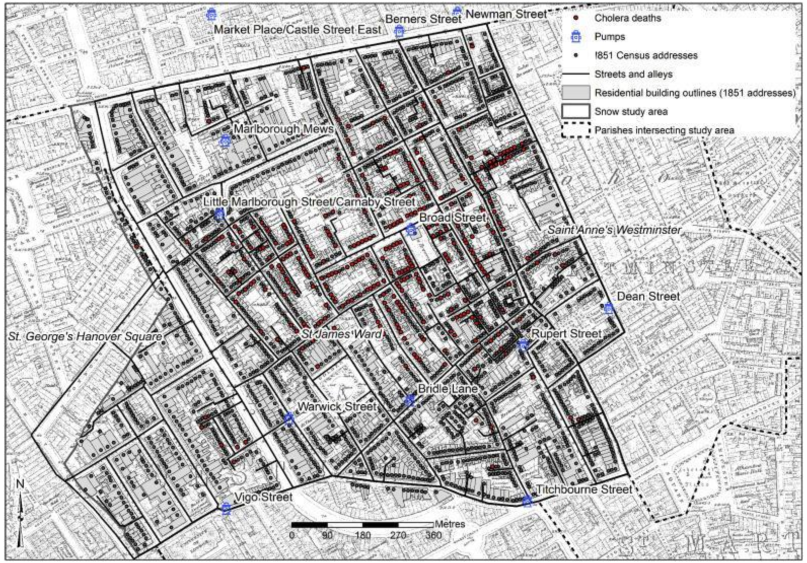
Agenda

- Introduction to genomic epidemiology
- Introduction to phylogenetics
- Creating multiple sequence alignments
- Phylogenetic reconstruction methods and tools
- Interpreting Trees
- Tree generation pitfalls

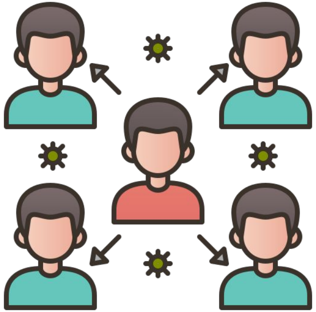
Genomic Epidemiology

What is epidemiology?

Study of the ***occurrence*** *and* ***causes of diseases*** *in a* ***population***

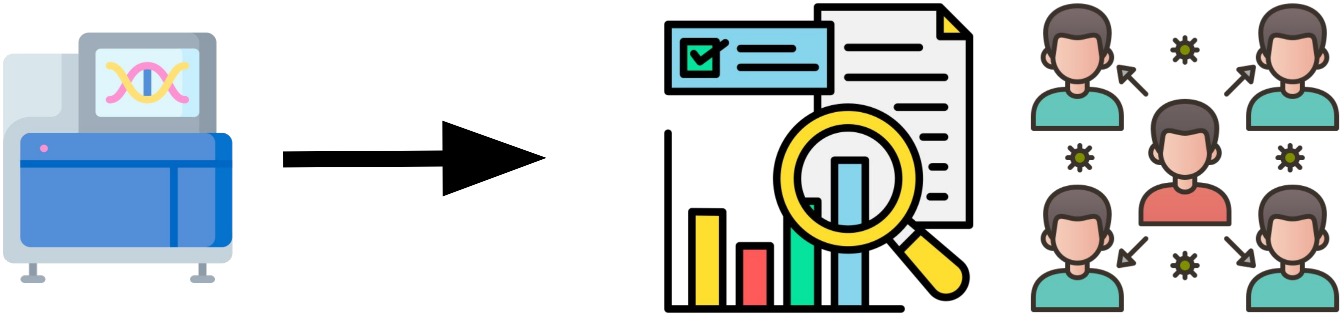


Newsom et al. 2006.



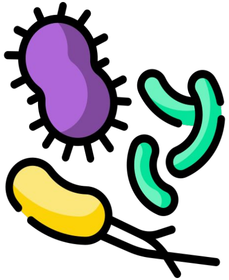
What is genomic epidemiology?

<u>Genomic Epidemiology</u>: Use of **pathogen genomic data** to study of the occurrence and causes of diseases in a population

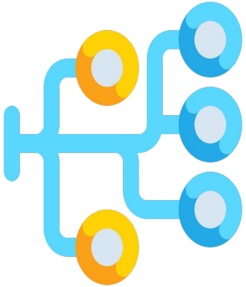


Genomic epidemiology comprises:

1. **Surveillance** and **typing** of pathogens over time and space

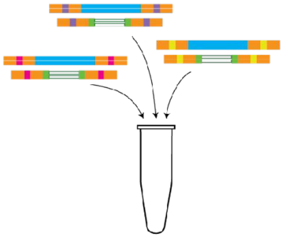


1. **Evolutionary history** (phylogenetics)

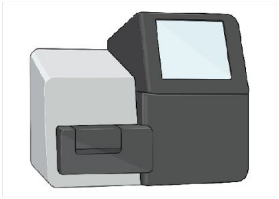


Genomic epidemiology workflow

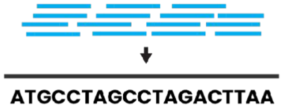
Sample collection



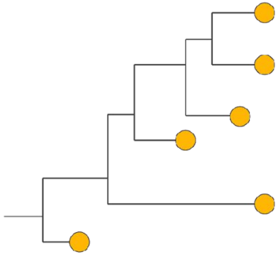
Library preparation



Sequencing



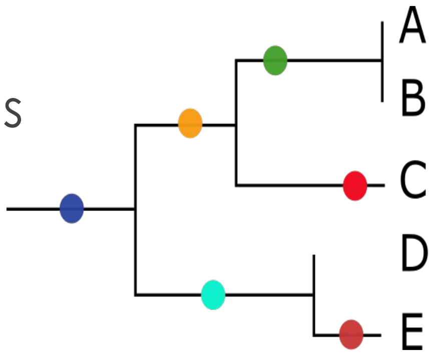
Genome assembly



Phylogenetic analysis

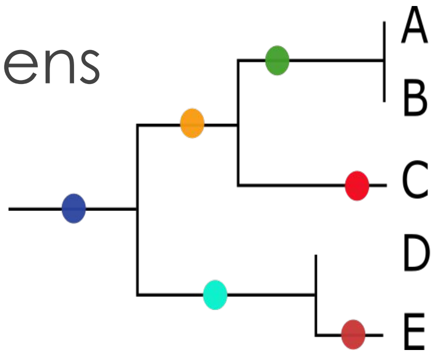
Genomic epidemiology facilitates…

- Identification of introductions of new pathogens
  - Identify the pathogen taxon
  - Characterize the pathogen
  - Identify the primary reservoir(s) (animal, human, or environmental source)
- Understanding the transmission dynamics
  - Determine clusters of closely related cases
  - Track the timeline of pathogen introduction



Genomic epidemiology facilitates...

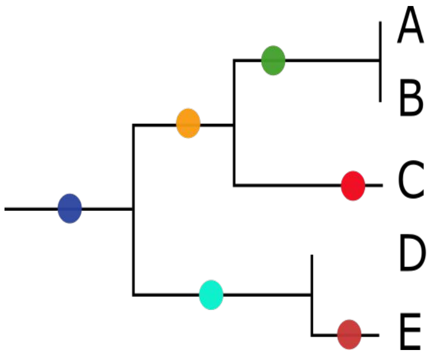
Identification of Introductions of New Pathogens



- Identify the pathogen of interest
  - Genus and species taxonomic ID assignment
- Characterize the pathogen
  - Strain, serotype, genotype, variants, lineage, etc.
- Identify the primary reservoir(s) (animal, human, or environmental source)
  - Determine the pathogen spills over into human populations
  - Highlight potential intermediate hosts involved in transmission

Genomic epidemiology facilitates...

Understanding Transmission Dynamics

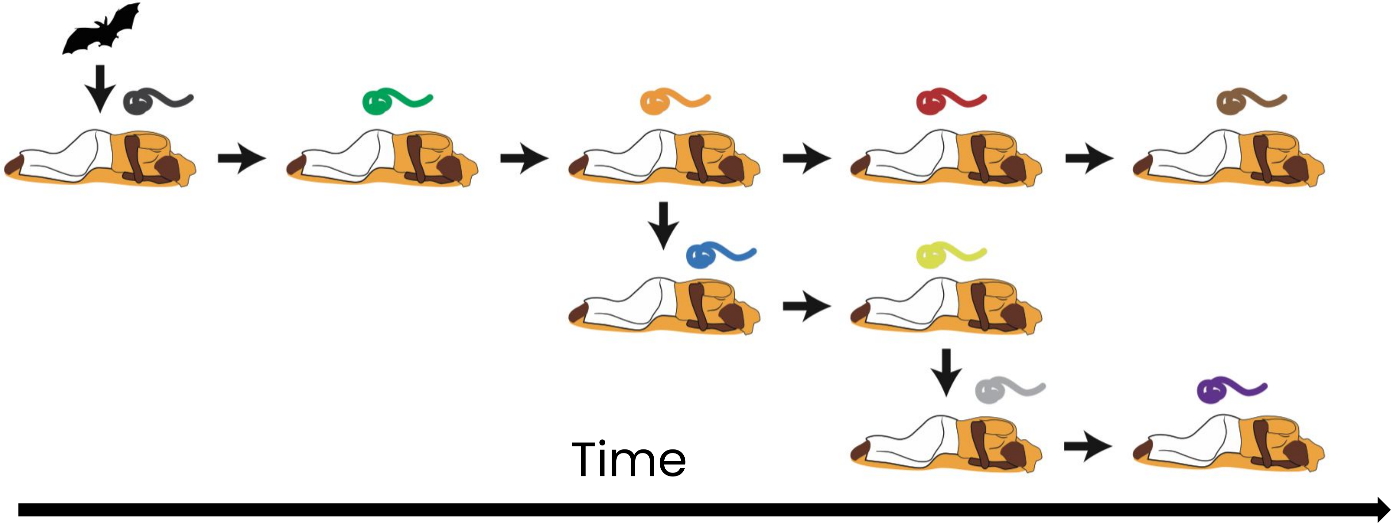


- Determine clusters of closely related cases
  - Identify clusters of cases using genomic relatedness
  - Investigate if cases share common exposures or routes of infection
  - Assess if the pathogen is spreading between humans or via other routes
- Track the timeline of pathogen introduction
  - Analyze the timeline of the introduction
  - Estimate the date of introduction based on mutation rates or case detection

Pathogen Evolution During Spread

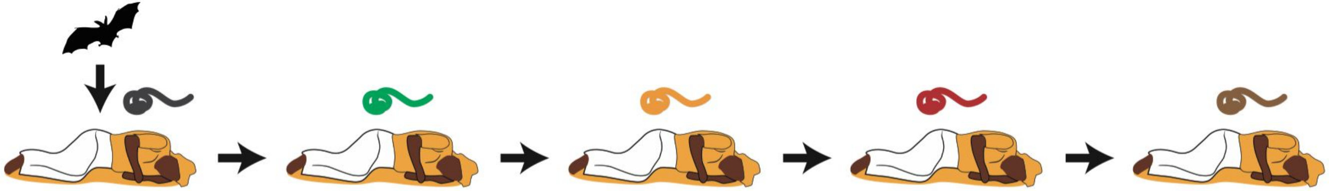
Pathogen genome evolution during spread

The pathogen evolves as it spreads between hosts

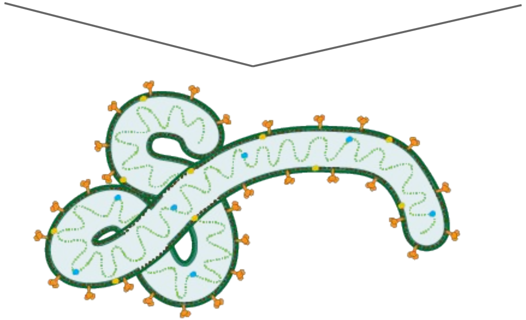


Pathogen genome evolution during spread

Mutations accumulate in the pathogen genome

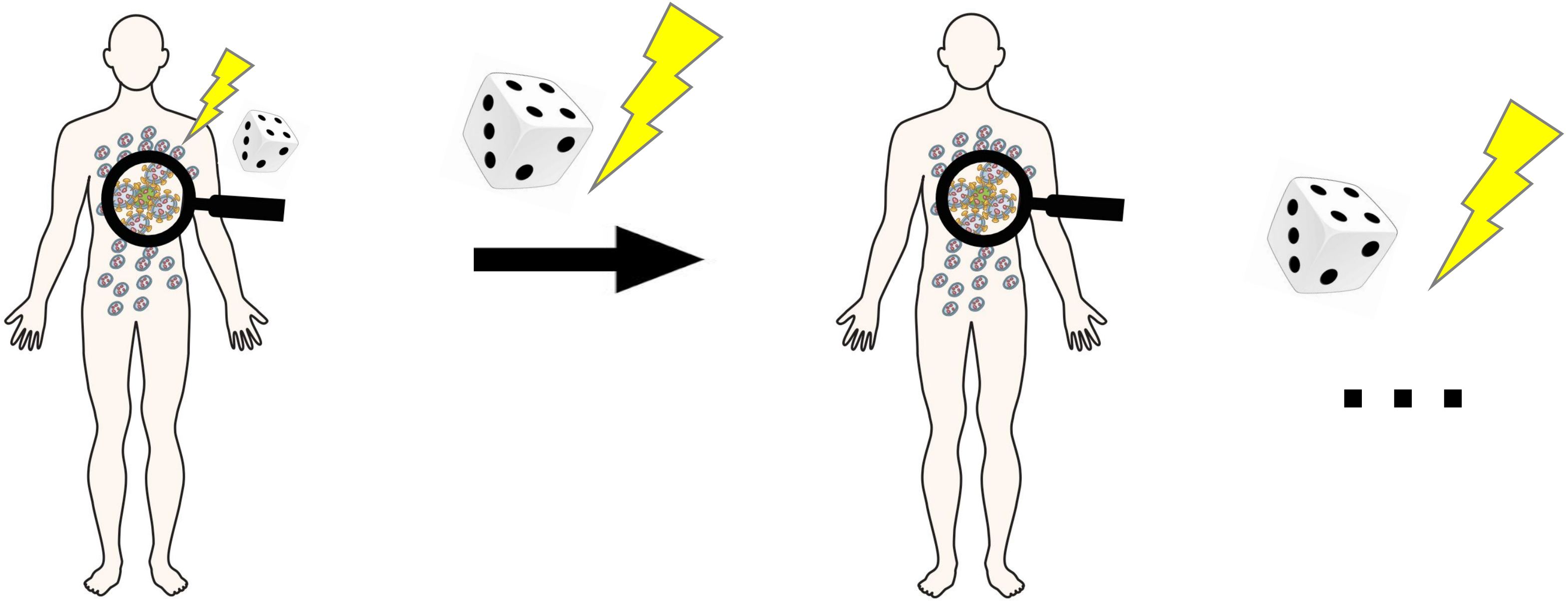


ATGCAG**C**TAGCT**G**ATGCTGAC**T**GACTGAC**T**G



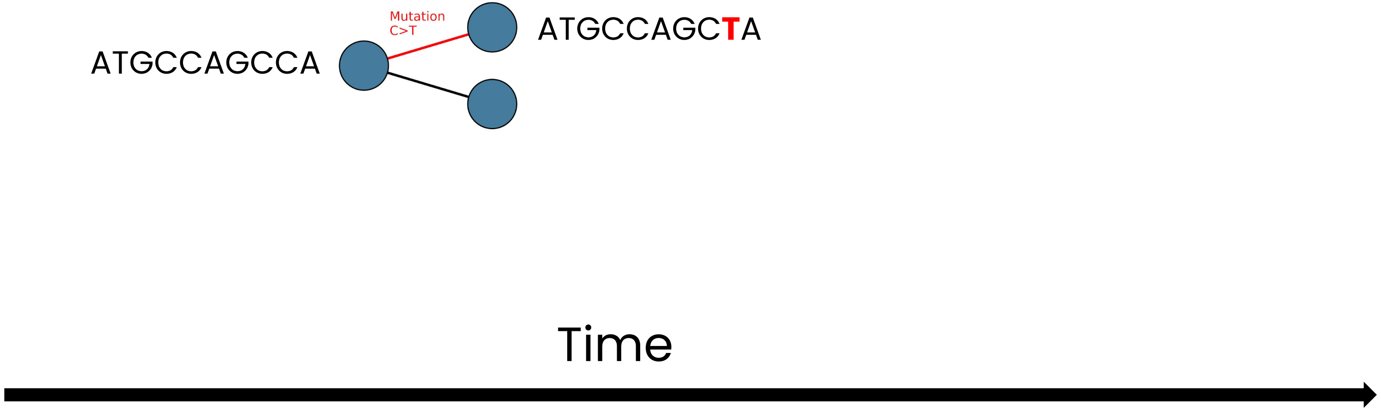
Mutations occur randomly and naturally

- Some mutations are transmitted, others are not



Pathogen genome evolution example

Tracking mutations during pathogen spread



ATGCCAGCCA

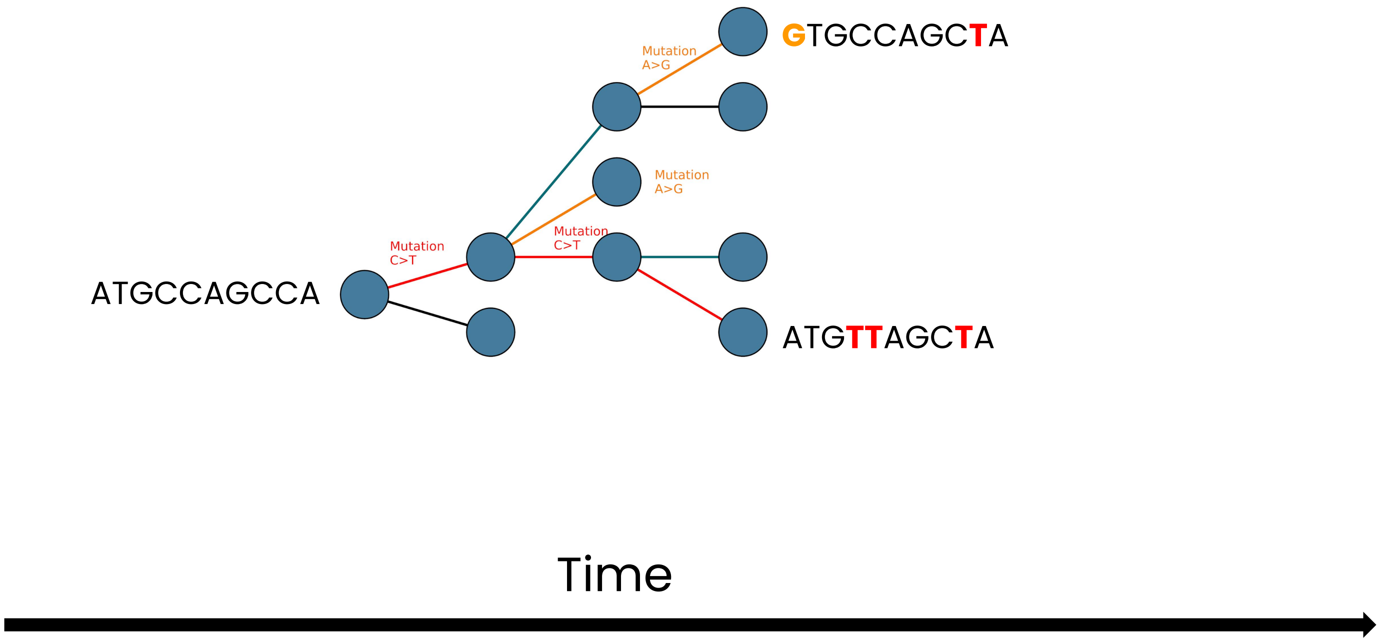
Mutation
C>T

ATGCCAGC**T**A

Time

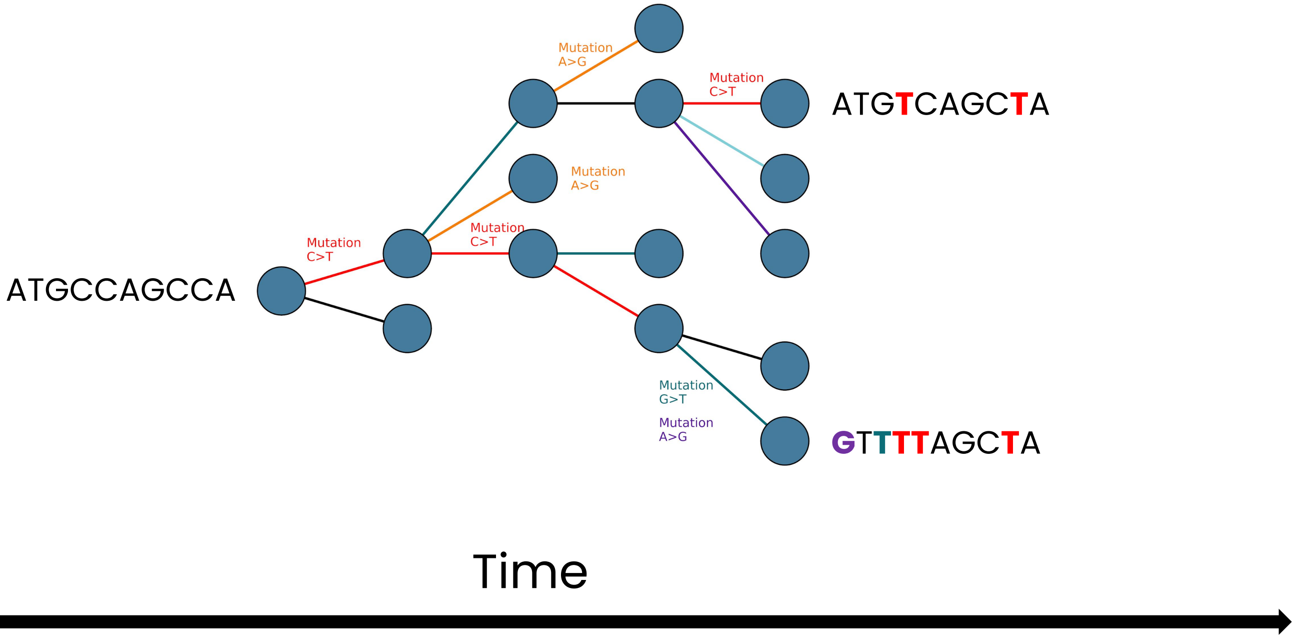
Pathogen genome evolution example

Tracking mutations during pathogen spread



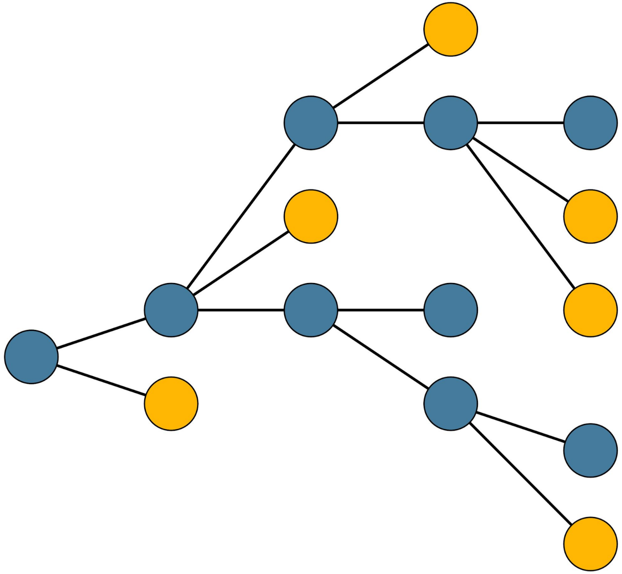
Pathogen genome evolution example

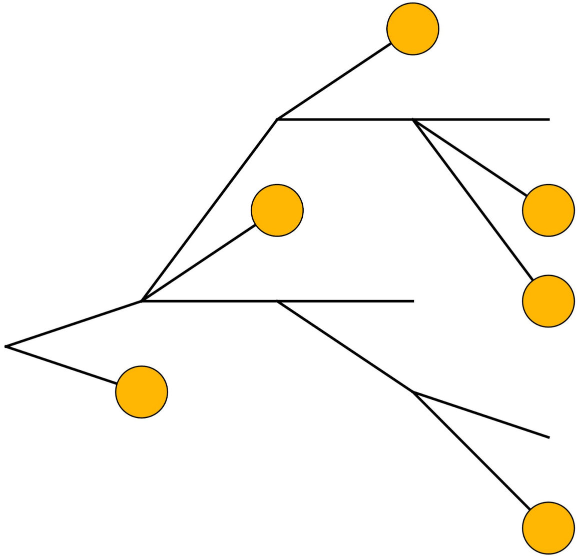
Tracking mutations during pathogen spread



Pathogen genome evolution example

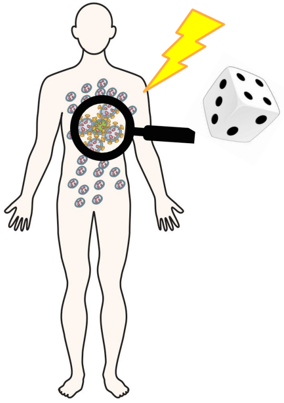
Only some cases in a transmission chain are sequenced





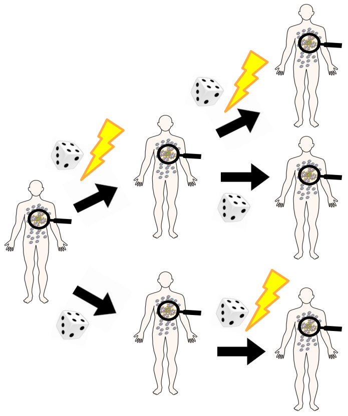
Evolutionary processes play out across multiple scales

Within hosts



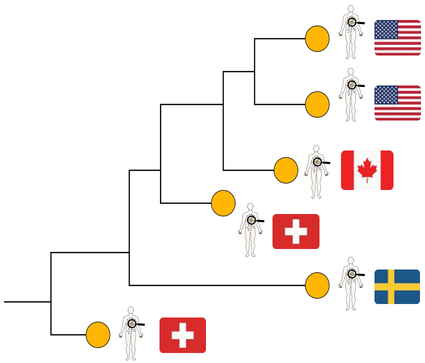
Short term evolution

Between hosts



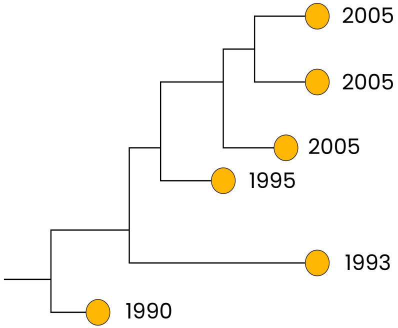
Longer term evolution

Across space



Time scale?

Across time

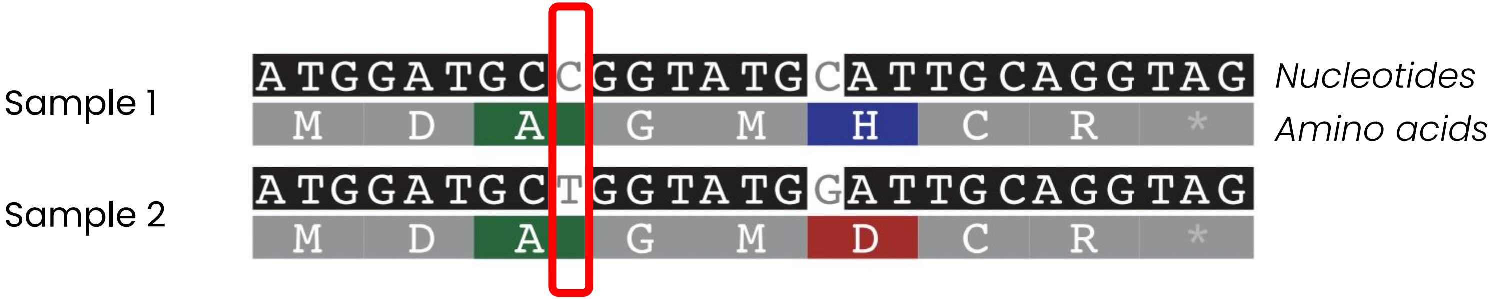


Time scale?

Types of mutations

1. Synonymous (‘S’) mutations

- No change on protein level



Sample 1

ATGGATGCCGGTATGCATTGCAGGTAG *Nucleotides*

M D A G M H C R * *Amino acids*

Sample 2

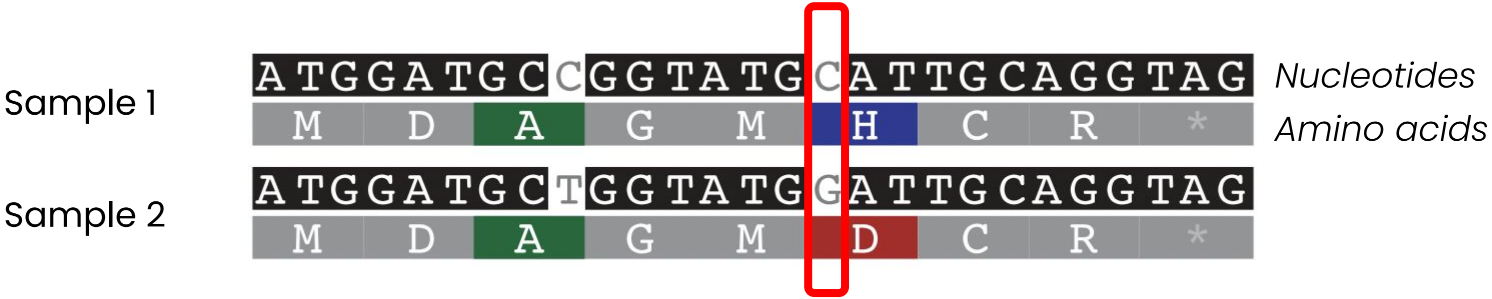
ATGGATGCTGGTATGGATTGCAGGTAG

M D A G M D C R *

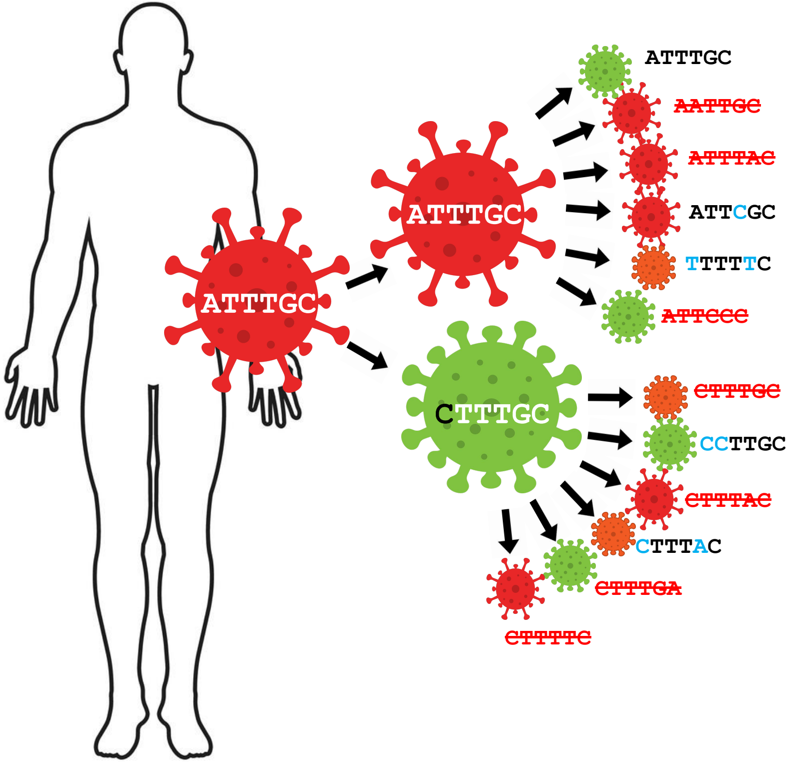
Types of mutations

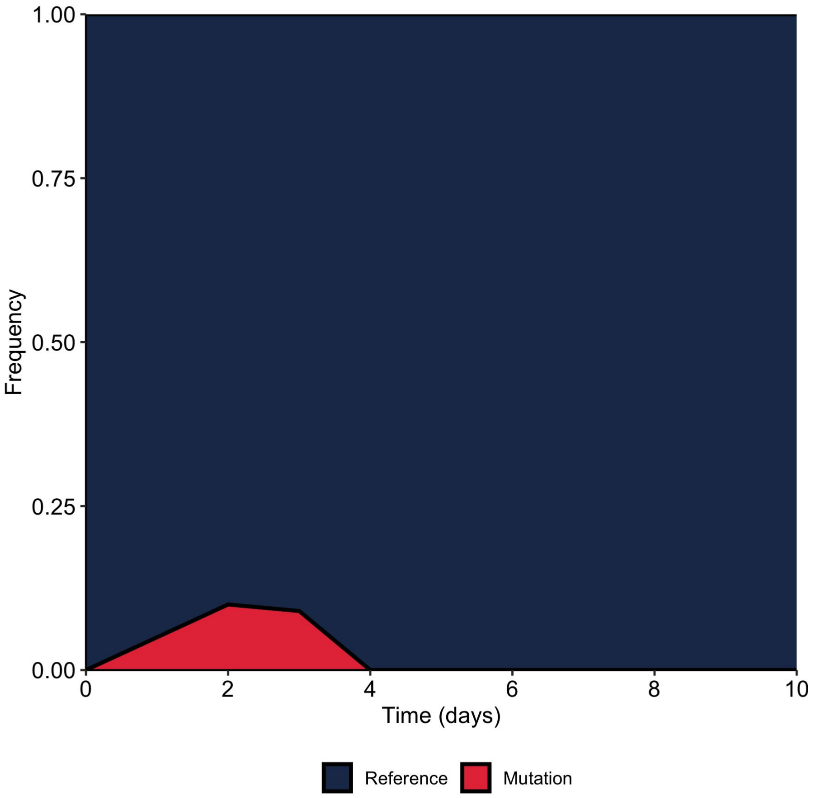
2. Non-synonymous (‘NS’) mutations

- Change on protein level

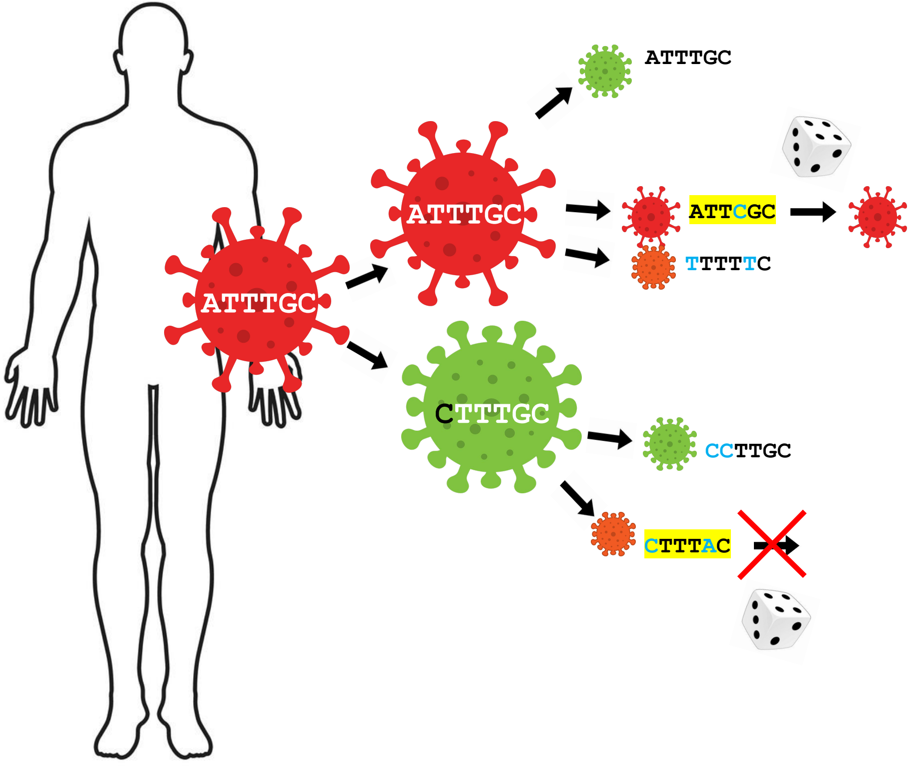


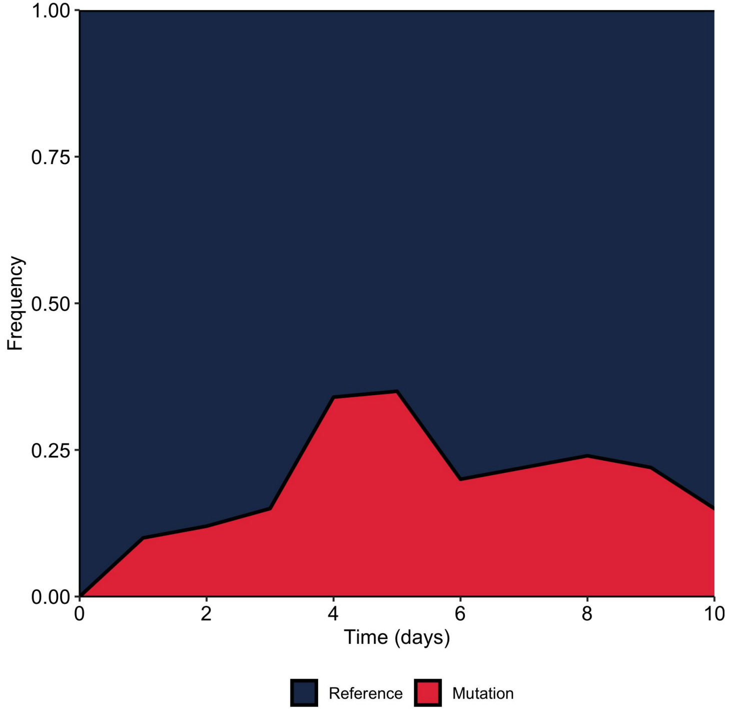
Most mutations are deleterious



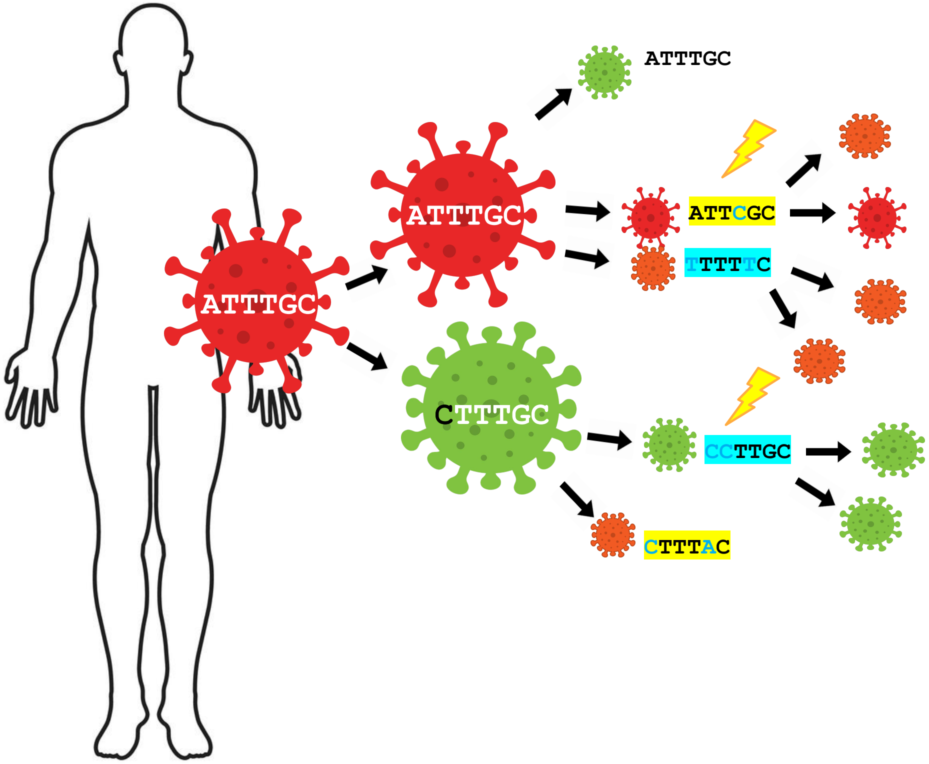


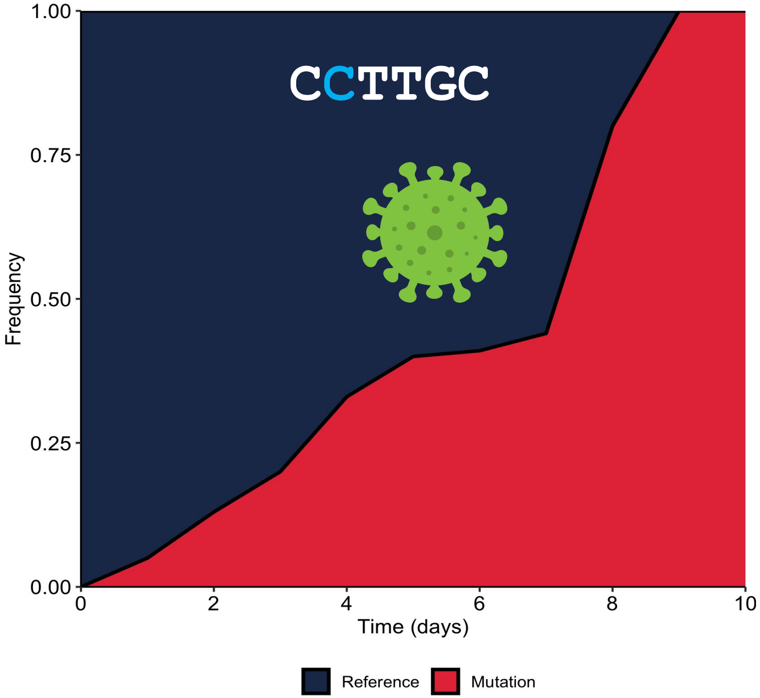
Some mutations are neutral





Some mutations increase viral fitness





What do we mean by fitness advantage?

- Transmission?
- Replication?
- Immune evasion?
- Survival in water?

.....etc

Introduction to Phylogenetics

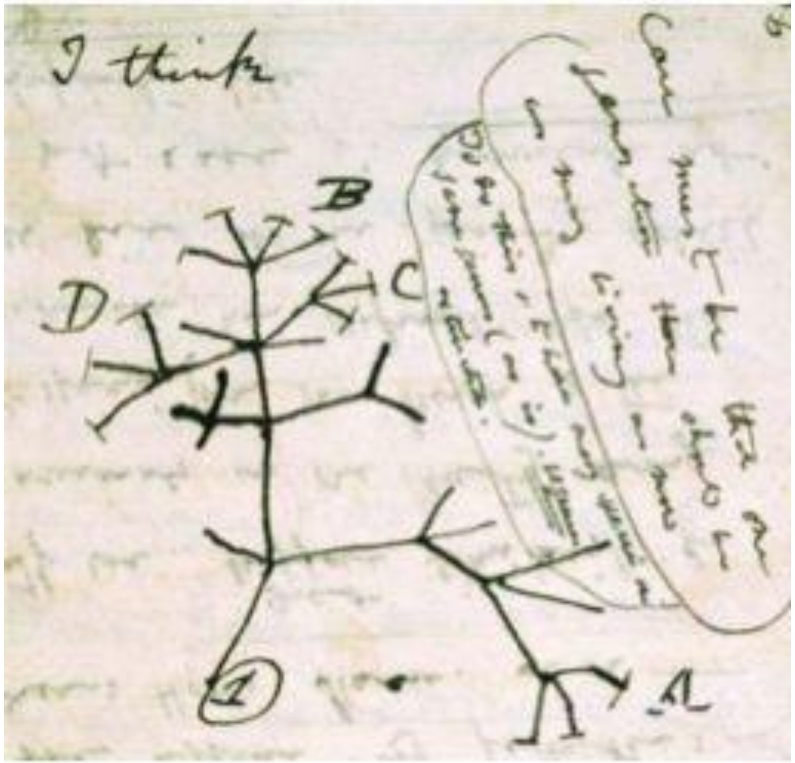
What is phylogenetics?

- Phylogenetics is the process of estimating evolutionary relationships among organisms (pathogens) by analyzing their genetic sequences.
  - Key Principle: Organisms with more similar sequences are more closely related, meaning they share a more recent common ancestor than those with more divergent sequences.
  - Genomic Distance: The number of differences between sequences.

Phylogenetics and Genomic Epidemiology

- Genomic Epidemiology leverages these phylogenetic principles to track transmission pathways.
  - Transmission Relationships: Genomic distance is used as a proxy for determining how pathogens spread from one host to another.
  - Closer Genomic Distance = More recent transmission or shared outbreak source.
  - Divergent Genomic Distance = Suggests different transmission chains or independent introductions.

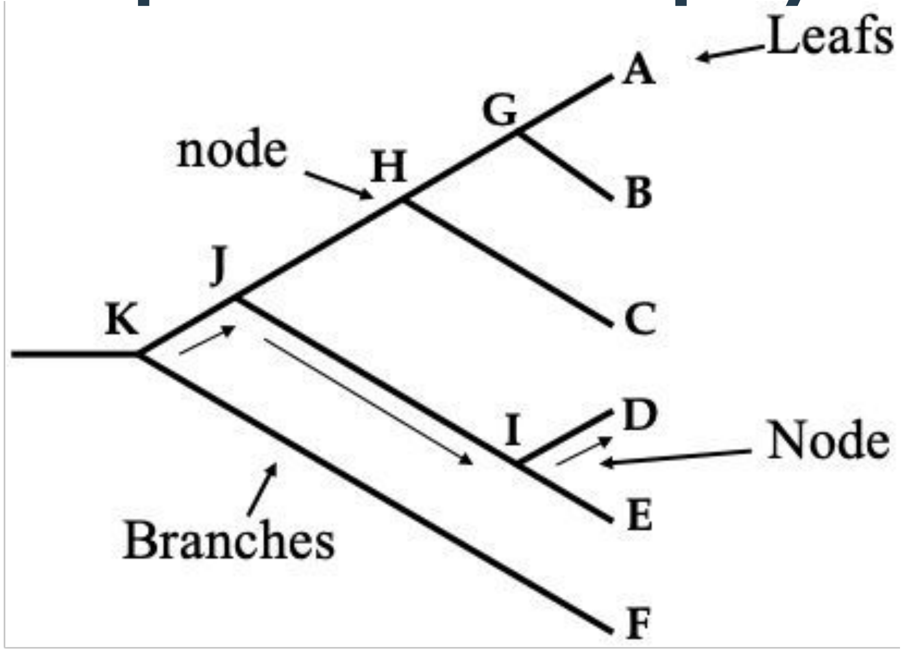
What is a phylogenetic tree?



Darwin, 1837

A diagram used for depicting evolutionary relationships from common ancestors

Components of a phylogenetic Tree



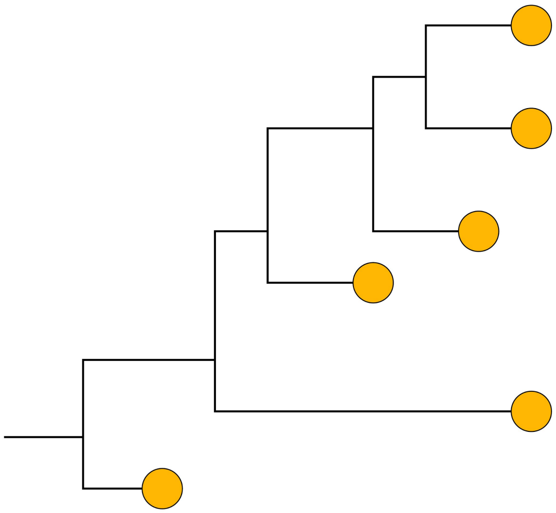
Ancestors
Descendants

- **Leafs/Taxa/Tips:**
  A, B, C, D, E, F
- **Nodes (External/Internal):**
  G, H, I, J, K
- **Branches**

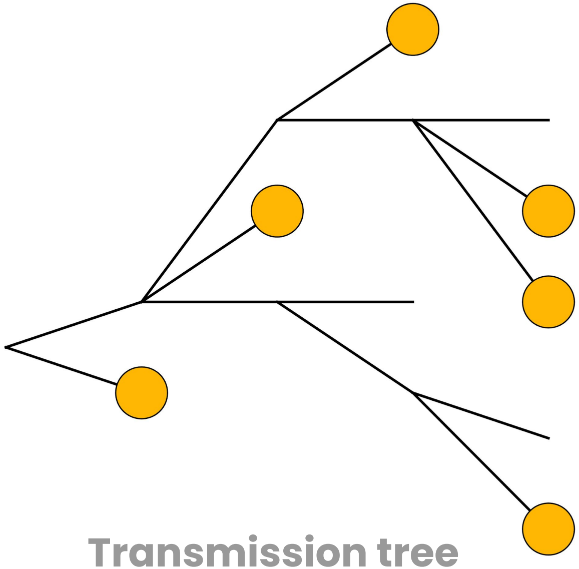
http://dx.doi.org/10.13140/RG.2.1.3433.1049

Phylogenetic trees are not transmission trees

Phylogenies reconstruct evolutionary relationships among *sampled genomes*

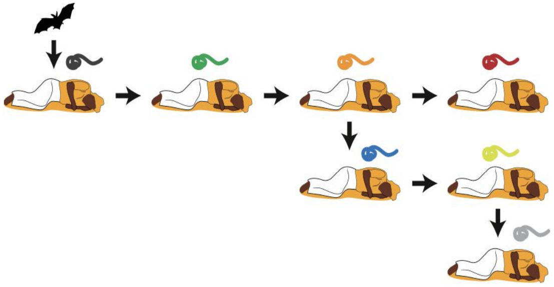


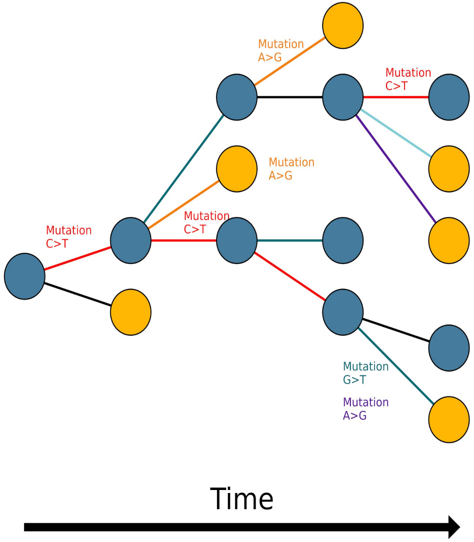
Phylogenetic tree

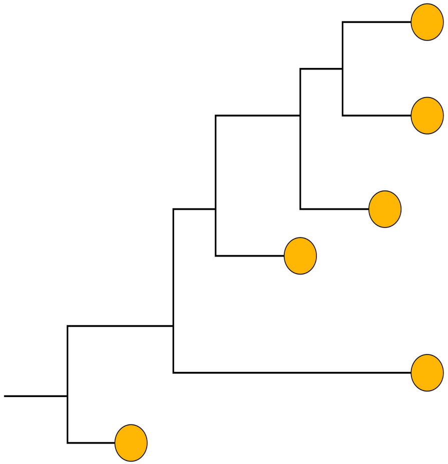


Transmission tree

What can phylogenetic trees tell us about transmission?



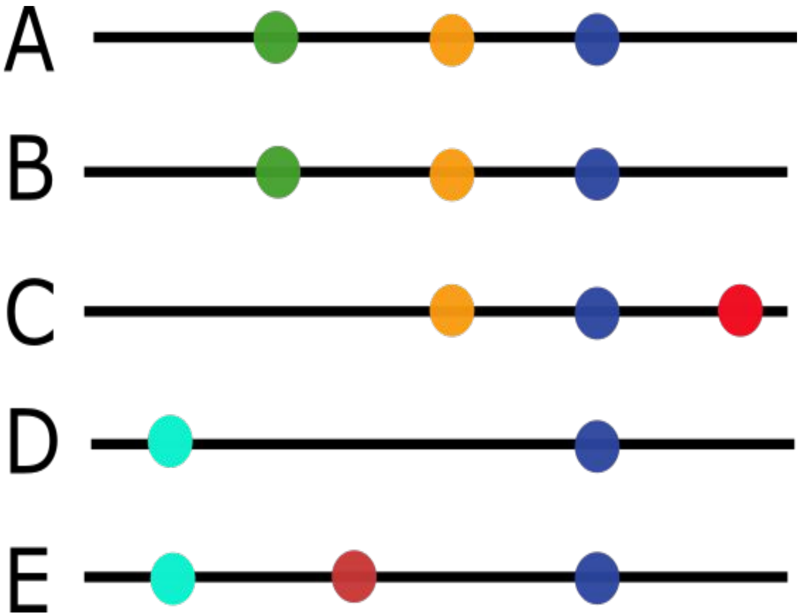




How phylogenies are built

Built from DNA, RNA or protein sequences

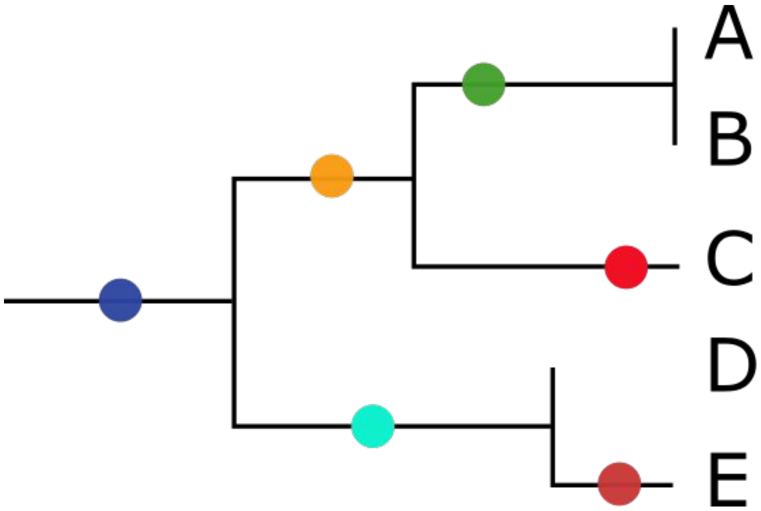
How phylogenies are built



1. Align genes/genomes and identify SNPs

1. Possibly mask out regions of the genome

How phylogenies are built

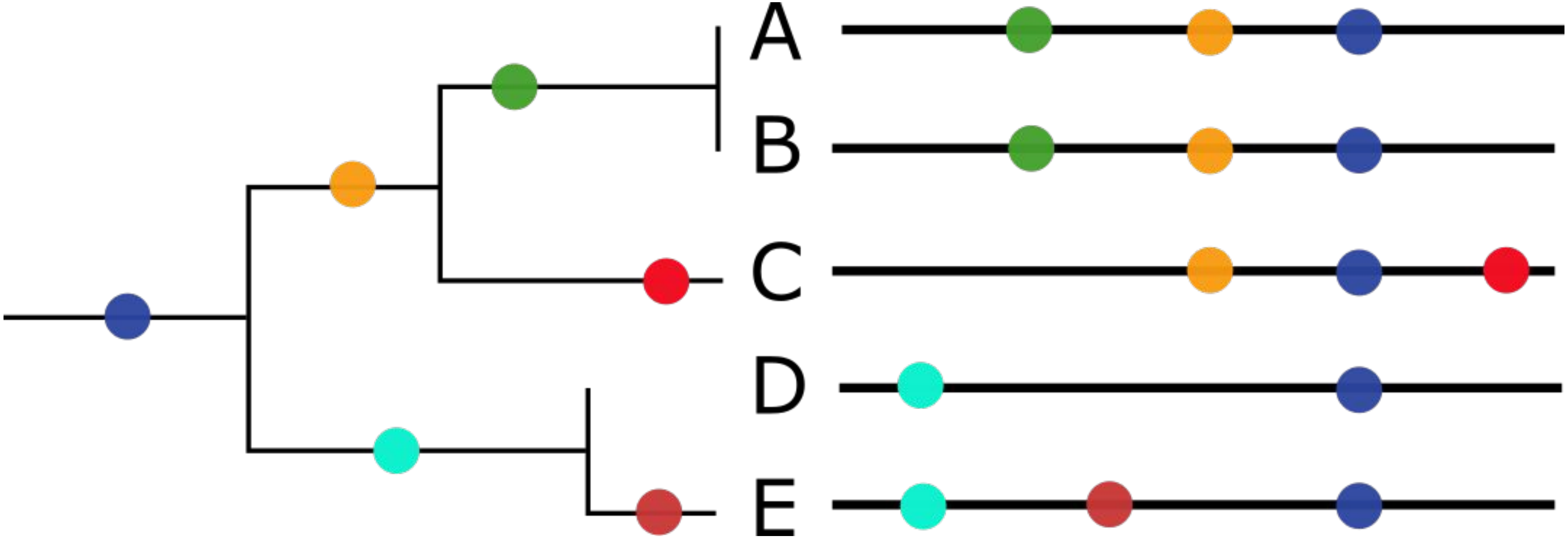


3. Infer the tree from the alignment

How phylogenies are built

Multiple Methods Available

- Best method dependent on data and purpose
- May be trade-offs between speed and accuracy
- Results highly dependent on quality of data being analyzed


A
B
C
D
E

Choosing samples to include in a phylogenetic tree

Phylogenetic Sample selection is critical

- Critical to have **high quality samples that pass QC thresholds**
- Need samples that can answer the question you want to address

What epidemiological question you are trying to answer?

- Do these sequences represent an outbreak?
  - Include all high-quality sequences of a specific lineage or clade

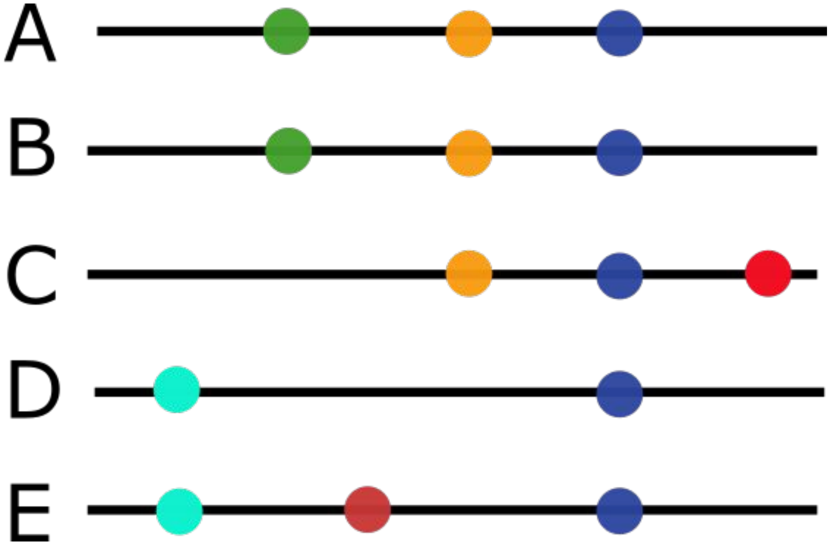
- Are these sequences part of an ongoing outbreak?
  - Include all high-quality sequences of the known genotype, plus some sequences known to be part of the outbreak

- Does this sequence belong to a novel genotype?
  - Include high-quality sequences from multiple genotypes

- No question, simply descriptive
  - Include all high-quality sequences of the species you are aiming to describe

Creating the multiple sequence alignment

Alignment: Assemblies



Reference genome selection

Reference selection is critical

- If there’s a section of your genome(s) that is not found in the reference, it will not be aligned or used in the phylogeny
- Using a reference genome that is close to your samples will enable you to find more SNPs to build the phylogeny

User-defined region masking

Users may choose to mask repetitive regions

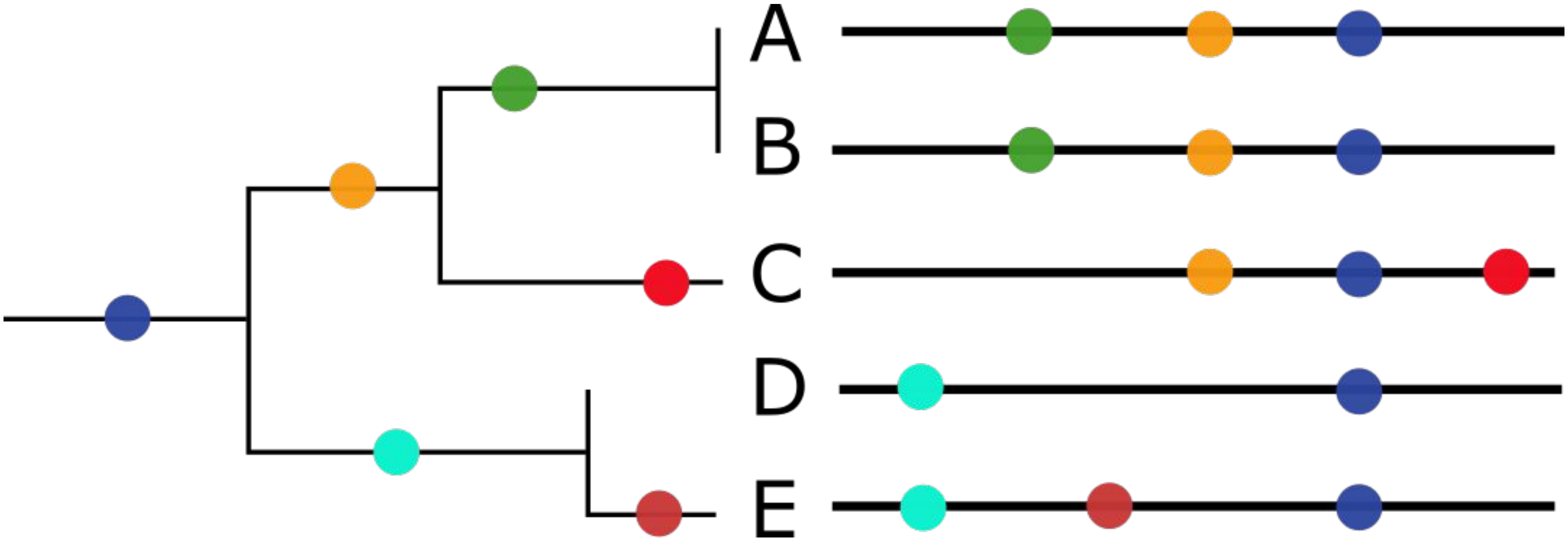
We want to build the phylogeny from positions in the genome that have **ancestral origin**, not:

- Repeat regions (likely to misalign to reference and may introduce erroneous SNPs)
- Recombinant regions (not ancestral)

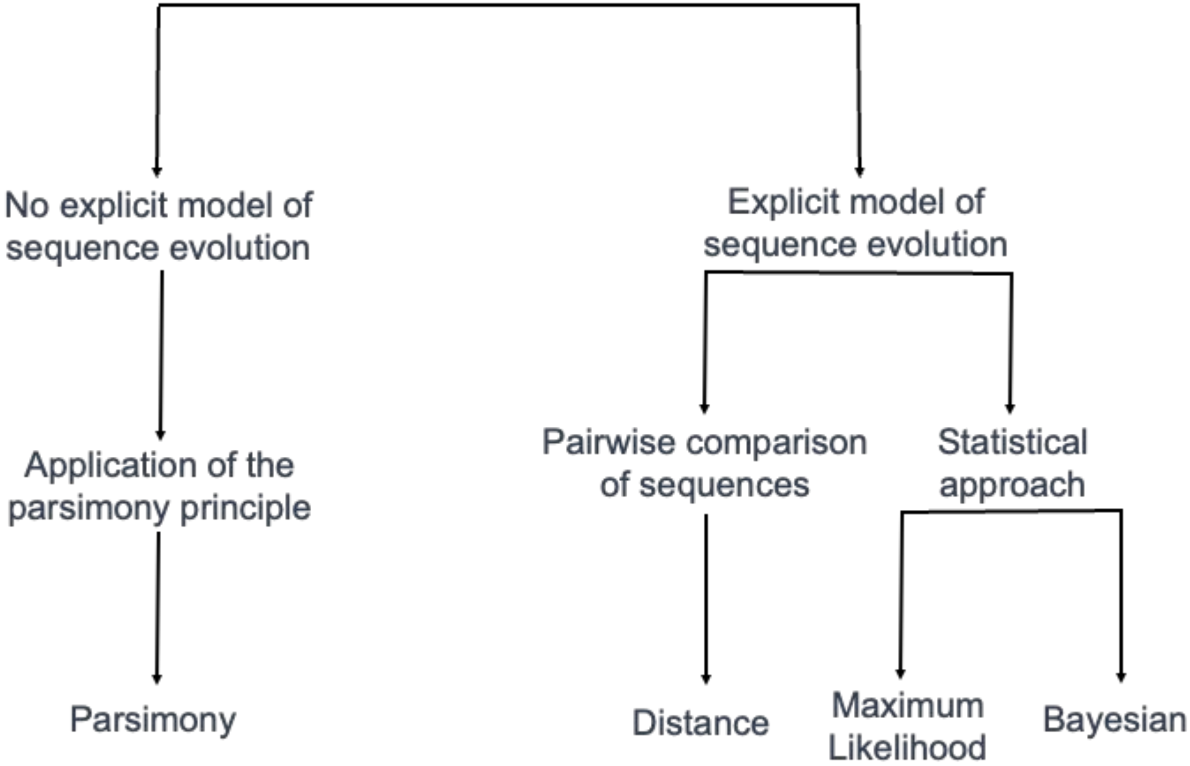
Provide a **bed file to mask regions** that we do not want to include in the MSA to build the phylogeny

Phylogenetic reconstruction methods

To tree, from alignment



Phylogenetic reconstruction methods



No explicit model of sequence evolution

Application of the parsimony principle

Parsimony

Explicit model of sequence evolution

Pairwise comparison of sequences

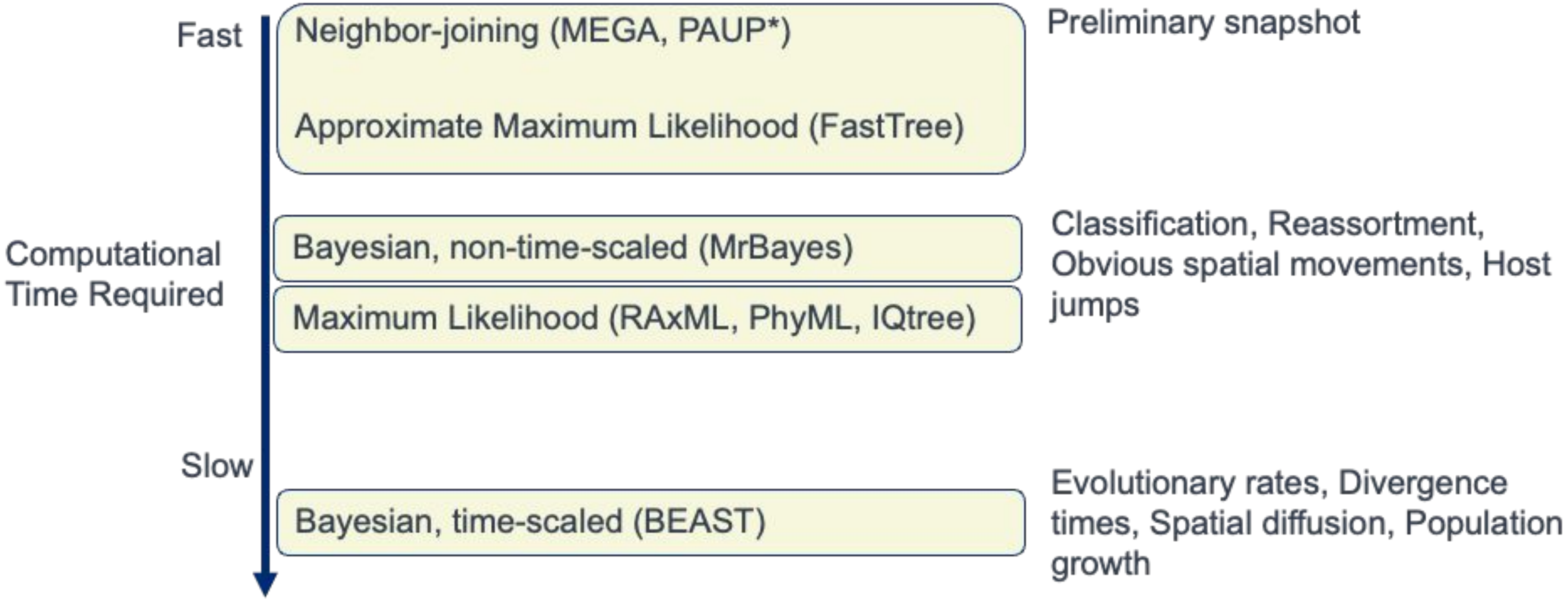
Distance

Statistical approach

Maximum Likelihood

Bayesian

Choosing the right software for the right analysis



Fast

Computational Time Required

Slow

Neighbor-joining (MEGA, PAUP*)

Approximate Maximum Likelihood (FastTree)

Preliminary snapshot

Bayesian, non-time-scaled (MrBayes)

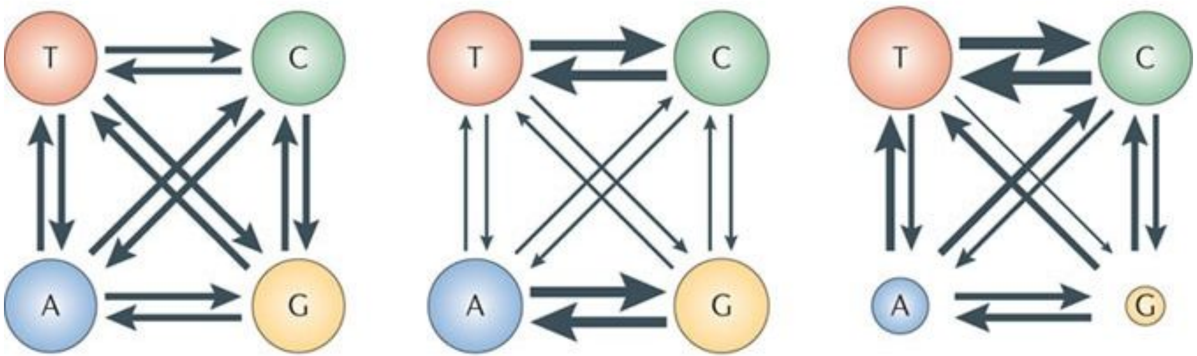
Maximum Likelihood (RAxML, PhyML, IQtree)

Classification, Reassortment, Obvious spatial movements, Host jumps

Bayesian, time-scaled (BEAST)

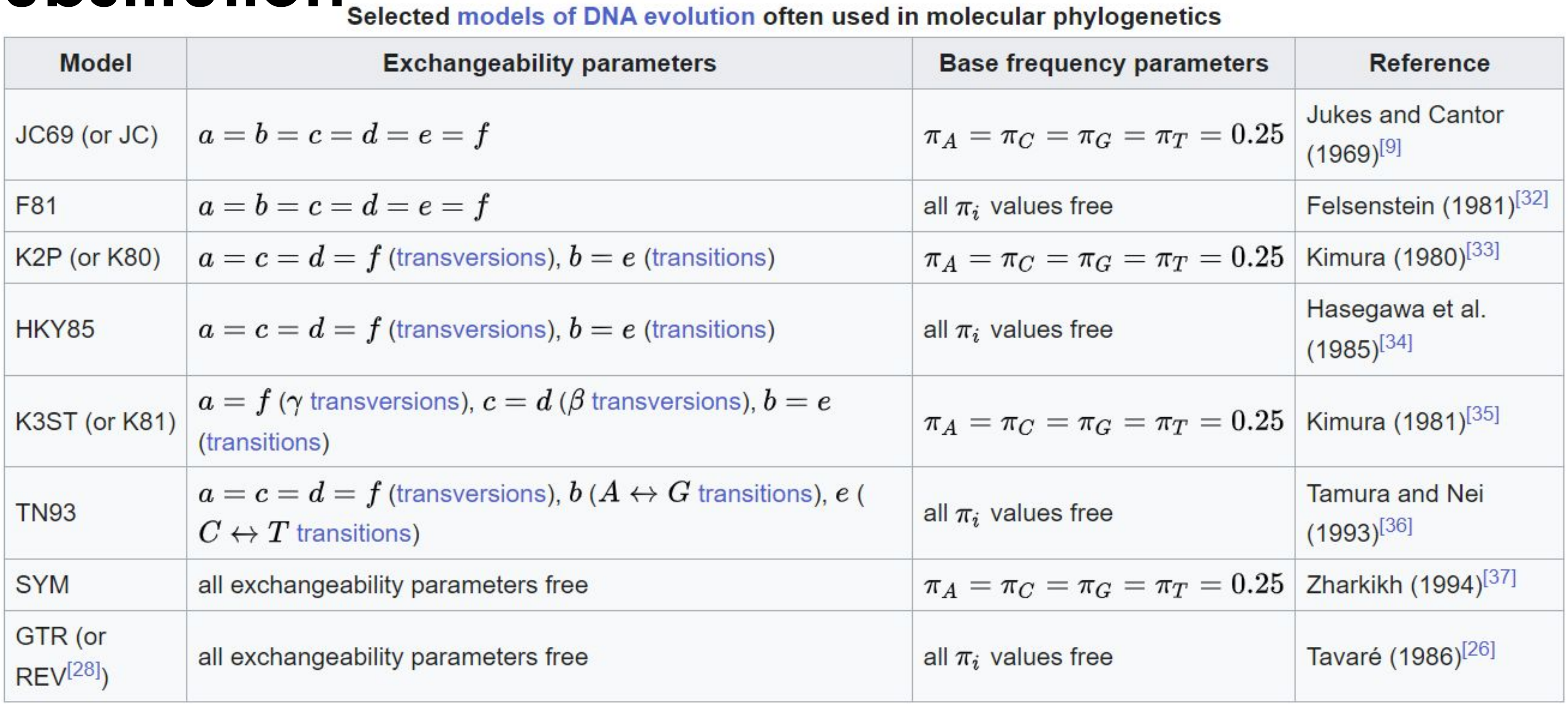
Evolutionary rates, Divergence times, Spatial diffusion, Population growth

Phylogenetic Tree Building Models



- Models describe the process of evolution
- Allows the correction of genetic distances for multiple substitutions (unseen evolution)
- Relative frequencies of the four nucleotides, e.g. A=26%, G=24%, etc
- Relative rates of pairwise nucleotide substitutions, e.g. rate of A to C mutation

Choosing the right model of nucleotide substitution



Selected models of DNA evolution often used in molecular phylogenetics

Model	Exchangeability parameters	Base frequency parameters	Reference
JC69 (or JC)	$a = b = c = d = e = f$	$\pi_A = \pi_C = \pi_G = \pi_T = 0.25$	Jukes and Cantor (1969)[9]
F81	$a = b = c = d = e = f$	all π_i values free	Felsenstein (1981)[32]
K2P (or K80)	$a = c = d = f$ (transversions), $b = e$ (transitions)	$\pi_A = \pi_C = \pi_G = \pi_T = 0.25$	Kimura (1980)[33]
HKY85	$a = c = d = f$ (transversions), $b = e$ (transitions)	all π_i values free	Hasegawa et al. (1985)[34]
K3ST (or K81)	$a = f$ (γ transversions), $c = d$ (β transversions), $b = e$ (transitions)	$\pi_A = \pi_C = \pi_G = \pi_T = 0.25$	Kimura (1981)[35]
TN93	$a = c = d = f$ (transversions), b ($A \leftrightarrow G$ transitions), e ($C \leftrightarrow T$ transitions)	all π_i values free	Tamura and Nei (1993)[36]
SYM	all exchangeability parameters free	$\pi_A = \pi_C = \pi_G = \pi_T = 0.25$	Zharkikh (1994)[37]
GTR (or REV[28])	all exchangeability parameters free	all π_i values free	Tavaré (1986)[26]

https://en.wikipedia.org/wiki/Substitution_model

Phylogenetic reconstruction tools

Software for Inferring Phylogenetic Trees

- **Parsimony (PAUP*, Mesquite, PHYLIP)**
  Find tree with the minimum number of mutations between sequences (i.e. choose tree with the least convergent evolution)

- **Neighbor-Joining (PAUP*, MEGA, PHYLIP, ClustalW, BioNJ)**
  Estimate genetic distances between sequences and cluster these distances into a tree that minimizes genetic distance over the whole tree

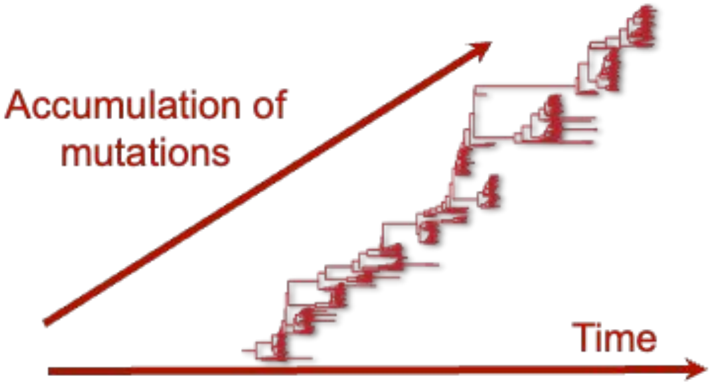
- **Maximum Likelihood (PAUP*, GARLi, PhyML, MEGA, TREE-PUZZLE, PAML, ExaML, RaxML, FastTree, IQ-Tree)**
  Determine the probability of a tree (and branch lengths) given a particular model of molecular evolution and the observed sequence data

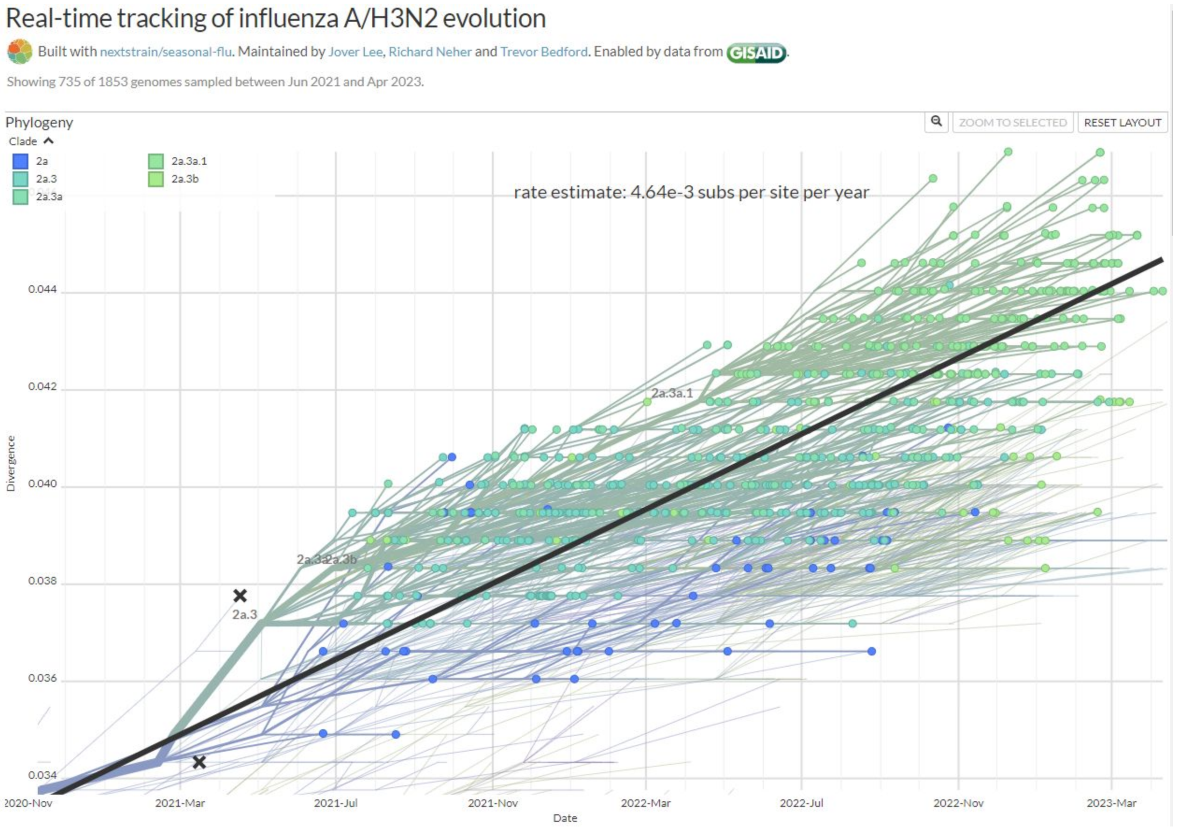
- **Bayesian (BEAST, BEAST2, Mr.Bayes, PhyloBayes, RevBayes)**
  Similar to likelihood but where there is information about the prior distribution of parameters. Also returns a (posterior) distribution of trees

Temporal Signal

Temporal Signal

- Pathogen genomes are usually sampled at different points in time (heterochronous sequences)
- Transmission history is estimated on a real time-scale (e.g. days, months or years)
- Before building a time-scaled phylogenetic tree from heterochronous sequences, confirm that the sequences contain sufficient '**temporal signal**' or '**clockiness**' for reliable estimation.
- In other words, there must be **sufficient genetic change between sampling times** to reconstruct a statistical relationship between genetic divergence and time.





Real-time tracking of influenza A/H3N2 evolution

Built with nextstrain/seasonal-flu. Maintained by Jover Lee, Richard Neher and Trevor Bedford. Enabled by data from GISAID.

Showing 735 of 1853 genomes sampled between Jun 2021 and Apr 2023.

- **Linear trend:** evolution will be adequately represented by a *strict molecular clock*. A linear trend with greater scatter from the regression line suggests a *relaxed molecular clock* model may be most appropriate.
- **Non-linear trend**: evolutionary rate has systematically changed through time.
- **No trend at all:** data contains little temporal signal and is unsuitable for inference using phylogenetic molecular clock models.

Rooting a Phylogenetic Tree

Rooting a phylogenetic tree

Phylogenetic trees are either rooted or unrooted.

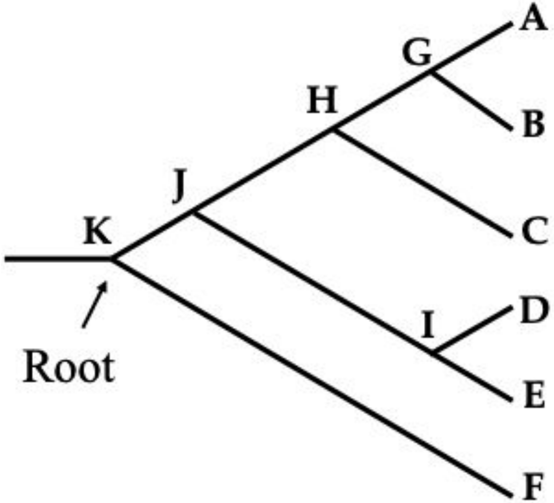
The root gives directionality to evolution within the tree

To root an unrooted tree:

- **root by outgroup**, e.g. use F as an outgroup
- **midpoint rooting** – the midpoint of the path joining the two most dissimilar taxa

Outgroup should:

- Not belong to ingroup, thus branched off before ingroup (e.g. judged using a priori biological/paleontological information)
- Be the most distantly related of the taxa
- Not be too distantly related to ingroup
- Be homologous to ingroup

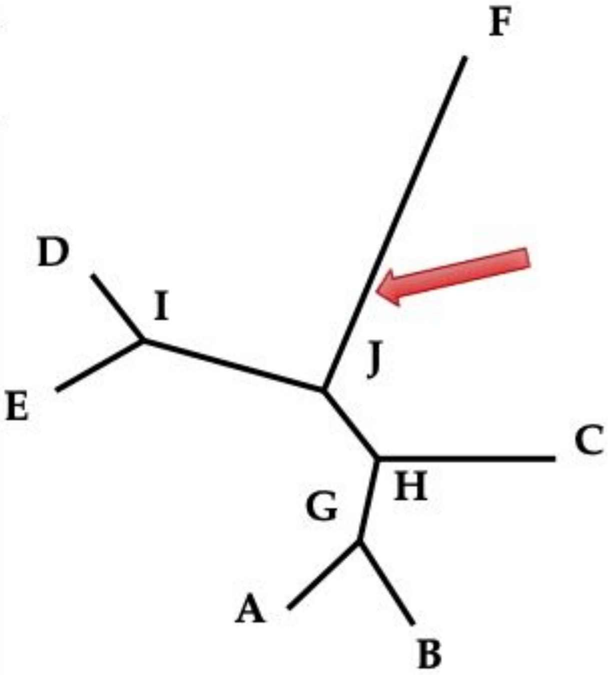


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7149615/

Unrooted Phylogenetic Tree

- Root node K disappeared

- Unrooted tree is focused **only on relationships** among the taxa rather than on the directionality of evolutionary change.



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7149615/

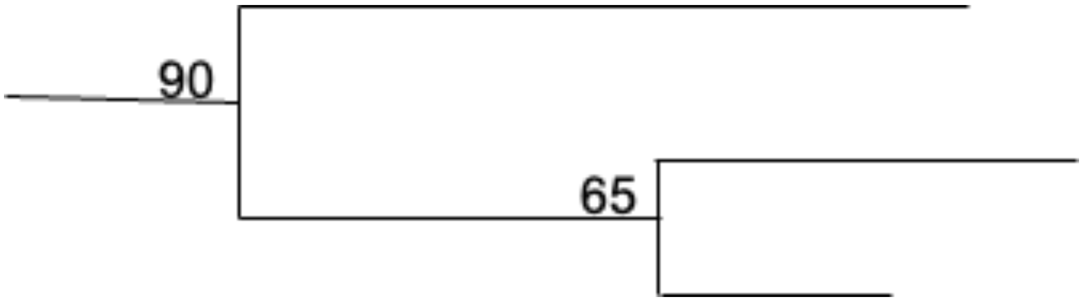
Bootstrapping

How Robust is the Tree?

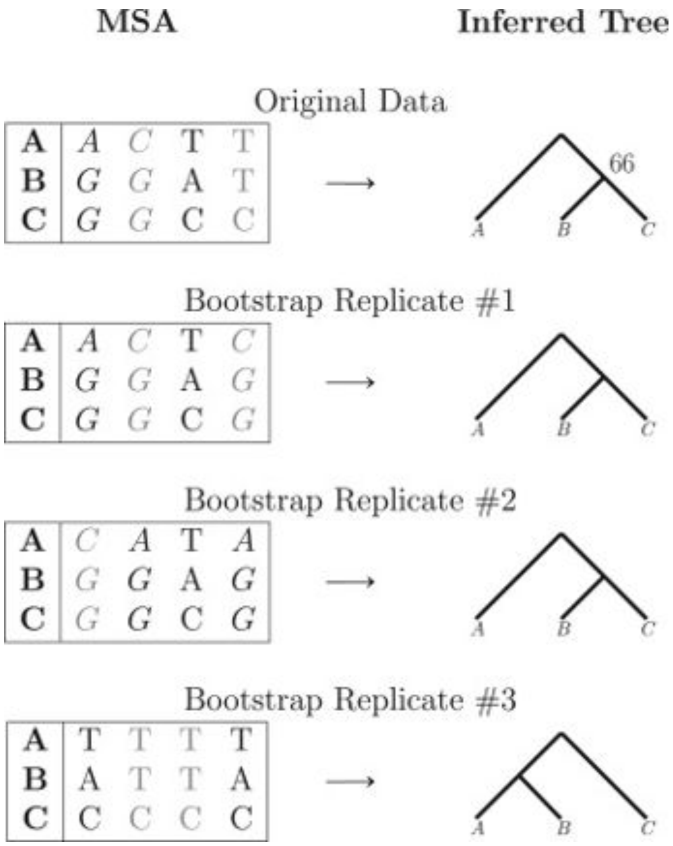
- **Bootstrapping** is a statistical technique that uses **random resampling** of data to determine sampling error.

- Gives an idea about the **‘reliability’** of branches and clusters.

- Usually considered significant is higher than 70% (or 0.7 or 70/100)



Bootstraps



Original Data

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

Bootstrap Replicate #1

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

Bootstrap Replicate #2

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

Bootstrap Replicate #3

A	T	T	T	T
B	A	T	T	A
C	C	C	C	C

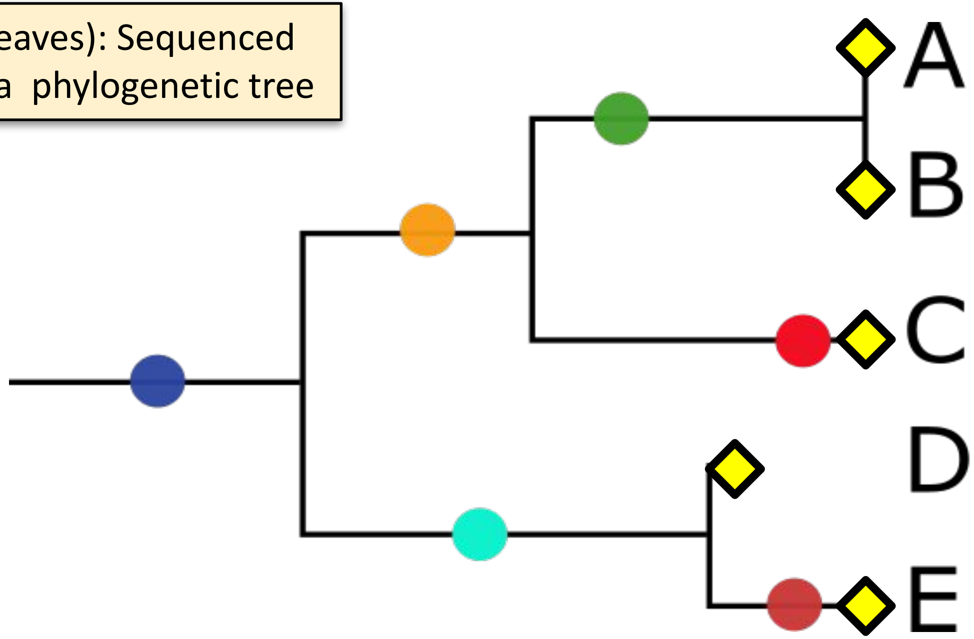
- Characters are **resampled with replacement to create many replicate data sets.** A tree is then inferred from each replicate.

- Agreement among the resulting trees is **summarized with a consensus tree.** The frequencies of occurrence of groups, **bootstrap** proportions (BPs), are a measure of support for those groups

Interpreting phylogenetic trees

Anatomy of a Phylogenetic Tree

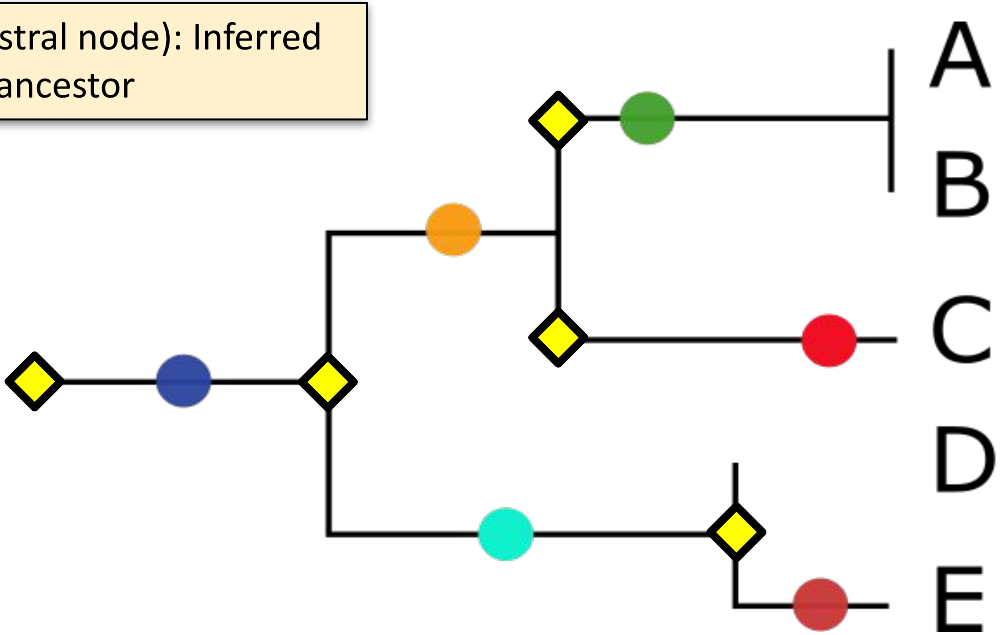
Terminal Nodes (tips/leaves): Sequenced sample represented on a phylogenetic tree



These are all of the **terminal nodes** on this phylogenetic tree

Anatomy of a Phylogenetic Tree

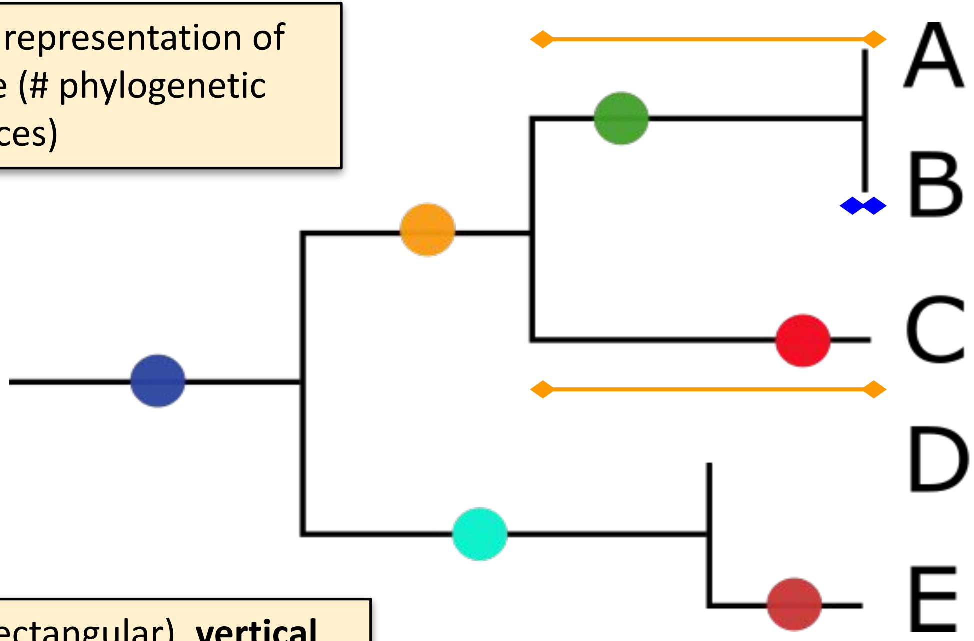
<u>Internal Nodes</u> (ancestral node): Inferred common ancestor



These are all of the **internal nodes** on this phylogenetic tree

Anatomy of a Phylogenetic Tree

<u>Branch Length</u>: Visual representation of evolutionary distance (# phylogenetic differences)



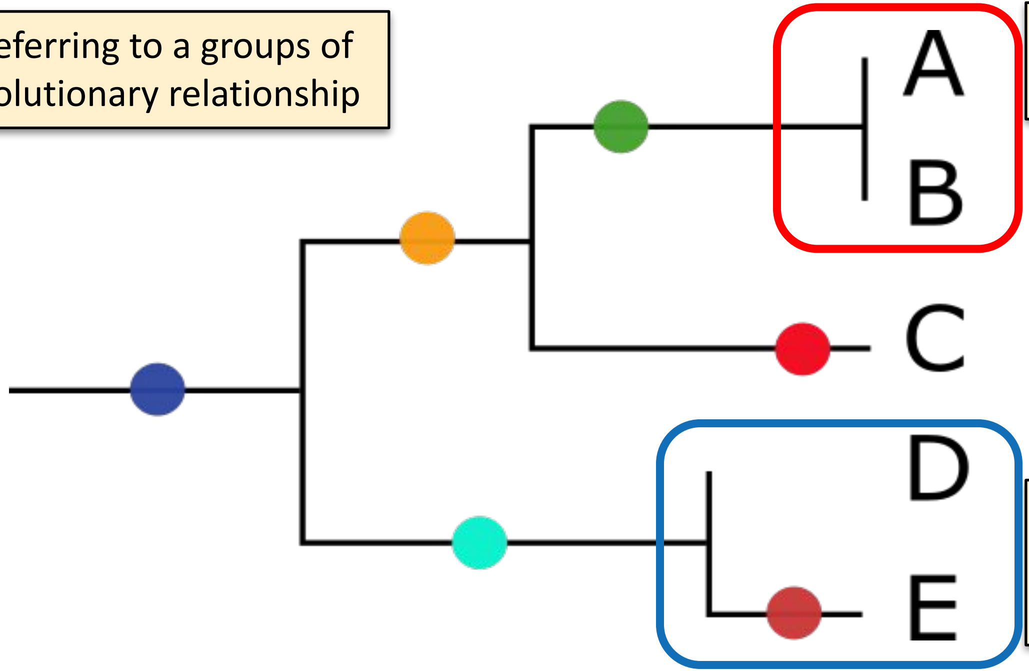
Branch length between A and B is zero, indication phylogenetic identity*

Branch length between A and C is larger, indicating a distant evolutionary relationship

Note: In this format (rectangular), **vertical branch length** serve strictly **as a visual aid** to enhance clarity and organization

Anatomy of a Phylogenetic Tree

Cluster: General term referring to a groups of samples with a close evolutionary relationship

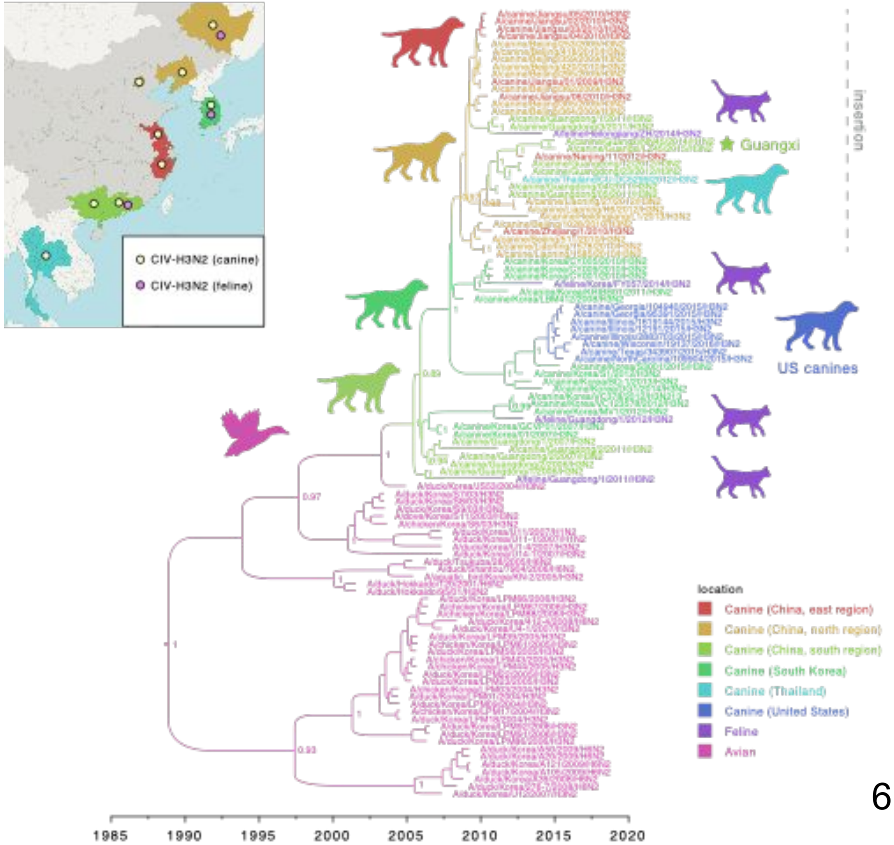


Samples A and B make up a cluster of samples

Samples D and E make up another cluster of samples

What can be inferred from a tree?

- Host switches
- Spatial movements
- Transmission
- When a novel lineage emerged
- When novel mutations occurred



Common Phylogenetic Tree Building Pitfalls

- Not using enough background data
- Not enough phylogenetic signal in dataset
- Annotation errors on sequence alignments and dates
- Not removing recombinants
- Not removing viruses with sequencing errors/contaminants
- Not questioning odd results
- Over-interpreting gaps in tree

Summary

- Inferring phylogenies enables you to **reconstruct ancestral relationships** and **recover hidden information**
- It is important to use **a good tree-building method**
- **Branch lengths** represent genetic distances between sequences. Can represent time on time trees
- With the right **root**, we can infer ancestor-descendent relationships
- We can test the reliability of an inferred tree using **bootstrapping**
- Building a tree is only half the challenge. The real challenge is **how to interpret the tree**
- New tools are allowing for more sophisticated interpretation of phylogenetic patterns

What questions do you have?