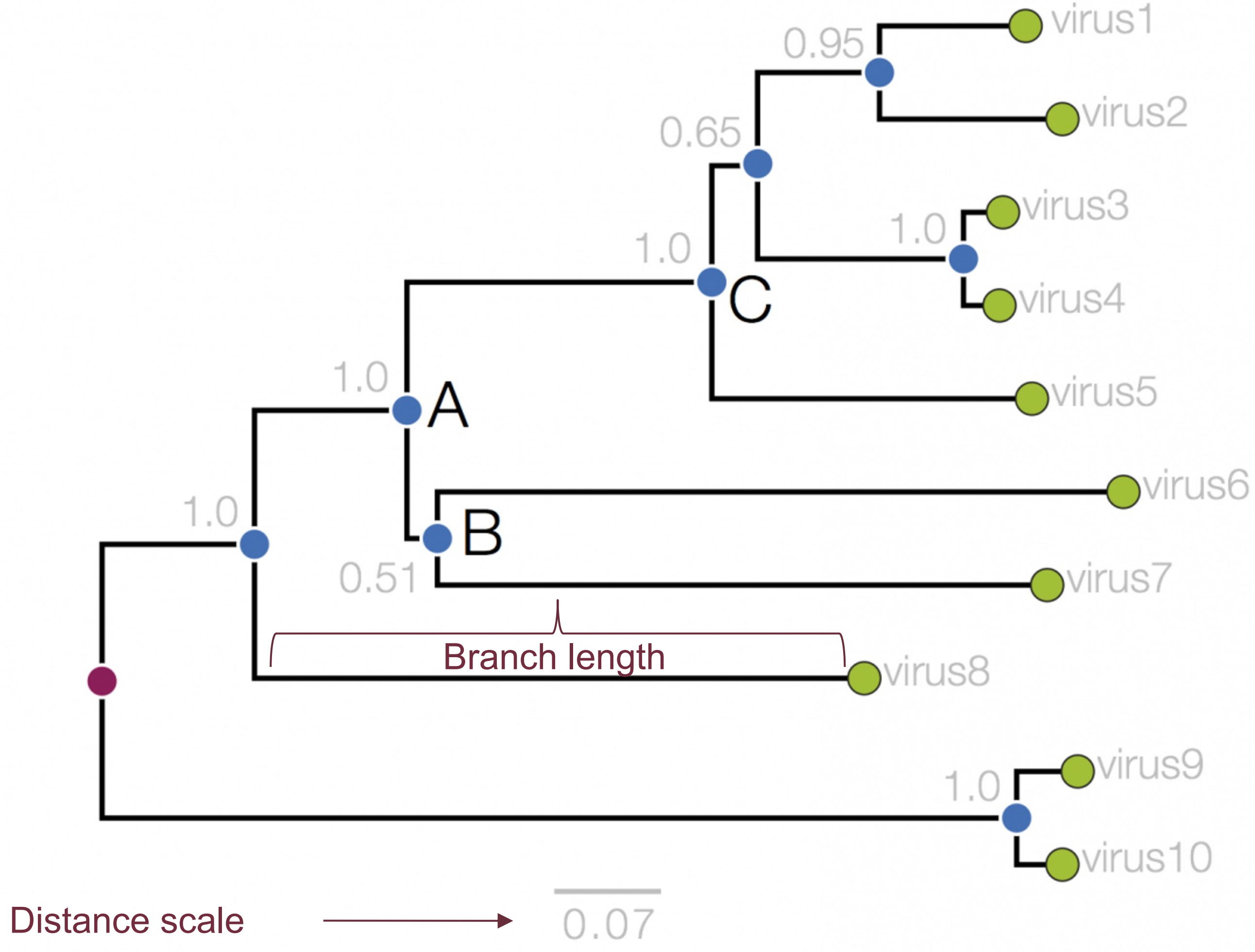


Figure 1. A fictional rooted phylogenetic tree showing the different parts of the tree.
Source. https://artic.network/how-to-read-a-tree.html

Elements of a phylogenetic tree

- The horizontal dimension gives the amount of genetic change.
- The horizontal lines are branches and represent evolutionary lineages changing over time.
- The bar at the bottom of the figure provides a scale for amount of genetic change.
  - nucleotide substitutions per site – number of changes or ‘substitutions’ divided by the length of the sequence.

Elements of a phylogenetic tree

Nodes

I. External nodes (tips or leaves)

Tips are represented by green circles and are actual viruses sampled and sequenced. Associated metadata of the tree data such as collection date, host, clinical features of the disease, collection location are always known.

II. Internal nodes.

Internal nodes are represented by blue circles and these represent putative ancestors for the sampled viruses.

Elements of a phylogenetic tree cont'd

Branches

The branches represent this chain of infections.

Root

This tree is rooted which suggests we know where the ultimate common ancestor of all the sampled viruses was (the red circle). Knowing this gives the tree an order of branching events in the horizontal dimension.

Measure of support

Numbers between 0 and 1 (but may be given as percentages) where 1 represents maximal support. These can be computed by a range of statistical approaches including bootstrapping and Bayesian posterior probabilities.

Rooting a phylogenetic tree

I. Use an outgroup – one or more sequences known to lie outside the diversity of the sequences of interest.

II. Use a method that implicitly assumes a time scale – a molecular clock mode.

Better lives through livestock

Evolutionary dynamics of viral infectious diseases

Phylodynamics

Infectious disease behavior that arise from a combination of evolutionary and ecological processes.

Variability in RNA viruses arise due to high viral mutation and replication rates. RNA viruses lack the proofreading mechanism often observed in vertebrates.

Viruses also undergo recombination thereby increasing genetic diversity.

Drivers of viral evolutionary analysis

- Increasing availability and quality of viral genome sequences.
  - 4,214,896 available SARS-CoV-2 genomes (21st March 2022) (https://www.covid19dataportal.org)
- Growth in computer processing power.
  - GPU enabled servers
  - CLIMB BIG DATA https://www.climb.ac.uk/overview/system/
- Development of sophisticated statistical methods.
  - Bayesian Evolutionary Analysis by Sampling Trees (BEAST)

Integrating evolutionary, ecological and spatial data

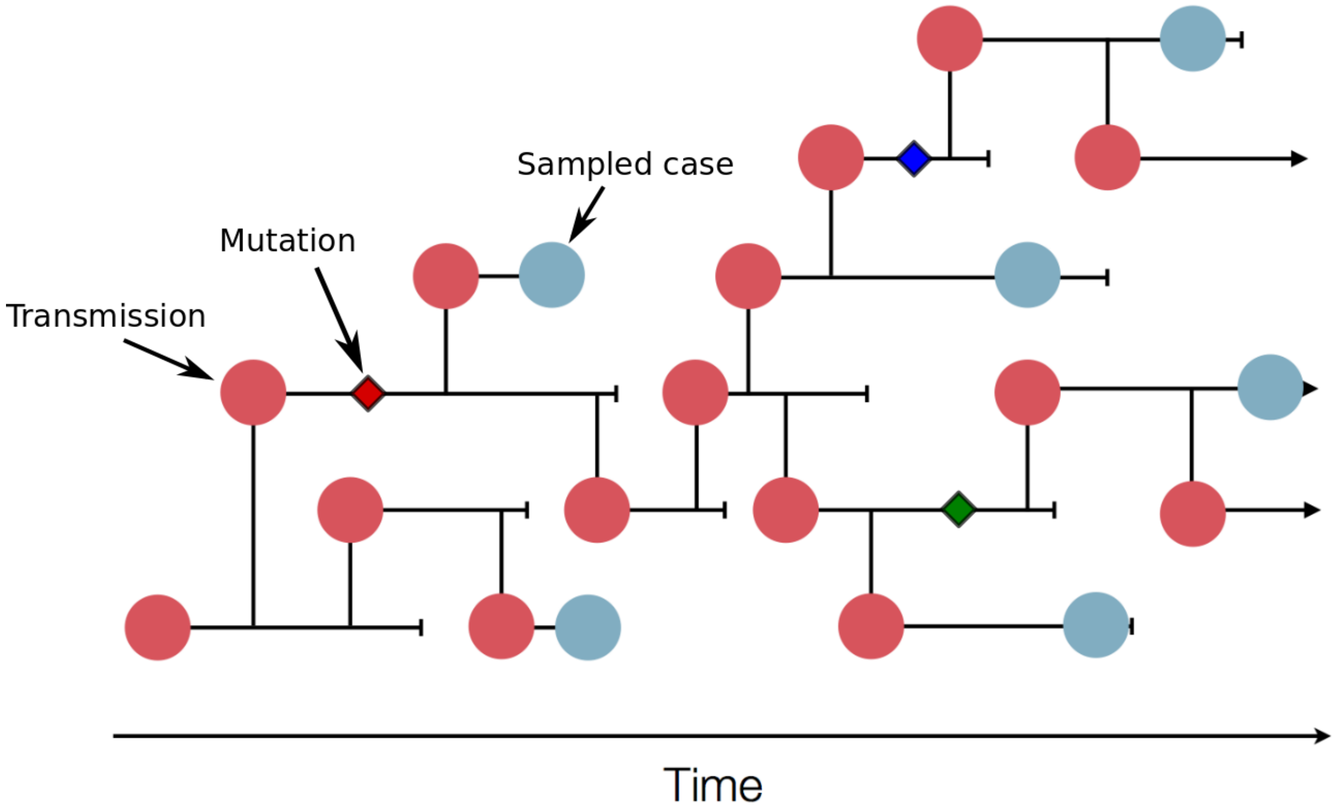
- Molecular clock: The hypothesis or observation that the evolutionary rate is constant over time or across lineages. Reconstructing evolutionary change on a natural timescale of months or years
  - Date epidemiologically important events such as zoonotic transmissions.
  - Allows pathogen evolution to be directly compared with known surveillance data.

- Phylogeography: geographic or spatial distribution of disease isolates
  - Reveal the location of origin of emerging infections
  - Discern the route of transmission
  - Rate of geographic spread

Reconstructing transmission chains

If an outbreak or infection cluster occurs on small-scale then we can realistically expect to sample viruses from all or most of the individuals involved. Studies of such outbreaks tend to fall into two categories: those for which the transmission history (that is, who infected whom, and when) is **mostly or wholly known**, and those for which it is **unknown**.

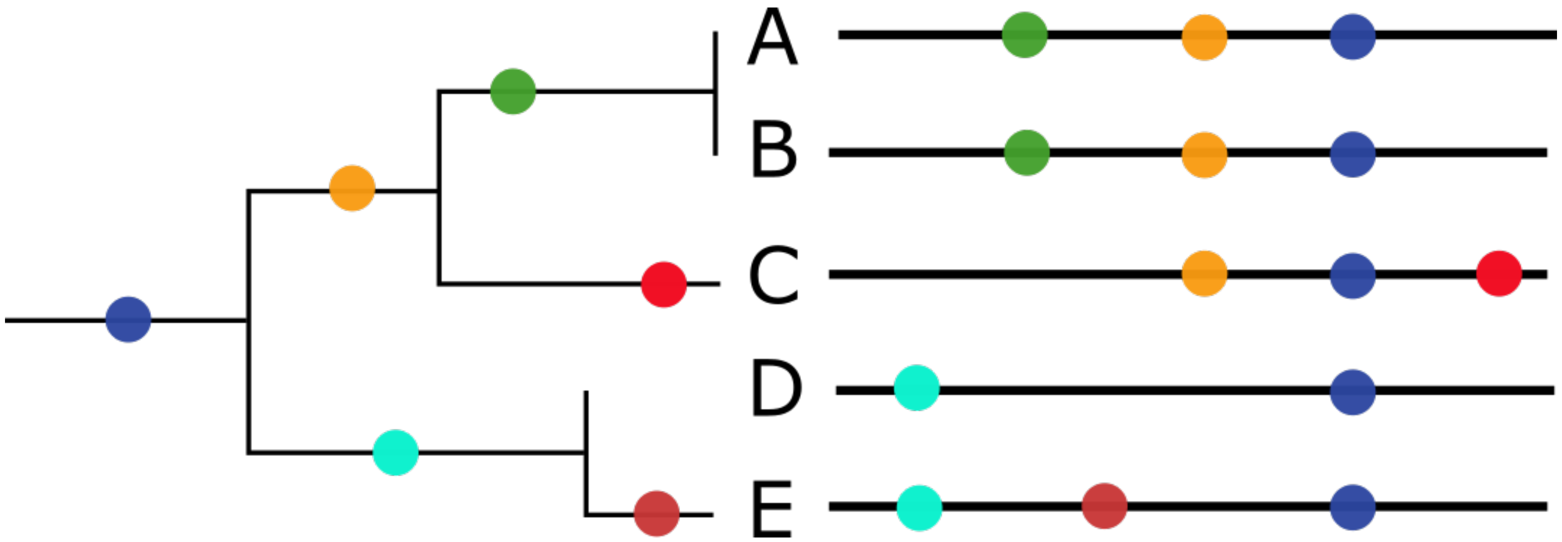
Transmission tree



A sketch of a transmission tree with a subset of cases that were sampled (blue). In practice, the transmission tree is unknown and typically only rough estimates of case counts are available. Genome sequences allow us to infer parts of the transmission tree. In this example, three mutations (little diamonds) are indicated on the tree. Sequences that have the same mutations are more closely related, so these mutations allow us to group samples into clusters of closely related viruses that belong to the same transmission chains.

Source. https://docs.nextstrain.org/en/latest/learn/interpret/how-to-read-a-tree.html

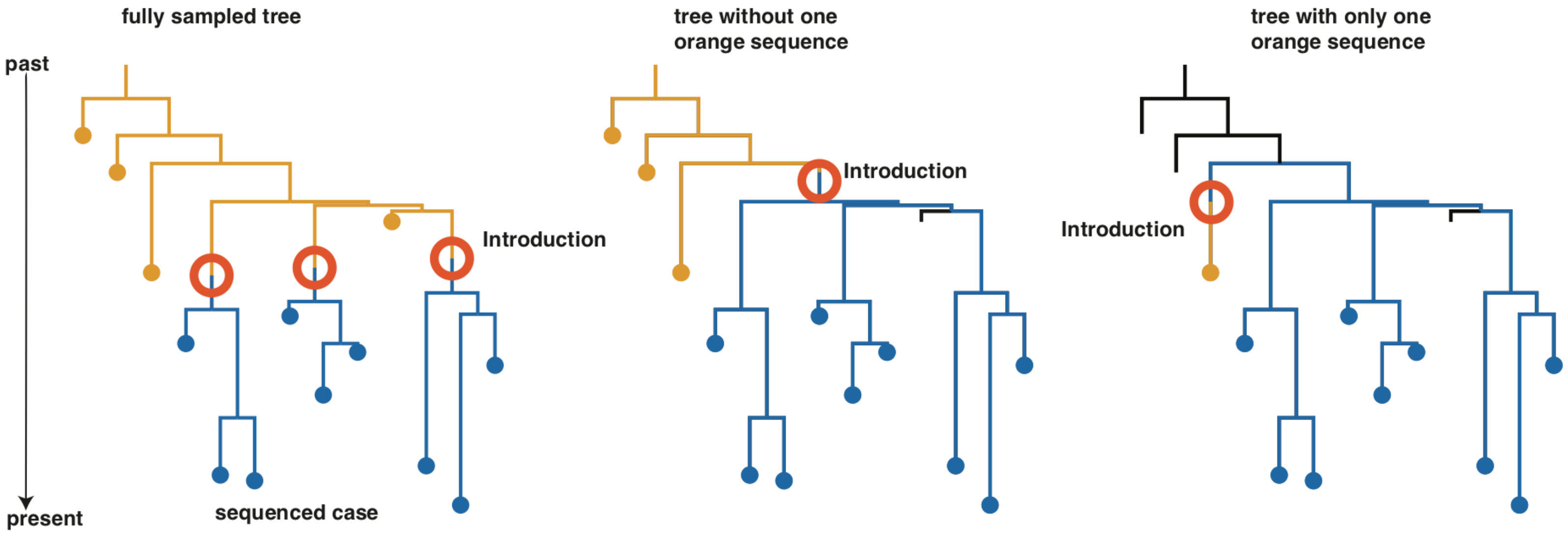
Phylogenetic tree



Source. https://docs.nextstrain.org/en/latest/learn/interpret/how-to-read-a-tree.html

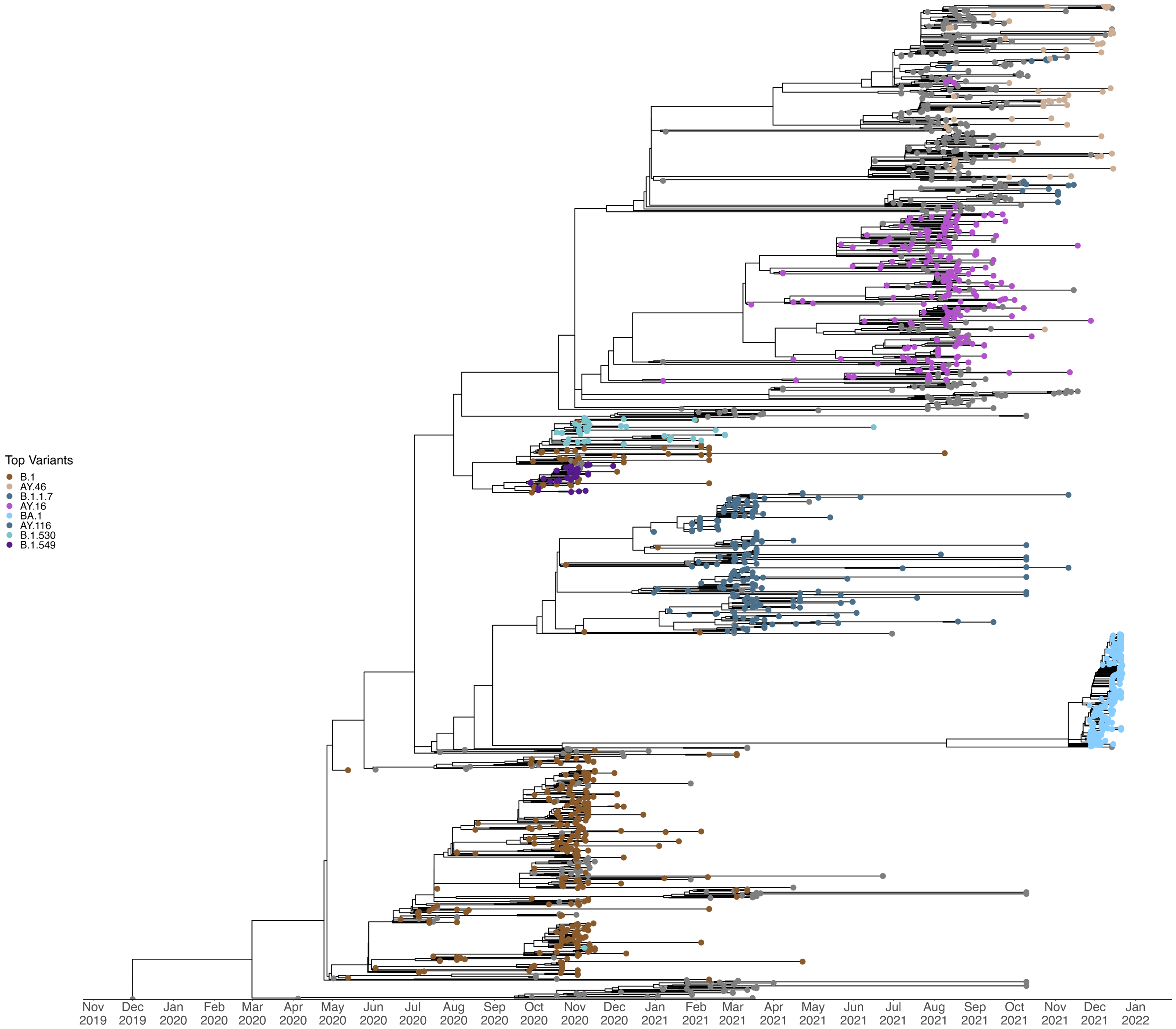
Reading a typed phylogenetic tree

Possible locations of internal nodes can be inferred using mathematical models. Interpreting these should, however, be done with caution, as the sampling and sequencing or lack thereof can significantly influence the interpretation.



Inferring locations of where a lineage has been in the past to show introductions.
Source. https://docs.nextstrain.org/en/latest/learn/interpret/how-to-read-a-tree.html

SARS-CoV-2 time-scaled phylogenetic tree



ILRI
INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

CGIAR

References and further reading

- https://artic.network/how-to-read-a-tree.html

- Pybus, O., Rambaut, A. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet* **10**, 540–550 (2009). https://doi.org/10.1038/nrg2583