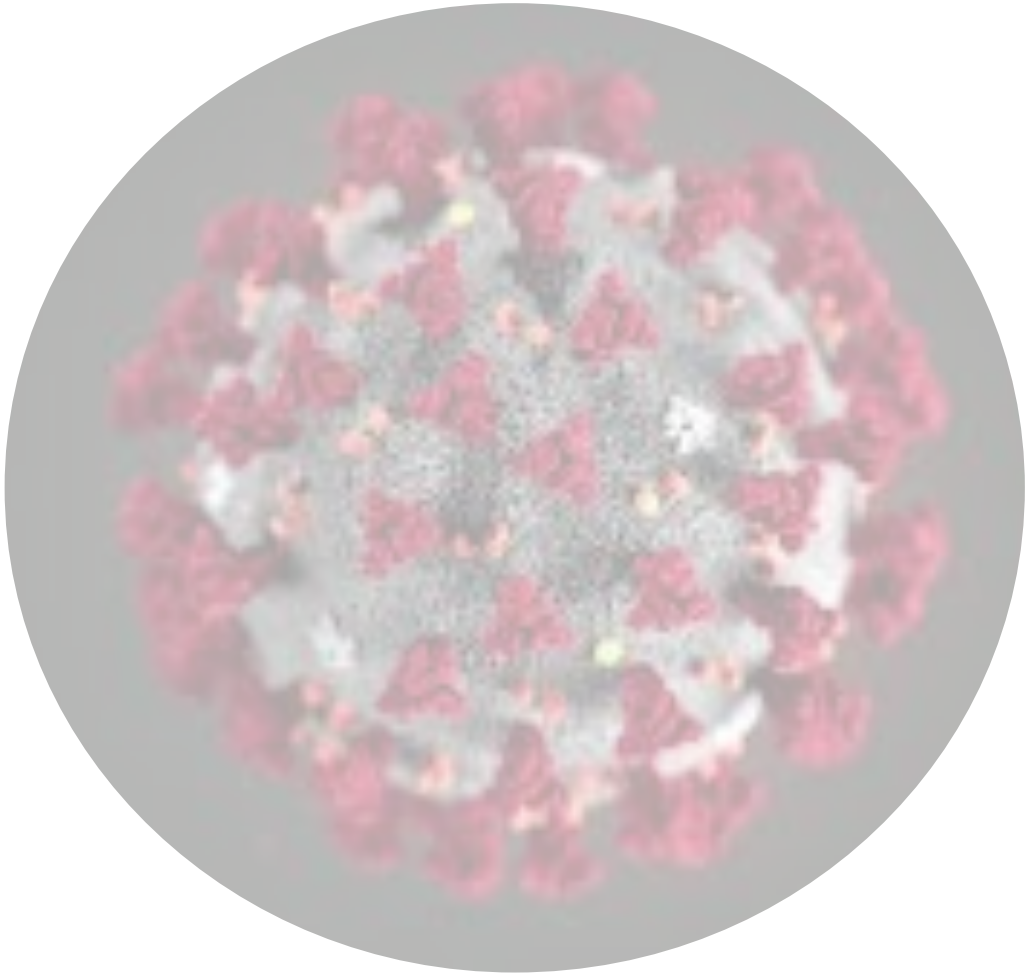
Understanding Modern Genomic Epidemiology Lecture 4



JOSEPH FAUVER, PH.D.

ASSISTANT PROFESSOR

UNMC COPH DEPARTMENT OF EPIDEMIOLOGY

5/7/2025

UNIVERSITY OF NEBRASKA MEDICAL CENTER℠
COLLEGE OF PUBLIC HEALTH

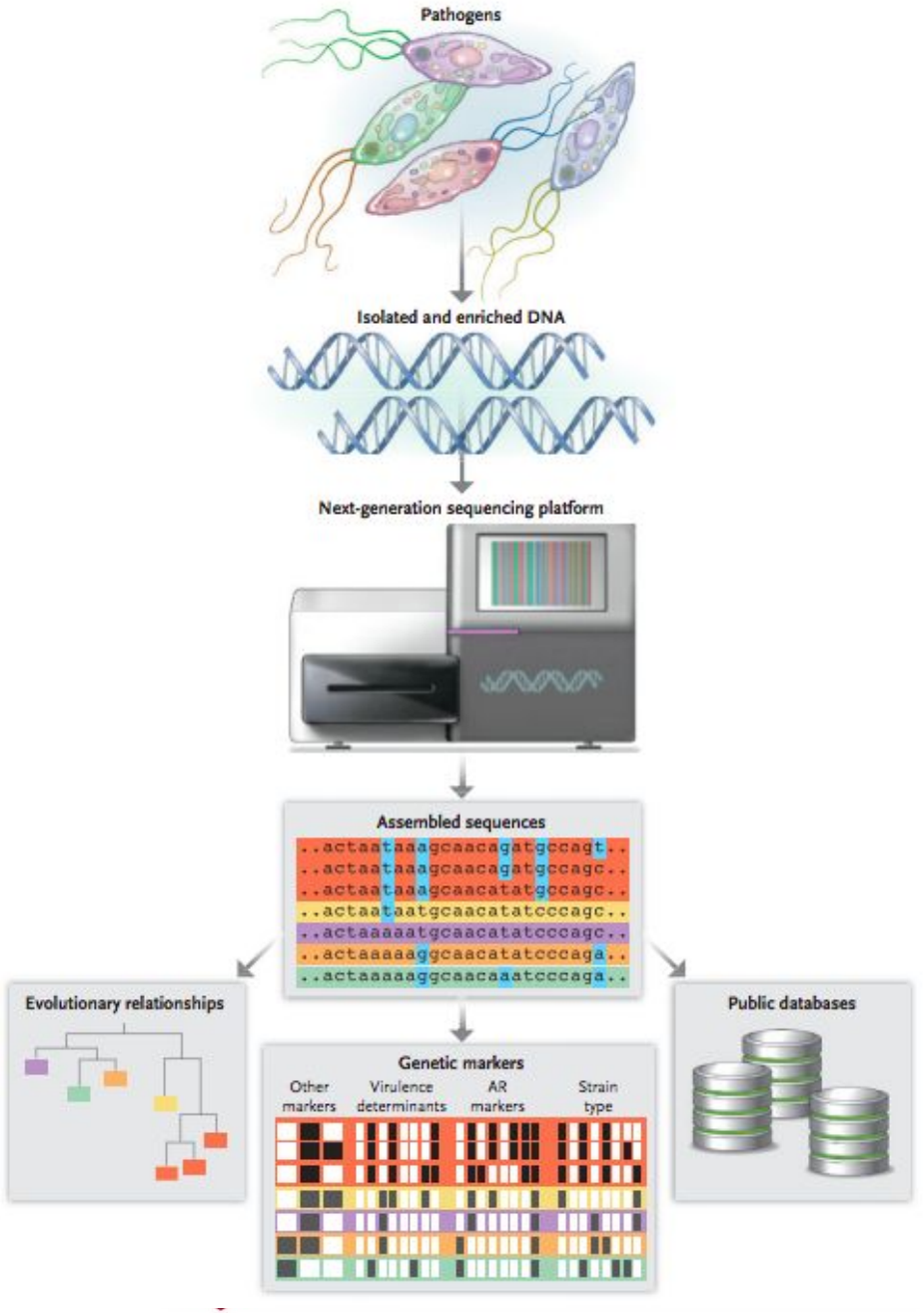
What does this look like?

"Wet" lab:

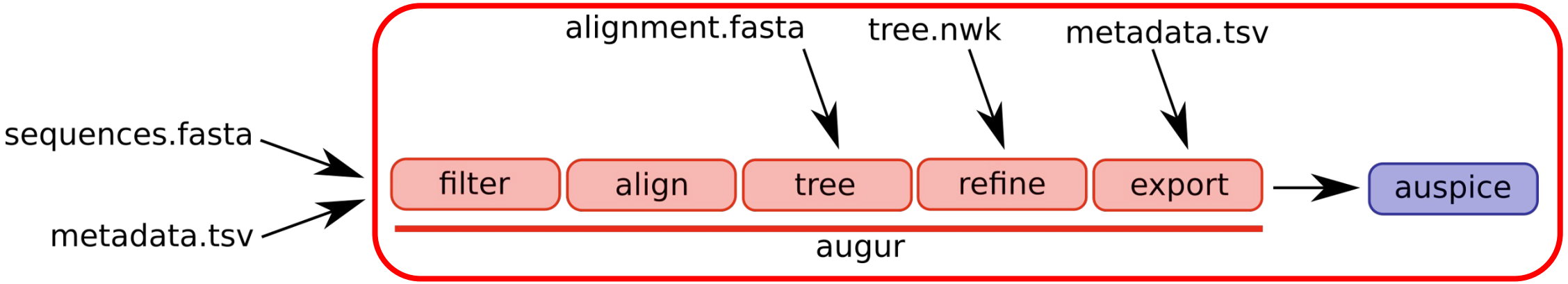
- Clinical samples/Microbial isolation
- RNA or DNA extraction
- Library preparation
- Next-Generation Sequencing

"Dry" lab:

- Data processing
- Primary genomic analyses
- Data interpretation
- Submission to repositories

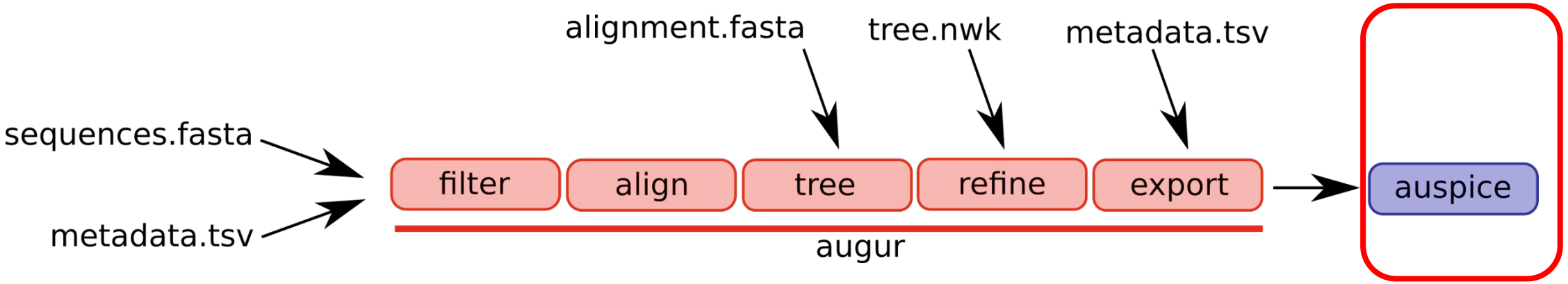


Augur- the engine behind Nextstrain



- 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 format (JSON)

Augur- the engine behind Nextstrain

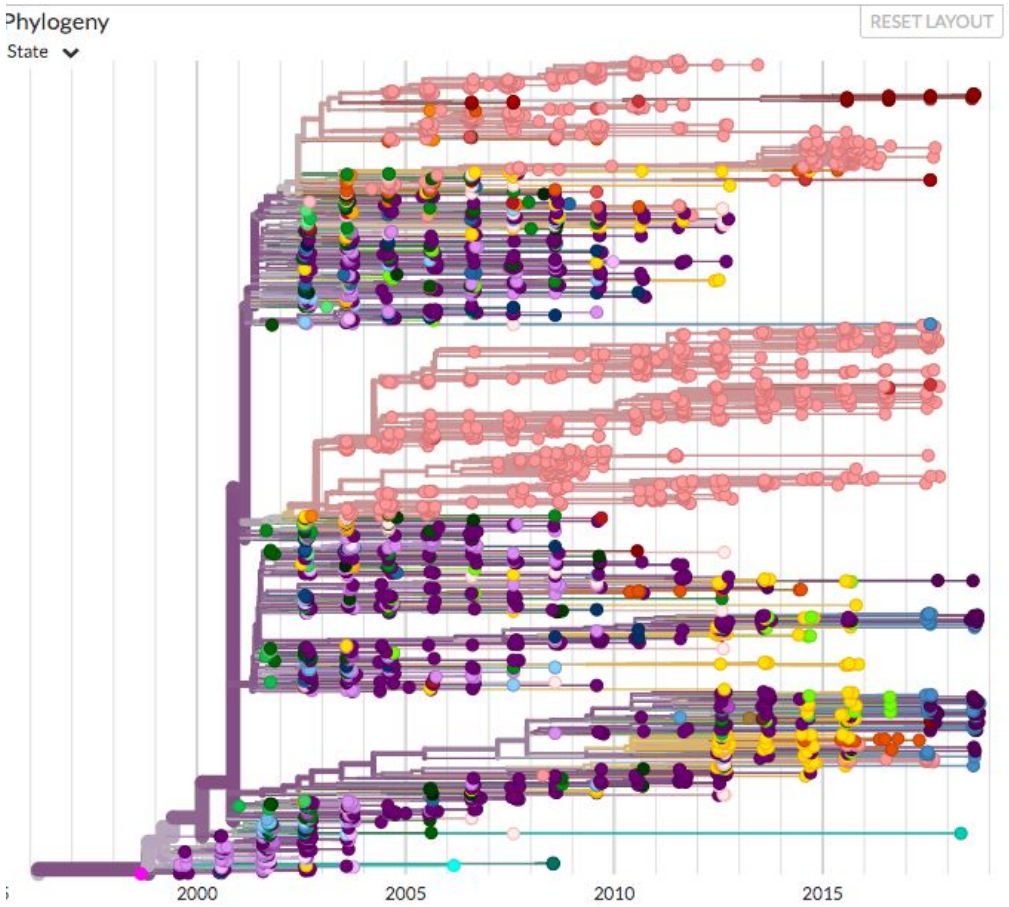


- What augur does
  - Prepare pathogen sequences and metadata
  - Align sequences
  - Construct a phylogeny from aligned sequences
  - Annotate the phylogeny with inferred ancestral pathogen dates, sequences, and traits
  - Export the annotated phylogeny and corresponding metadata into auspice-readable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

Auspice- interactive phylogenetic visualization

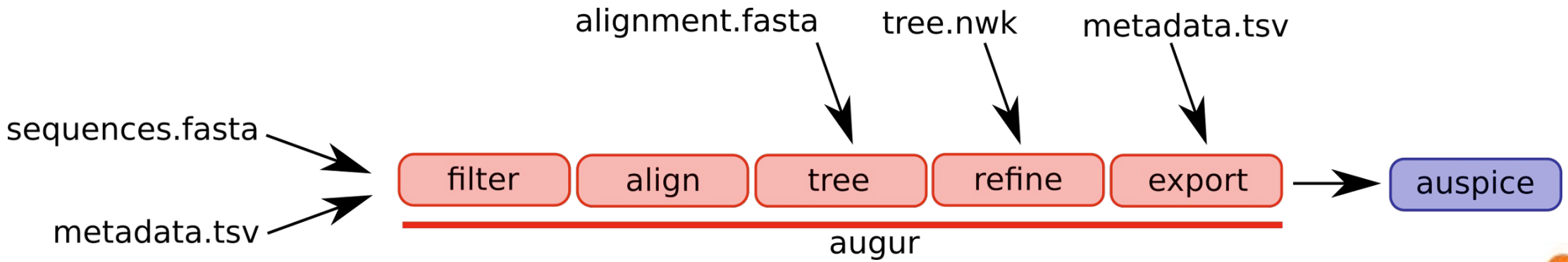
- Inputs .json files exported from augor


```
{
  "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",
        "clock_length": 0.001013391722153769,
        "country": "Israel",
        "date": "1998-08-28",
        "div": 0.001070880761851891,
        "division": "Israel/Israel/Eilat",
        "host": "Bird-other",
        "lineage": "pre-NY",
        "lineage_confidence": {
          "pre-NY": 1.0
```

What is next?

- NGS data, consensus sequence generation, data repositories
- Description of Nextstrain (what it is, what it isn't, how it works, etc.)
- **Basics of phylogenetic tree interpretations**



Nextstrain

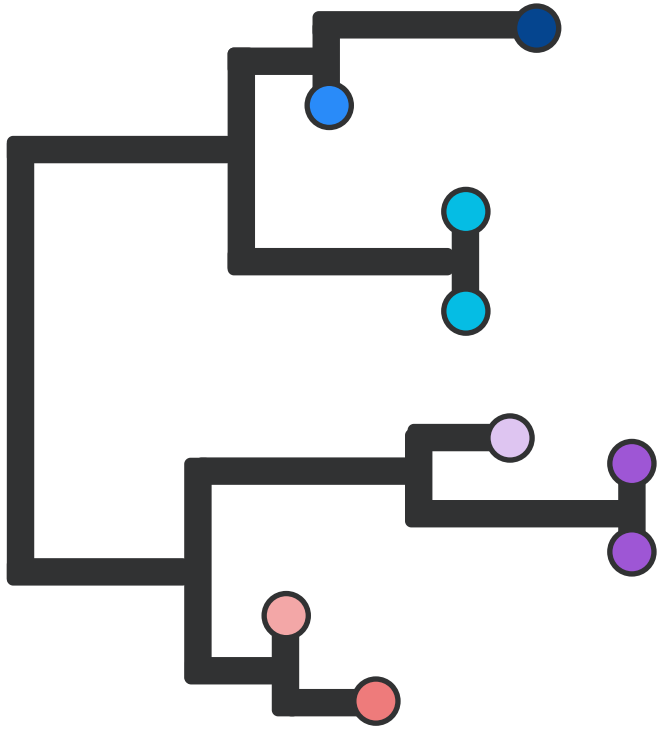
Real-time tracking of pathogen evolution

geneious prime

UNIVERSITY OF NEBRASKA MEDICAL CENTER
COLLEGE OF PUBLIC HEALTH

You got a tree- now what?

- Phylogenetics- the study of evolutionary relations among biological entities (species, populations, genes, etc.)
- These relationships are almost always inferred using molecular sequence data
  - Single genes
  - Complete genomes
  - Somewhere in between
- Relationships are visualized as bifurcating trees or phylogenies



Time or Diversity ⟹

MSA- Multi-Sequence Alignment

- Phylogenetic methods are used to determine the evolutionary relationship of organisms based on variations at sequence level (polymorphisms)
  - Due to polymorphisms, sequences derived from common ancestors (homologs) have distinct levels (%) of similarity
  - Only homologous sequences can be compared, and their variations can be used to determine how closely/distantly related the organisms are to each other
  - Goal is to get homologous sites arranged in columns.

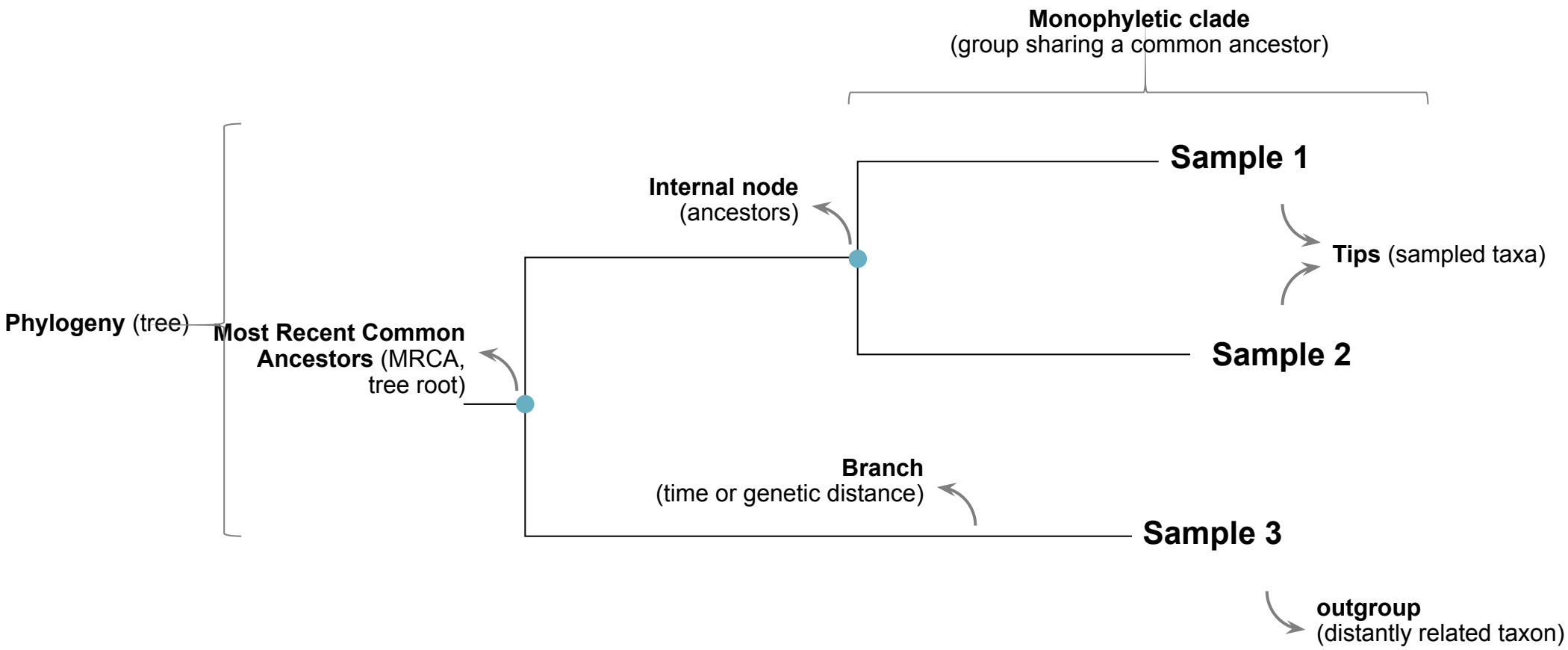
```
DENV1  AAAAGTCAGGTCAAACGCAGCTATTGGAGCAGTGTTCGTTGATGAAAATCA
DENV2  AAAGGTGAGAAGCAATGCAGCCTTGGGGGCCATATTCACTGATGAGAACAA
DENV3  AAAGGTCAGAACTAACGCAGCCATGGGCGCCGTTTTCACAGAGGAGAACCA
DENV4  AAAAGTTAGATCAAACGCAGCCATAGGCGCAGTCTTTCAGGAAGAACAGGG
ZIKV   CAAGGTGCGCAGCAATGCAGCACTGGGAGCAATATTTGAAGAGGAAAAAGA
WNV AAAAGTCAACAGTAATGCCGCCCTAGGAGCGATGTTTGAAGAACAGAACCA
YFV AAAAGTCCGAAGTCATGCAGCCATTGGAGCTTACCTGGAAGAACAAGAACA
POWV   GAAGGTGAGGTCCAACGCTGCTCTAGGTGCATGGTCGGATGAACAGAATAA

*AA*GT*********A*GC*GC**T*GG*GC**********GA**A**A***
```

Alignments of homologous sequences are the input into building a phylogenetic tree!

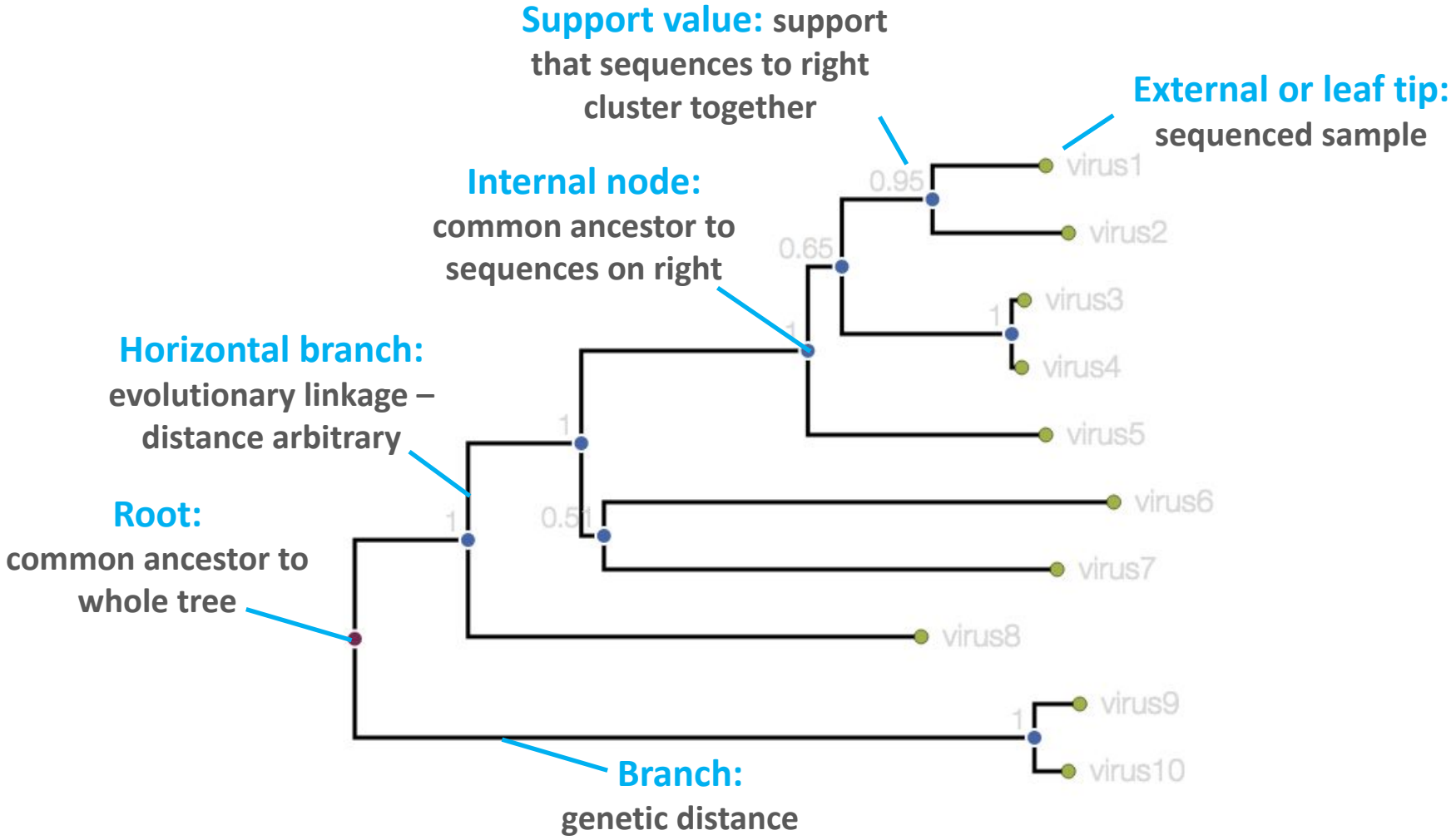
UNIVERSITY OF NEBRASKA MEDICAL CENTER℠
COLLEGE OF PUBLIC HEALTH

Basic phylogenetic nomenclature



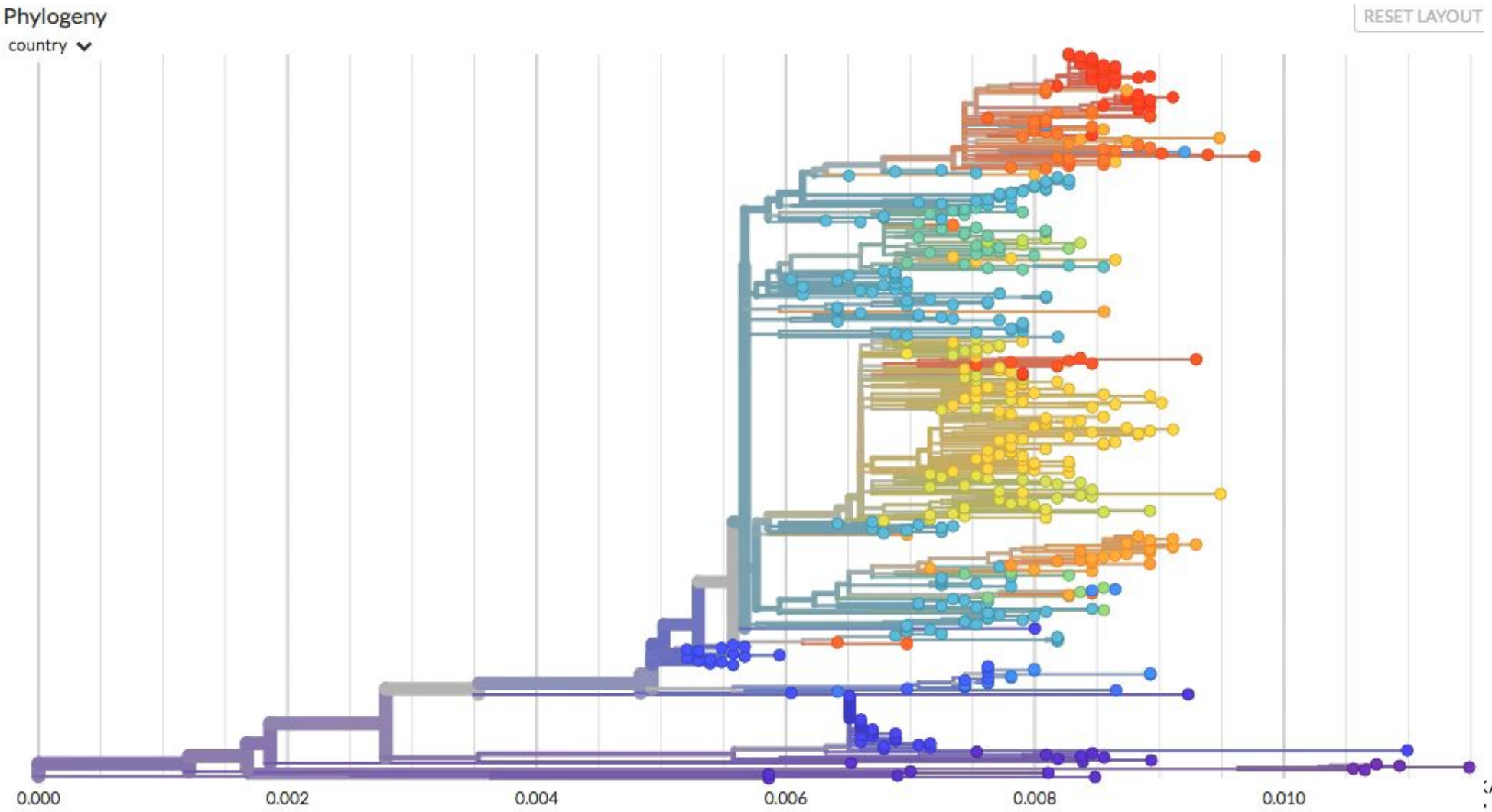
Interpreting a phylogenetic tree

Slide adapted from
Andrew Rambaut



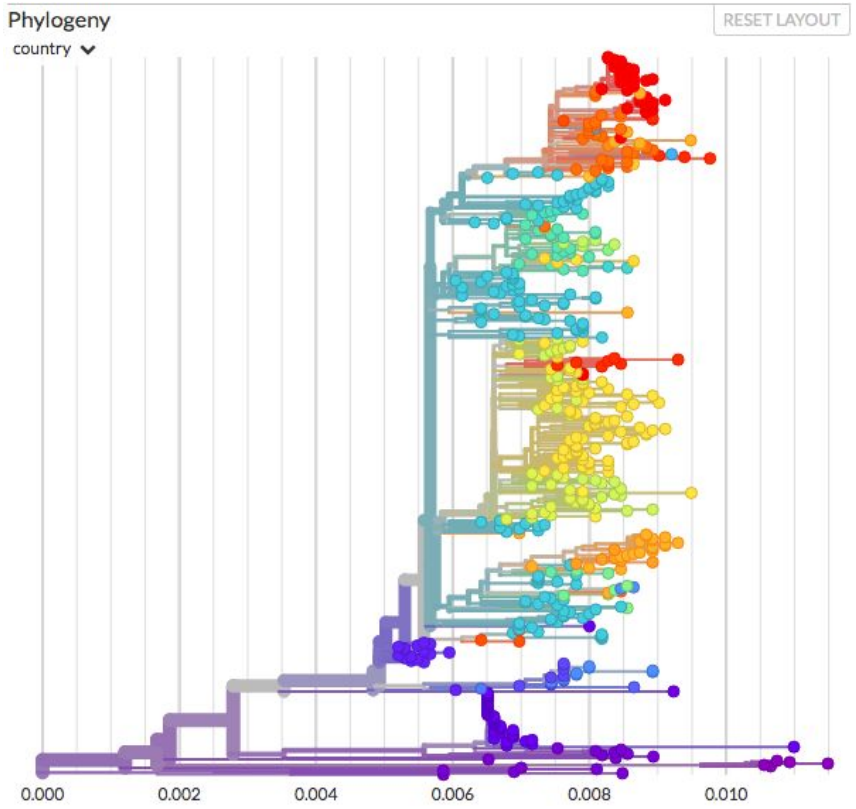
https://artic.network/how-to-read-a-tree.html#what-information-does-the-tree-contain

The same tree can be represented multiple ways

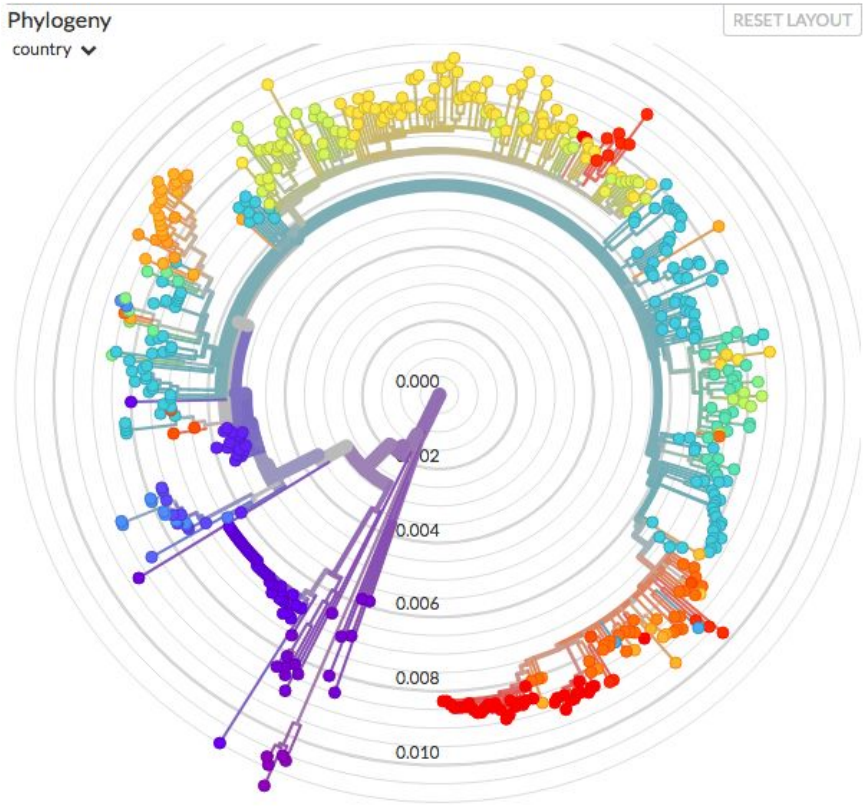


The same tree can be represented multiple ways

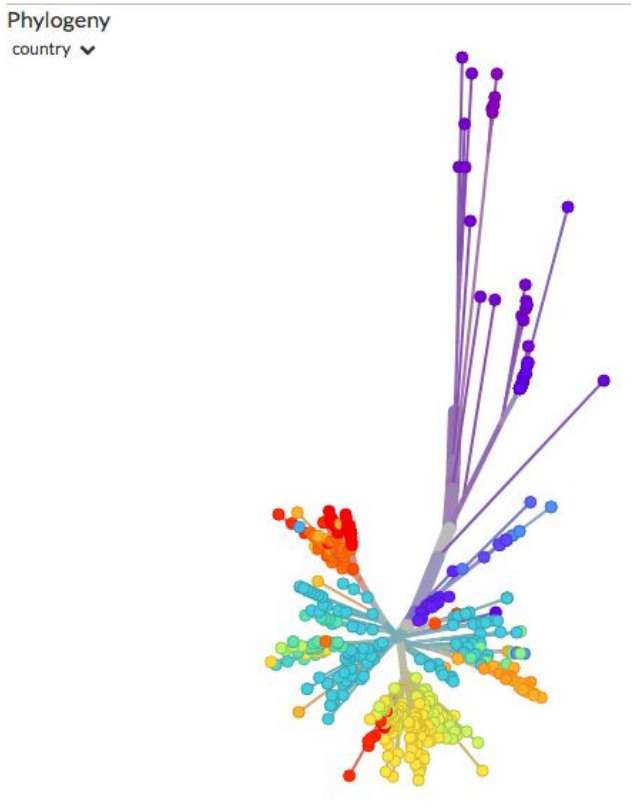
Source: nextsrain.org



Rectangular Rooted trees



Radial Rooted trees



Unrooted tree
(when outgroup is unknown)

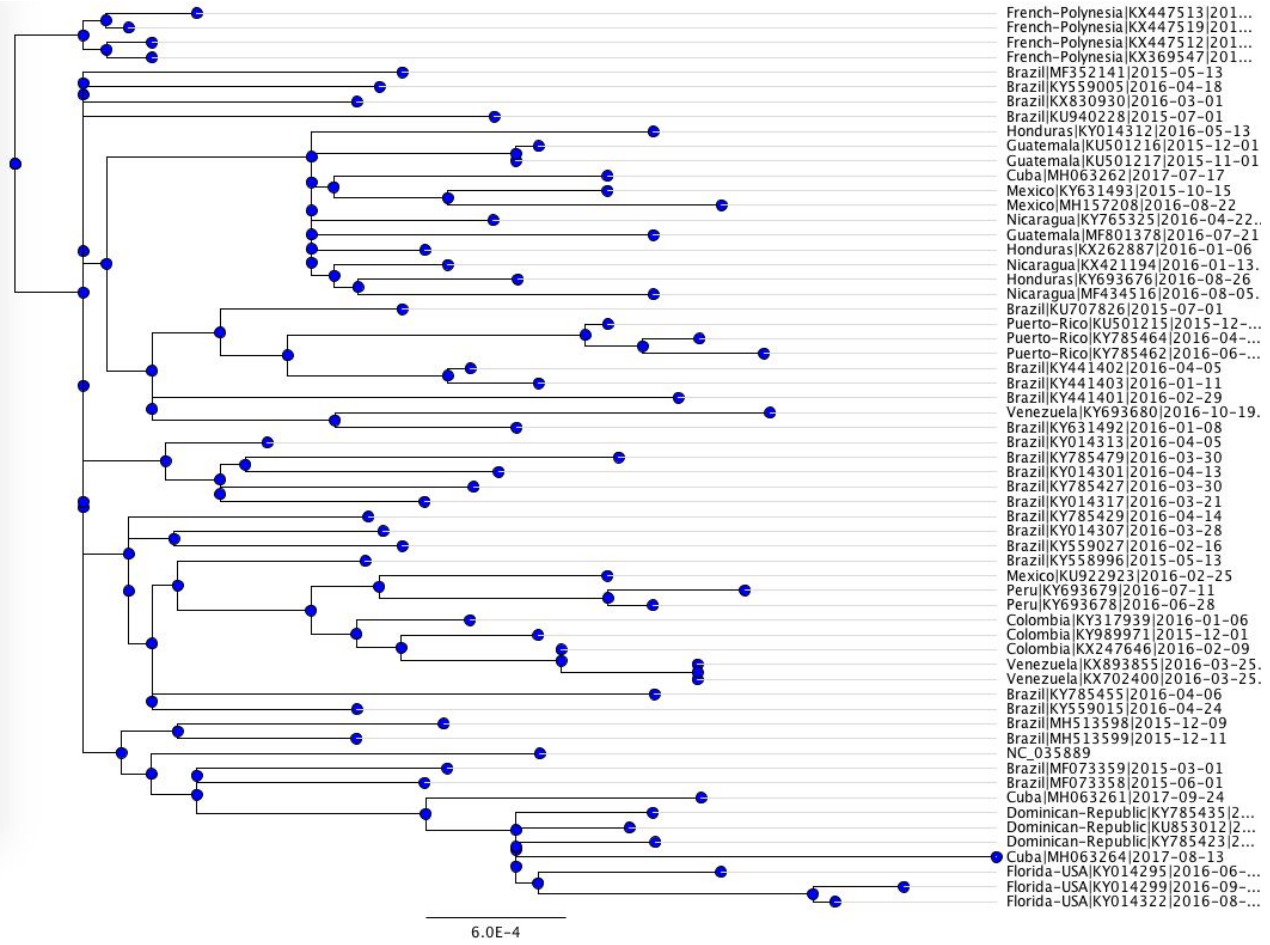
Multiple phylogenetic approaches

	Character-based methods	Non-character-based methods
Uses an explicit model of evolution	**Maximum likelihood** Bayesian inference	Pairwise distance (UPGMA & Neighbor Joining)
Do not use an explicit model of evolution	Maximum parsimony	

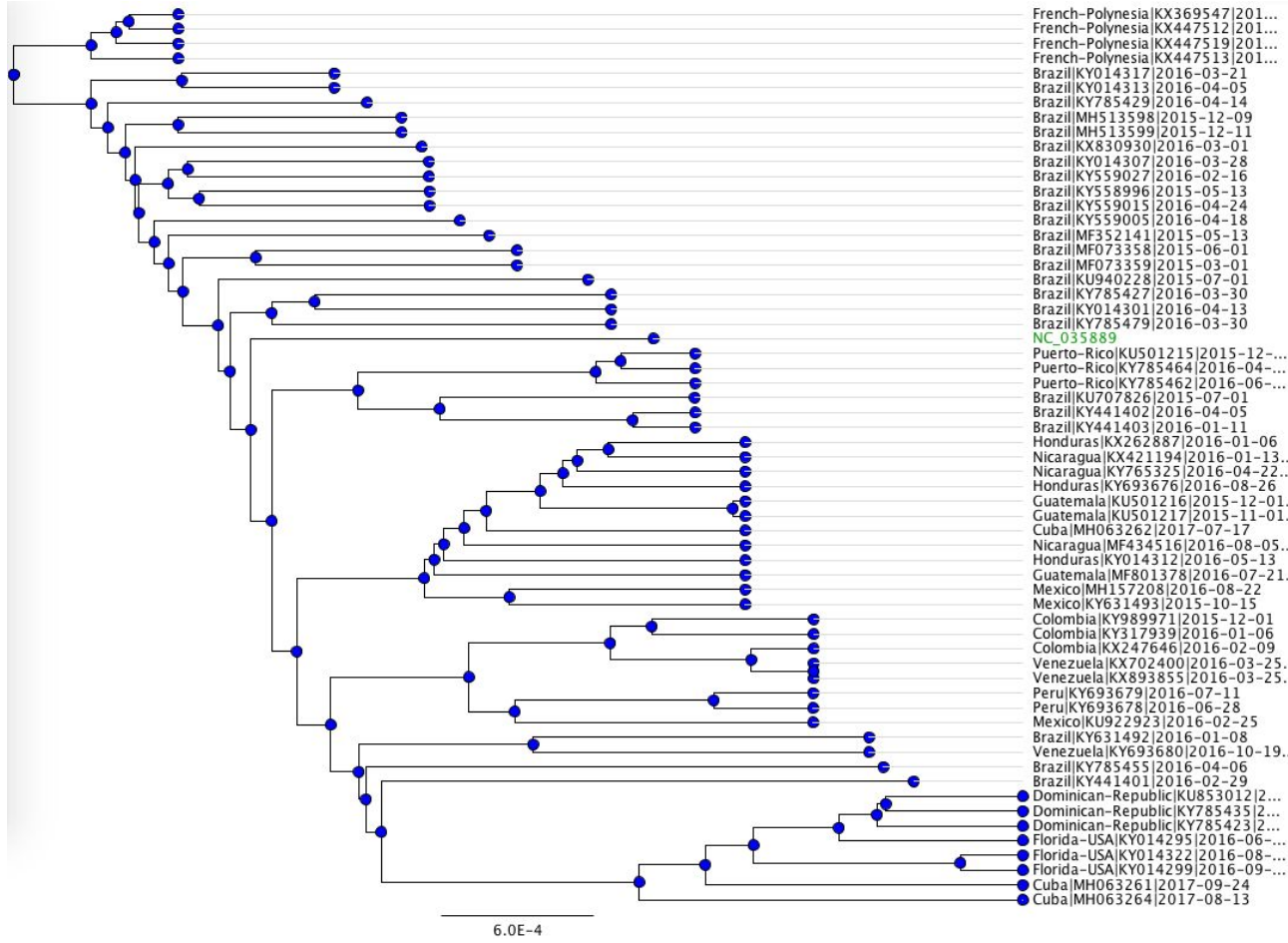
Which method to use?

- Maximum parsimony (MP)
  - Character-based, no model of evolution
  - Fast, shows general relationships, good for placing sequencing into existing trees.
- Neighbor-Joining (NJ)
  - Clusters sequences based on genetic distance
  - Fast, good for quick analysis (produces a single tree)
  - Fine for showing relationships over short time scales (days, months)
- Maximum-likelihood (ML)
  - Applies a model of sequence evolution (substitution rates)
  - More accurate than NJ, preferred for final analyses (along with Bayesian)
  - Computationally more expensive than NJ (takes more memory/time)
- Bayesian
  - Similar benefits as ML, but can include more information
  - Takes the longest to run (weeks for a tree with a few thousand samples)

ML


French-Polynesia|KX447513|201...
French-Polynesia|KX447519|201...
French-Polynesia|KX447512|201...
French-Polynesia|KX369547|201...
Brazil|MF352141|2015-05-13
Brazil|KY559005|2016-04-18
Brazil|KX830930|2016-03-01
Brazil|KU940228|2015-07-01
Honduras|KY014312|2016-05-13
Guatemala|KU501216|2015-12-01
Guatemala|KU501217|2015-11-01
Cuba|MH063262|2017-07-17
Mexico|KY631493|2015-10-15
Mexico|MH157208|2016-08-22
Nicaragua|KY765325|2016-04-22..
Guatemala|MF801378|2016-07-21
Honduras|KX262887|2016-01-06
Nicaragua|KX421194|2016-01-13.
Honduras|KY693676|2016-08-26
Nicaragua|MF434516|2016-08-05.
Brazil|KU707826|2015-07-01
Puerto-Rico|KU501215|2015-12-...
Puerto-Rico|KY785464|2016-04-...
Puerto-Rico|KY785462|2016-06-...
Brazil|KY441402|2016-04-05
Brazil|KY441403|2016-01-11
Brazil|KY441401|2016-02-29
Venezuela|KY693680|2016-10-19.
Brazil|KY631492|2016-01-08
Brazil|KY014313|2016-04-05
Brazil|KY785479|2016-03-30
Brazil|KY014301|2016-04-13
Brazil|KY785427|2016-03-30
Brazil|KY014317|2016-03-21
Brazil|KY785429|2016-04-14
Brazil|KY014307|2016-03-28
Brazil|KY559027|2016-02-16
Brazil|KY558996|2015-05-13
Mexico|KU922923|2016-02-25
Peru|KY693679|2016-07-11
Peru|KY693678|2016-06-28
Colombia|KY317939|2016-01-06
Colombia|KY989971|2015-12-01
Colombia|KX247646|2016-02-09
Venezuela|KX893855|2016-03-25.
Venezuela|KX702400|2016-03-25.
Brazil|KY785455|2016-04-06
Brazil|KY559015|2016-04-24
Brazil|MH513598|2015-12-09
Brazil|MH513599|2015-12-11
NC_035889
Brazil|MF073359|2015-03-01
Brazil|MF073358|2015-06-01
Cuba|MH063261|2017-09-24
Dominican-Republic|KY785435|2...
Dominican-Republic|KU853012|2...
Dominican-Republic|KY785423|2...
Cuba|MH063264|2017-08-13
Florida-USA|KY014295|2016-06-...
Florida-USA|KY014299|2016-09-...
Florida-USA|KY014322|2016-08-...
6.0E-4

NJ


French-Polynesia|KX369547|201...
French-Polynesia|KX447512|201...
French-Polynesia|KX447519|201...
French-Polynesia|KX447513|201...
Brazil|KY014317|2016-03-21
Brazil|KY014313|2016-04-05
Brazil|KY785429|2016-04-14
Brazil|MH513598|2015-12-09
Brazil|MH513599|2015-12-11
Brazil|KX830930|2016-03-01
Brazil|KY014307|2016-03-28
Brazil|KY559027|2016-02-16
Brazil|KY558996|2015-05-13
Brazil|KY559015|2016-04-24
Brazil|KY559005|2016-04-18
Brazil|MF352141|2015-05-13
Brazil|MF073358|2015-06-01
Brazil|MF073359|2015-03-01
Brazil|KU940228|2015-07-01
Brazil|KY785427|2016-03-30
Brazil|KY014301|2016-04-13
Brazil|KY785479|2016-03-30
NC_035889
Puerto-Rico|KU501215|2015-12-...
Puerto-Rico|KY785464|2016-04-...
Puerto-Rico|KY785462|2016-06-...
Brazil|KU707826|2015-07-01
Brazil|KY441402|2016-04-05
Brazil|KY441403|2016-01-11
Honduras|KX262887|2016-01-06
Nicaragua|KX421194|2016-01-13..
Nicaragua|KY765325|2016-04-22..
Honduras|KY693676|2016-08-26
Guatemala|KU501216|2015-12-01.
Guatemala|KU501217|2015-11-01.
Cuba|MH063262|2017-07-17
Nicaragua|MF434516|2016-08-05..
Honduras|KY014312|2016-05-13
Guatemala|MF801378|2016-07-21.
Mexico|MH157208|2016-08-22
Mexico|KY631493|2015-10-15
Colombia|KY989971|2015-12-01
Colombia|KY317939|2016-01-06
Colombia|KX247646|2016-02-09
Venezuela|KX702400|2016-03-25.
Venezuela|KX893855|2016-03-25.
Peru|KY693679|2016-07-11
Peru|KY693678|2016-06-28
Mexico|KU922923|2016-02-25
Brazil|KY631492|2016-01-08
Venezuela|KY693680|2016-10-19.
Brazil|KY785455|2016-04-06
Brazil|KY441401|2016-02-29
Dominican-Republic|KU853012|2...
Dominican-Republic|KY785435|2...
Dominican-Republic|KY785423|2...
Florida-USA|KY014295|2016-06-...
Florida-USA|KY014322|2016-08-...
Florida-USA|KY014299|2016-09-...
Cuba|MH063261|2017-09-24
Cuba|MH063264|2017-08-13
6.0E-4

Tree building softwares

- Pipelines: aligners, editors, tree building
  - Geneious (download, $$): https://www.geneious.com/
  - CLC Workbench (download, $$): https://digitalinsights.qiagen.com/products-overview/analysis-and-visualization/qiagen-clc-main-workbench/
  - UGENE (download, free): http://ugene.net/
  - MEGA (download, free): https://www.megasoftware.net/
- Tree software
  - RAxML GUI (download, free): https://sourceforge.net/projects/raxmlgui/
  - PHYLIP (download, free): http://evolution.genetics.washington.edu/phylip.html
  - BEAST (download, free): https://beast.community/index.html
  - IQ-TREE (download, free): http://www.iqtree.org/
  - NG Phylogeny (online, free) :https://ngphylogeny.fr/

Multiple phylogenetic approaches

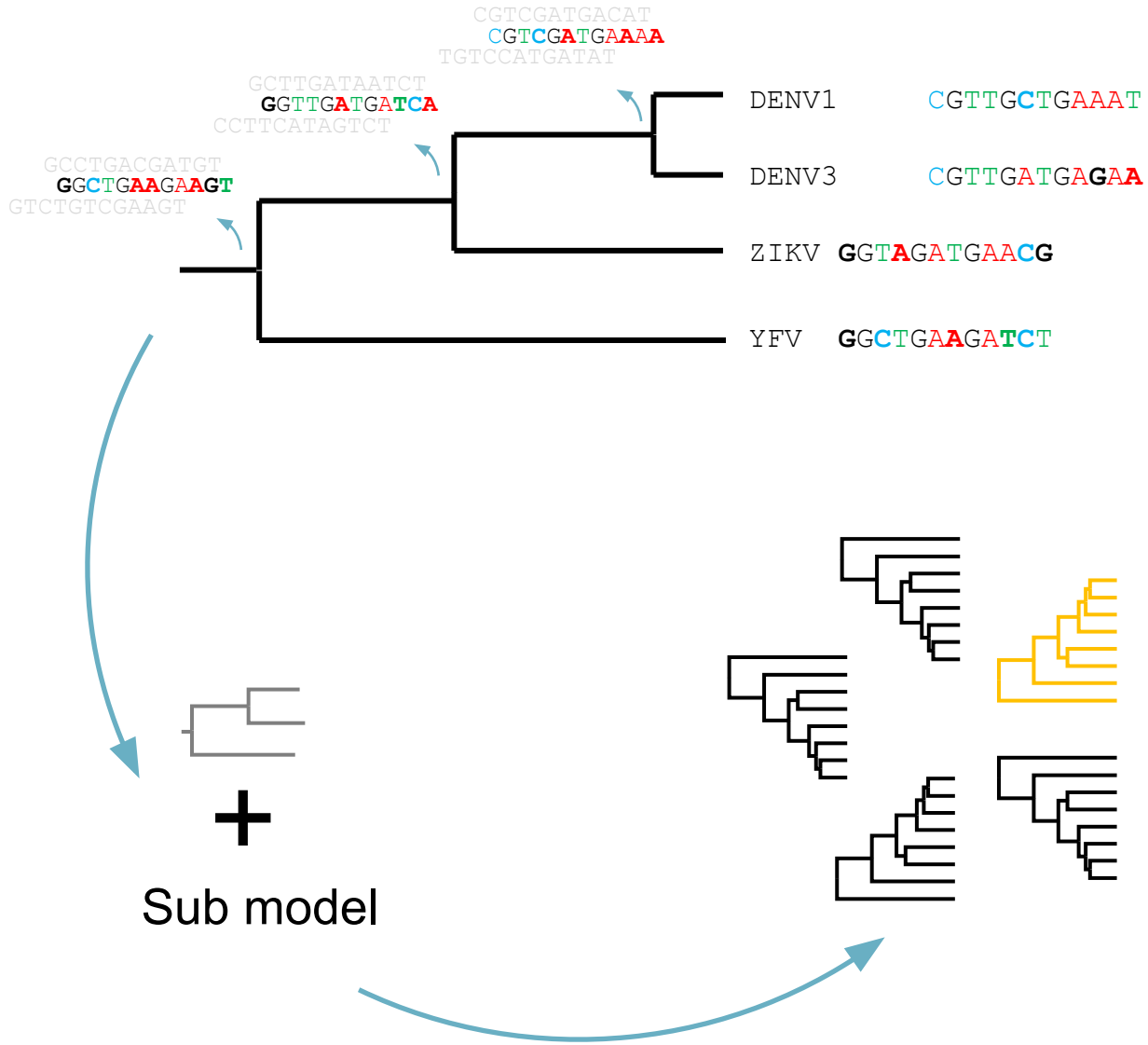
	Character-based methods	Non-character-based methods
Uses an explicit model of evolution	**Maximum likelihood** Bayesian inference	Pairwise distance (UPGMA & Neighbor Joining)
Do not use an explicit model of evolution	Maximum parsimony	

Maximum likelihood (ML) methods

ML calculates the probabilities of **ancestral sequences** using rates of substitution defined by **substitution models**

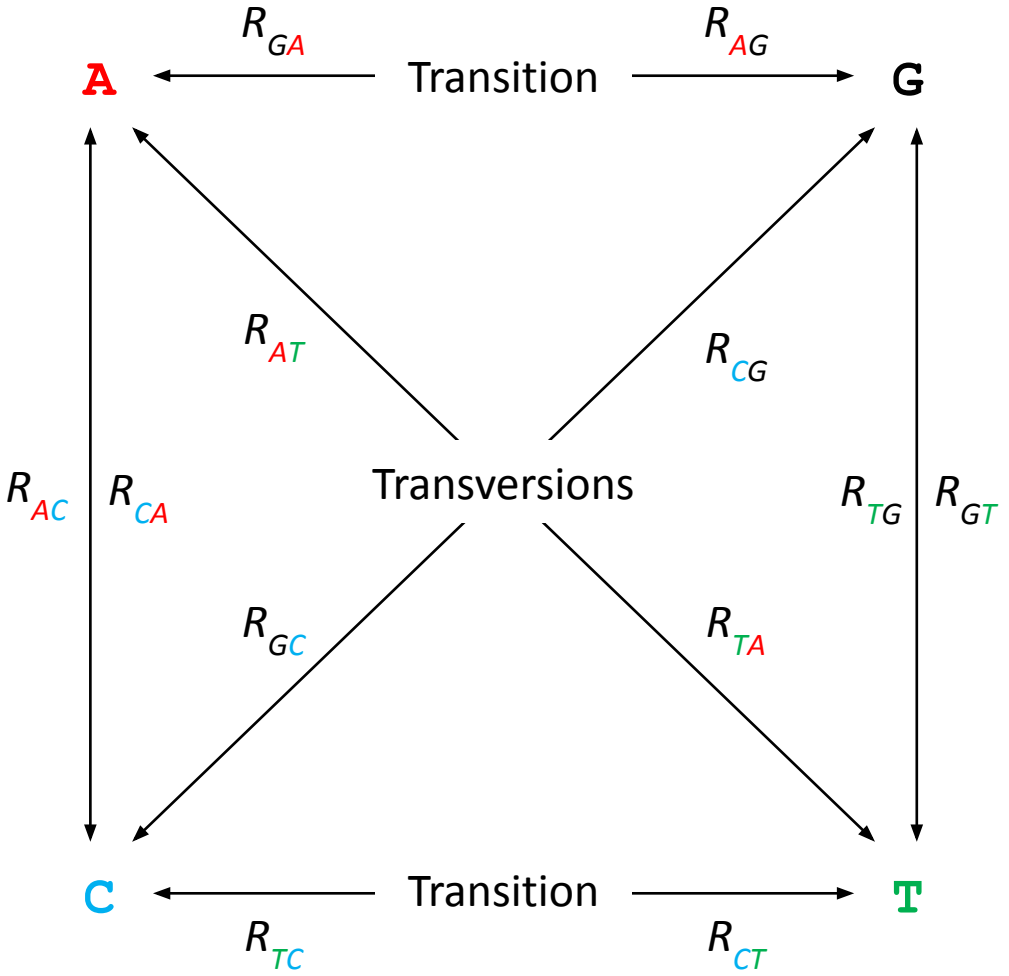
ML searches for the **most likely evolutionary scenario** that explain the sampled data (existing and ancestral sequences) given sampled tree topologies and

Many trees are proposed to explain the evolution of the sequences, and one of them is selected as **the most likely** hypothesis



Primer on substitution models

- To infer more accurate genetic distances, one must select a substitution model that best explains how the sequences evolved.
- Such models account for:
  - Rates of substitution between nucleotides
  - Rate heterogeneity among sites (1st, 2nd, and 3rd codon positions)
  - Transition/Transversion ratio
  - Nucleotide frequency



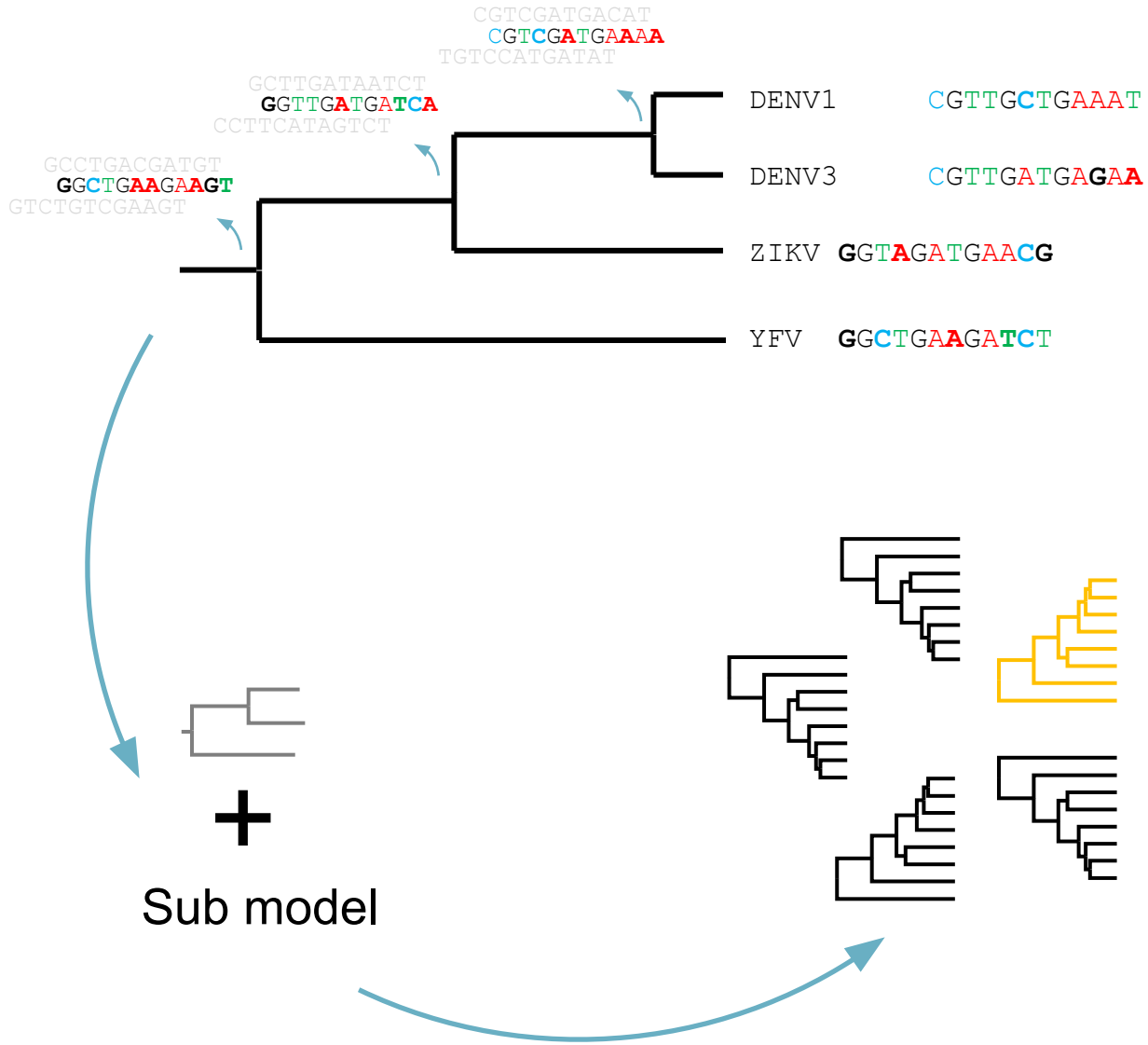
Substitution model
(explains how sequences changed)

Maximum likelihood (ML) methods

ML calculates the probabilities of **ancestral sequences** using rates of substitution defined by **substitution models**

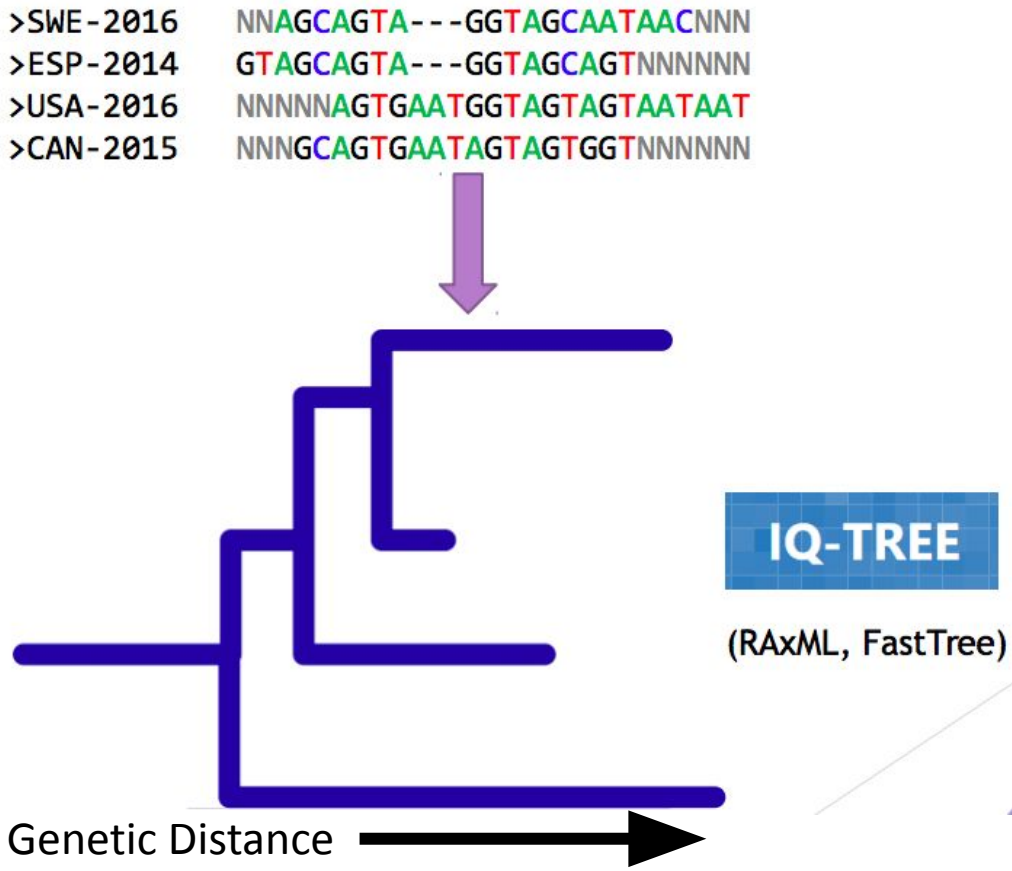
ML searches for the **most likely evolutionary scenario** that explain the sampled data (existing and ancestral sequences) given sampled tree topologies and

Many trees are proposed to explain the evolution of the sequences, and one of them is selected as **the most likely** hypothesis



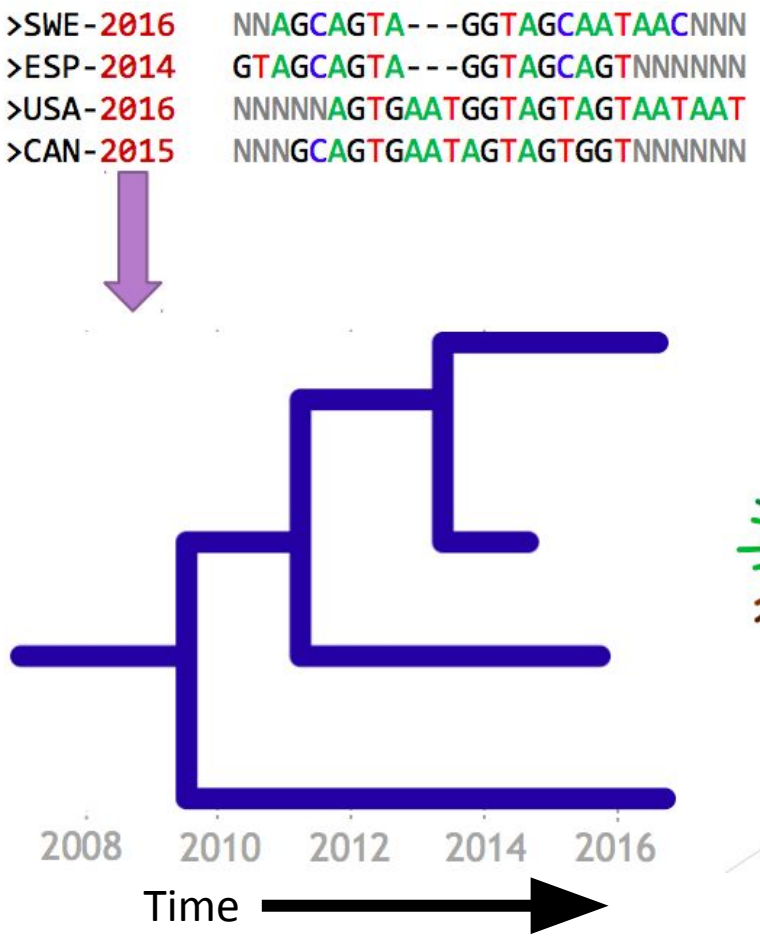
Building Phylogenies in Nextstrain

- IQ-TREE
  - Maximum-likelihood (ML)***; Fast
  - ***There are A LOT of ways to make phylogenetic trees



Building Phylogenies in Nextstrain

- IQ-TREE
  - Maximum-likelihood (ML); Fast
- TreeTime
  - Also ML – much faster than Bayesian!
  - What it does:
    - estimate discrete ancestral states
    - infer evolution models
    - reroot trees to maximize temporal signals
    - estimate molecular clock phylogenies
    - estimate population size histories



Useful manuscripts

- Nextstrain Manuscript: https://doi.org/10.1093/bioinformatics/bty407
- Tracking virus outbreaks in the twenty-first century: https://doi.org/10.1038/s41564-018-0296-2
- Resurgence of Ebola virus in 2021 in Guinea suggests a new paradigm for outbreaks: https://doi.org/10.1038/s41586-021-03901-9
- MERS-CoV spillover at the camel-human interface: https://doi.org/10.7554/eLife.31257
- Twenty years of West Nile virus spread and evolution in the Americas visualized by Nextstrain: https://doi.org/10.1371/journal.ppat.1008042
- Genomic Analysis of Lassa Virus during an Increase in Cases in Nigeria in 2018: https://doi.org/10.1056/nejmoa1804498
- The emergence of SARS-CoV-2 in Europe and North America: https://doi.org/10.1126/science.abc8169
- Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States: https://doi.org/10.1016/j.cell.2020.04.021
- Origins of the current outbreak of multidrug-resistant malaria in southeast Asia: a retrospective genetic study: https://doi.org/10.1016/S1473-3099(18)30068-9
- Phylogenomic characterization and signs of microevolution in the 2022 multi-country outbreak of monkeypox virus: https://doi.org/10.1038/s41591-022-01907-y

Questions?

Reach out via email: jfauver@unmc.edu