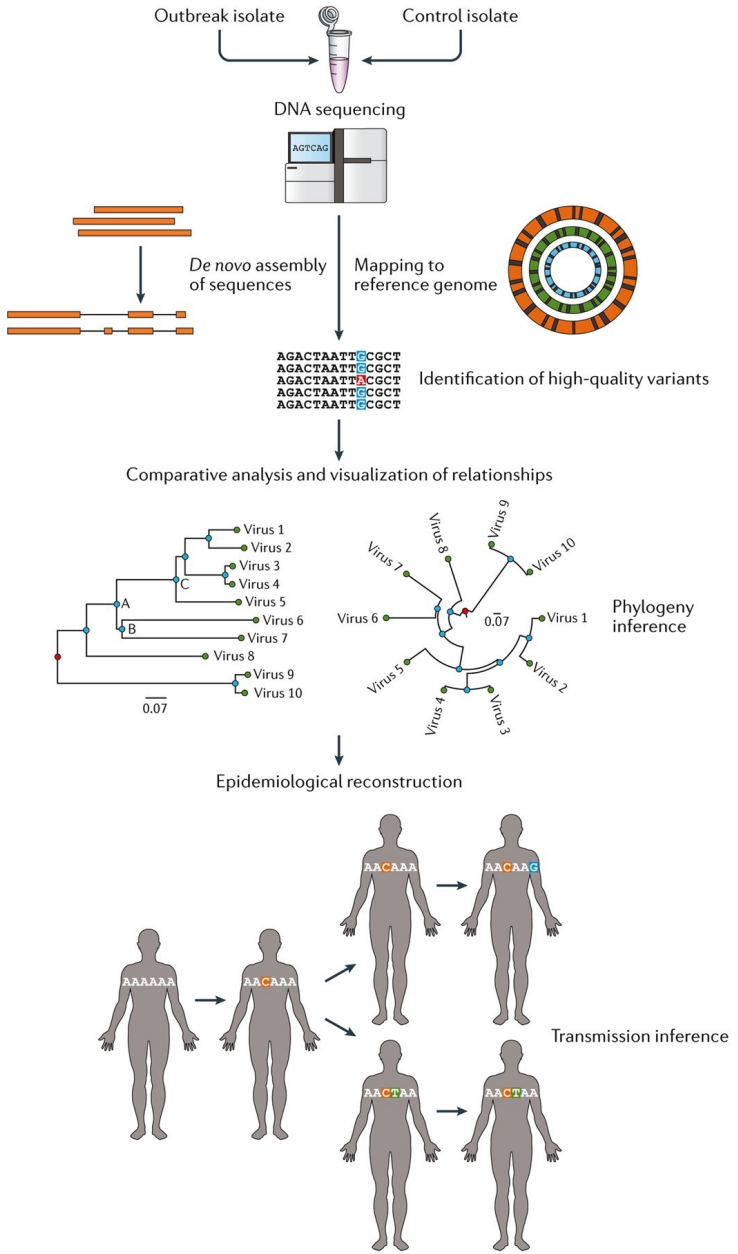
Pathogen genomic epidemiology

Pathogen genomic epidemiology: the use of pathogen genomes to study the spread of infectious diseases through populations.

1. High-throughput ('next-generation') sequencing: Routine and cost-effective near real time generation of full-length genomes.

2. Development of new computational and statistical methods in data analysis.



Gardy, J., Loman, N. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat Rev Genet* **19**, 9–20 (2018). https://doi.org/10.1038/nrg.2017.88

Nature Reviews | Genetics

Objectives

- Basics of phylogenetics: definition, traits, nodes, taxa, branches, lineage, clades, lengths.

- Modeling sequence evolution: turning pairwise sequence alignment to pairwise distance, substitutions models.

- Tree-building algorithms: distance and character-based methods

- Tree evaluation: Bootstrapping and Posterior probability

- Phylodynamics: molecular clock, phylogeography

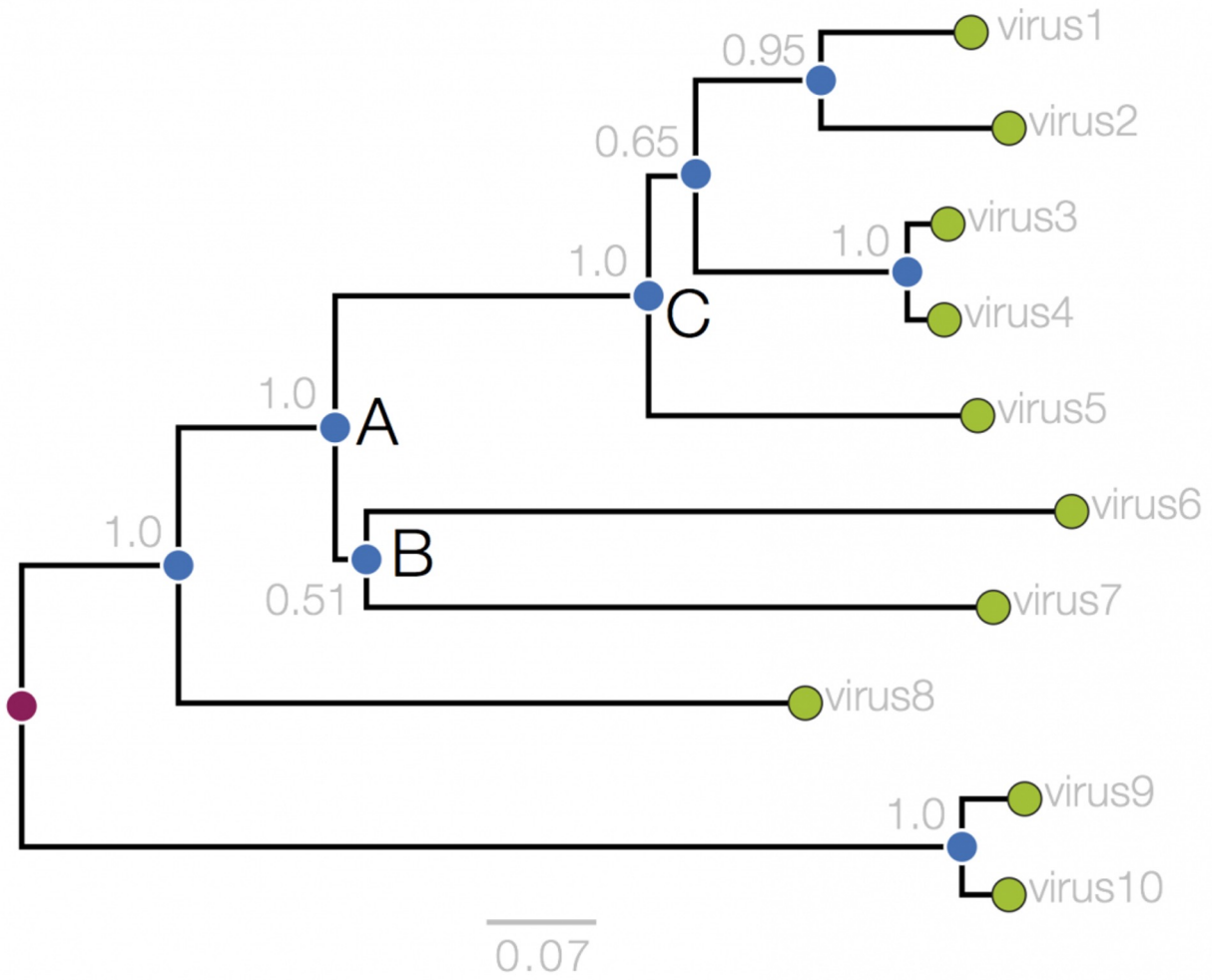
Phylogenetic inference

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).

Phylogenies are useful for:

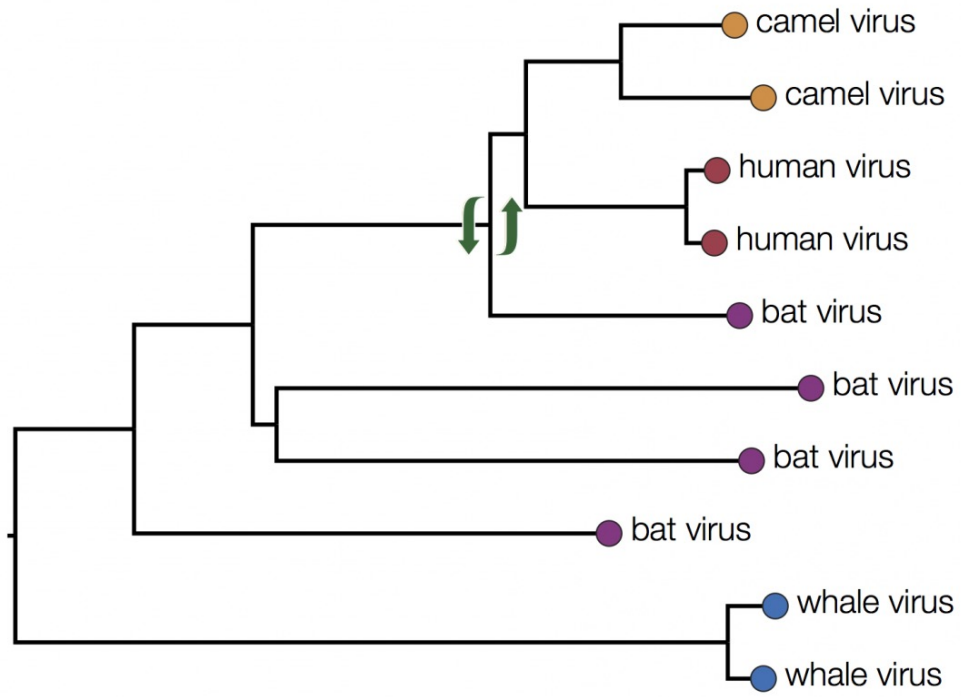
- Organizing knowledge of biological diversity.
- Structuring classifications.
- Providing insight into events that occurred during evolution.

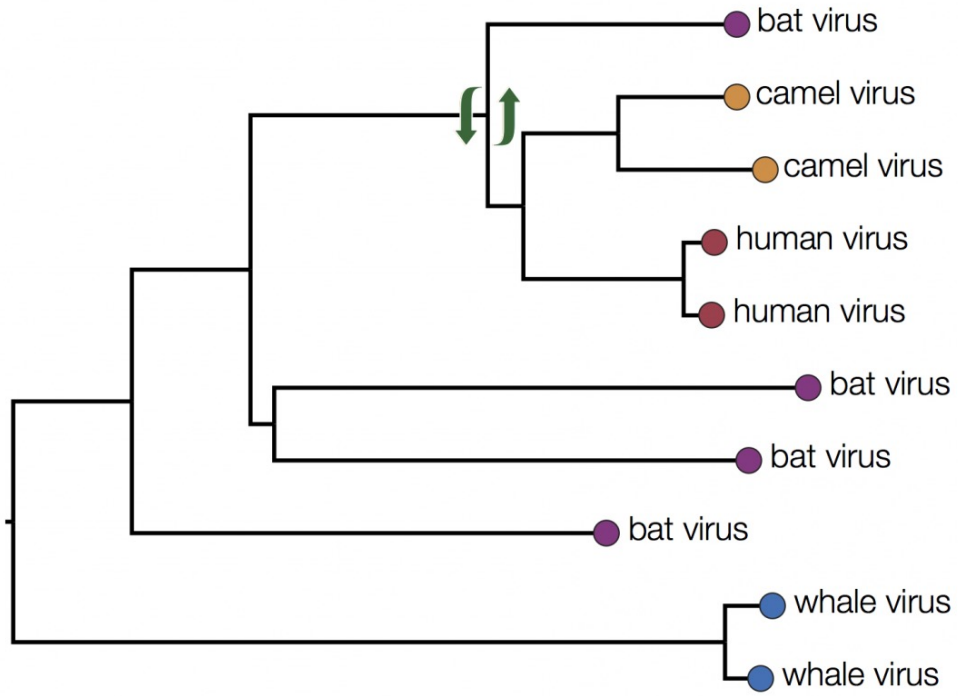
Phylogenetic tree structure



A fictional rooted phylogenetic tree showing the different parts of the tree.
Adapted from https://artic.network

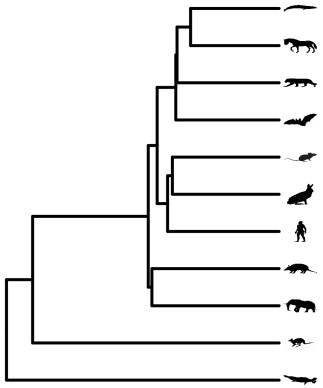
Topological representations



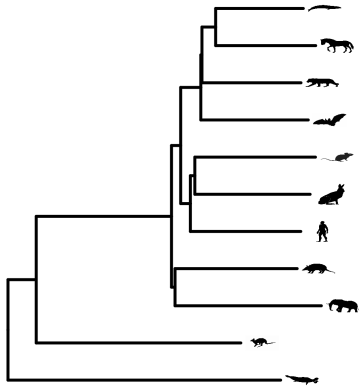


Types of trees

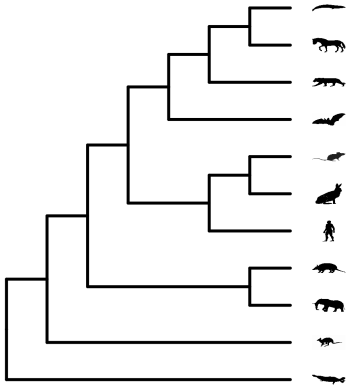
(A) chronogram



(B) phylogram

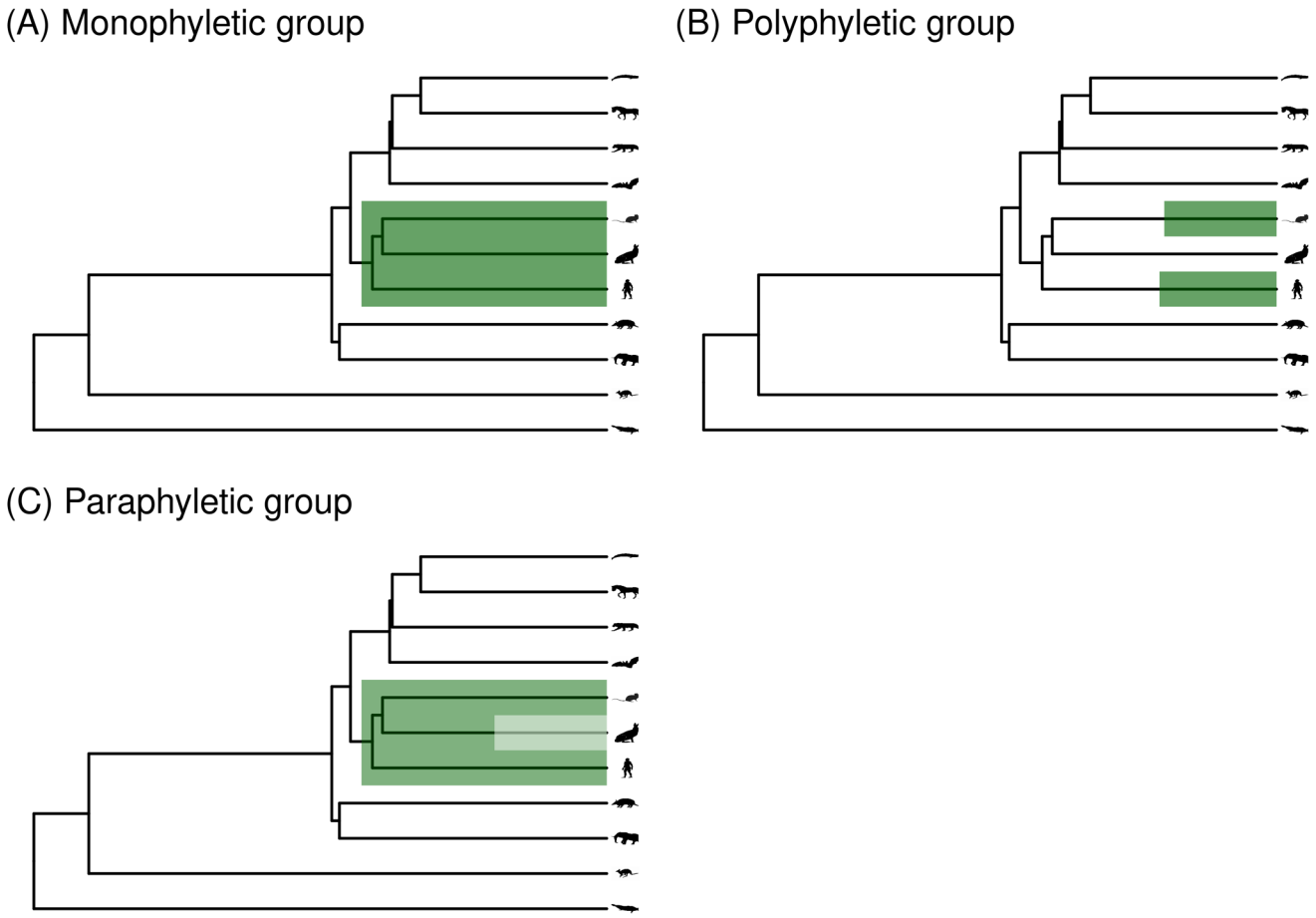


(C) cladogram



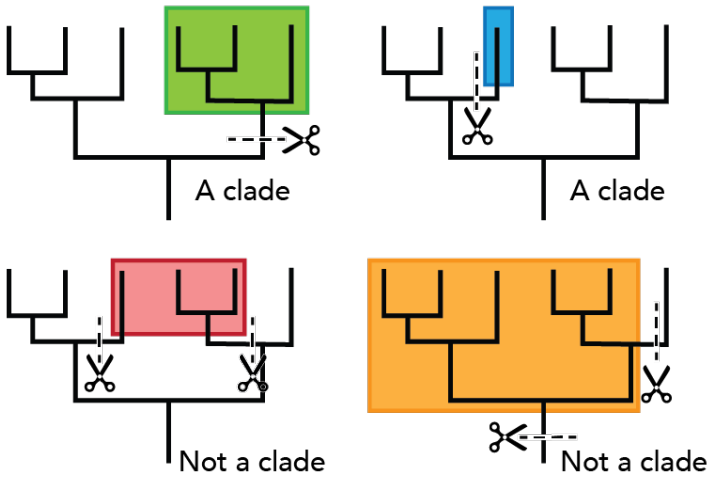
- Chronogram: Topology + Divergence times.
- Cladogram: Topology only.
- Phylogram: Topology + Divergence times + Divergence rates.

Clades/clusters

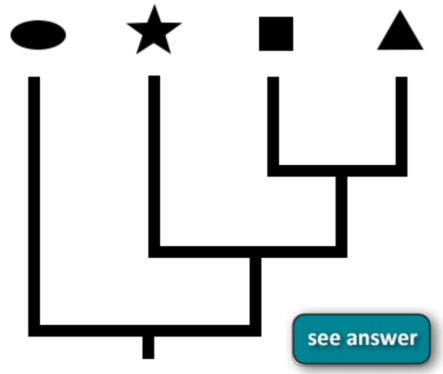


- Monophyletic (Clade): a taxon that consists of a most recent common ancestor and all its descendants.
- Paraphyletic: taxon that consists of a most recent common ancestor and some of its descendants.
- Polyphyletic: taxon that consists of unrelated organisms who are from a different recent common ancestor.

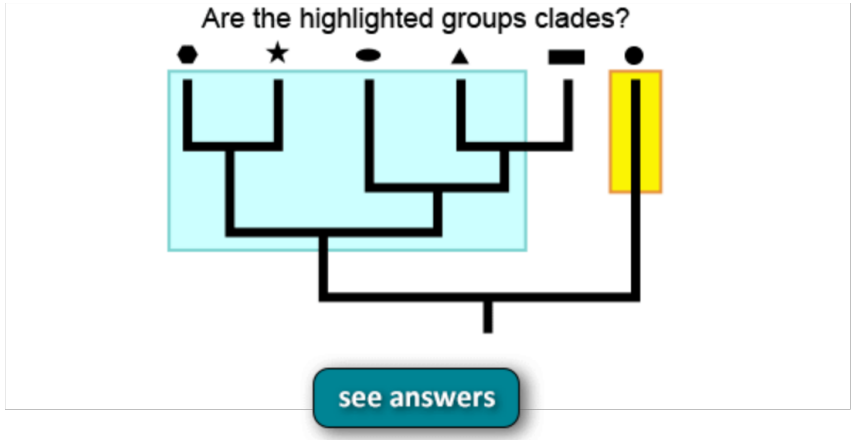
Clades/clusters



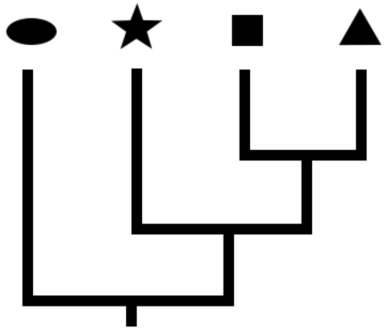
Which node represents the most recent common ancestor of the square taxon and the star taxon?



see answer



Which taxa are more closely related:
the oval and the triangle or
the triangle and the star?



see answer

Phylogenetic tree construction

1. Alignment (both building the data model and extracting a phylogenetic dataset).
2. Determining the substitution model
3. Tree building
4. Tree evaluation

Multiple sequence alignment

Why compare sequences?

- Given a new sequence, infer its **function** based on similarity to another sequence.
- Find important **molecular regions** – conserved across species.
- Determine the **evolutionary constraints** at work.
- Find **mutations** in a population or family of genes.

Tools: ClustalW, MUSCLE, MAFFT, Clustal Omega, T-Coffee, MULTALIN

Extraction of a phylogenetic dataset

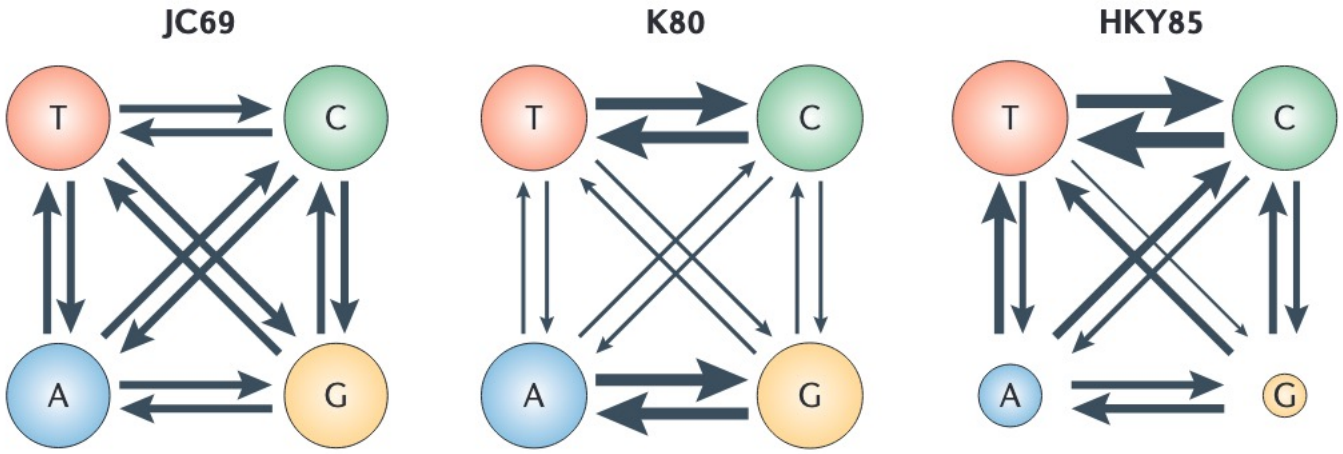
For length-variable sequences with insertions deletions (indels):

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

Tools: trimal, clipkit

Evolutionary models

- 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

Yang, Z and Rannala B. (2012). *Nature Genetics Reviews*, **13 (5)**, 303-314. doi: 10.1038/nrg3186

Tree building

1. 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)

Tree building: Distance-based method

```
A    ATCGTGGTACTG
     B    CCGGAGAACTAG
          C  AACGTGCTACTG
D    ATGGTGAAAGTG
     E    CCGGAAAACTTG
F    TGGCCCTGTATC
```

Tree building: Distance-based methods

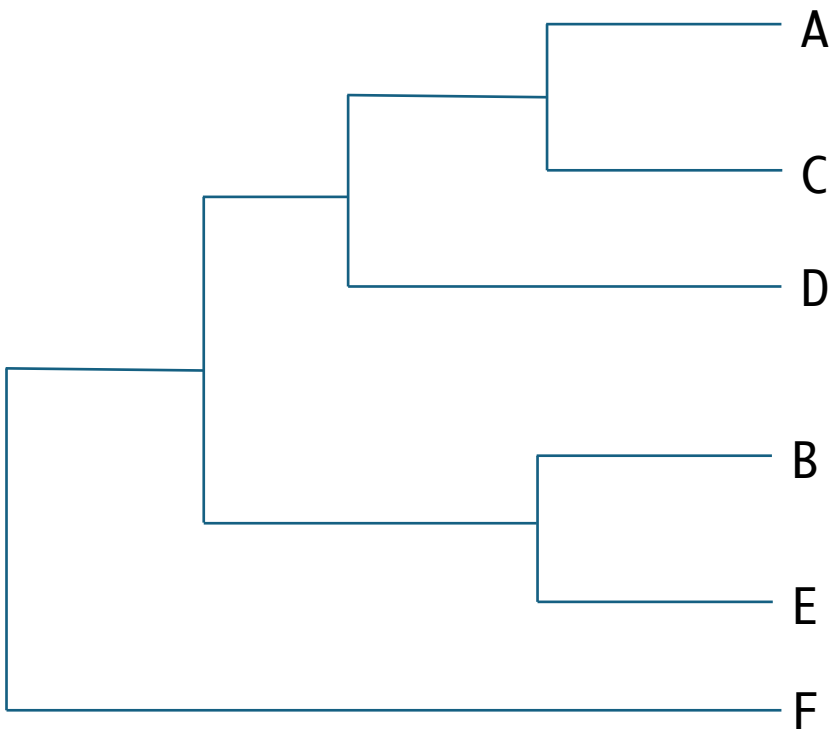
```
A    A T C G T G G T A C T G
B    C C G G A G A A C T A G
C    A A C G T G C T A C T G
D    A T G G T G A A A G T G
E    C C G G A A A A C T T G
F    T G G C C C T G T A T C
```

1. Align sequences

	A	B	C	D	E	F
A	0	9	2	4	9	10
B		0	9	6	2	10
C			0	5	9	10
D				0	6	10
E					0	10
F						0

2. Determine pairwise differences between all sequence pairs

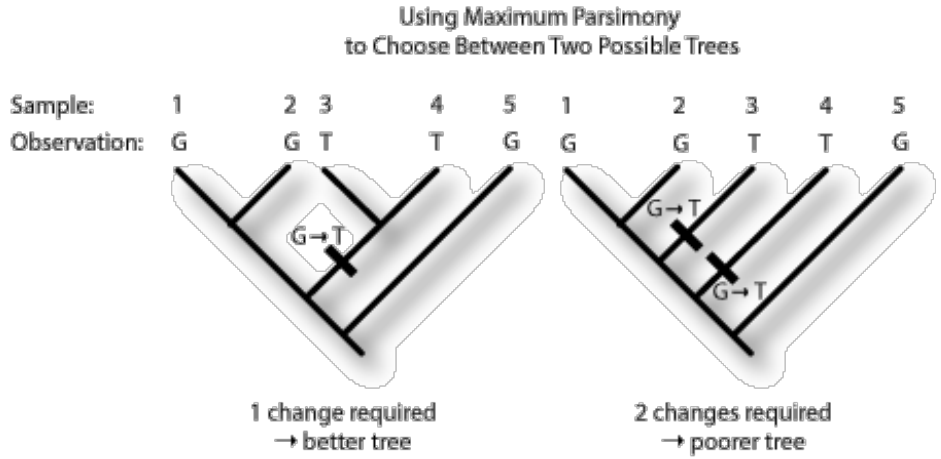
Tree building: Distance-based method



3. Draw the first groupings of the tree, recalculate distance of the joint pair by taking the average, Repeat this process until all species are connected in a single cluster

Tree building: Character-based method

Maximum Parsimony (MP): Predict the evolutionary tree that minimizes the number of steps required to generate the observed variation in the sequences.



Tools: PAUP, MEGA, PHYLIP

Tree building: Bayesian methods

Bayesian Inference

Bayesian inference is a probabilistic method. The method was developed by Thomas Bayes theorem (1701–1761) and consists of describing the probability of events based on a priori knowledge about the event. The theorem is described by the equation.

$$P(H|D) = \frac{P(D|H).P(H)}{P(D)}$$

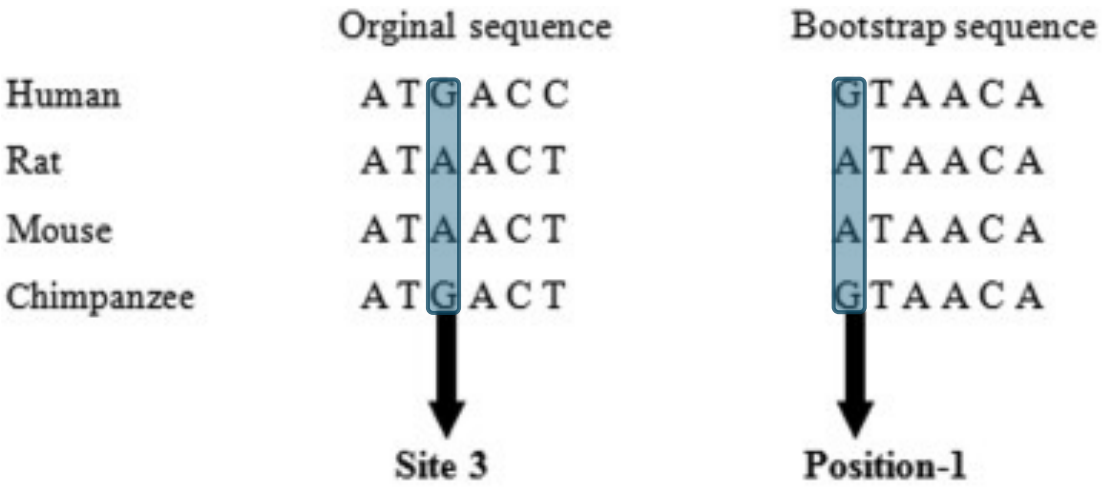
- $P(H|D)$ is the posterior probability
- $P(H)$ is the prior probability for hypothesis.
- $P(D|H)$ is the likelihood or the probability of the data given the hypothesis
- $P(D)$ is constant when comparing the fit of different models for a given data set and thus has no influence on Bayesian model selection under most circumstances.

Bayesian phylogenetic inference relies on MCMC algorithms to generate a sample from the posterior distribution.

Tools: BEAST, MrBayes, RevBayes

Tree evaluation: 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.



Site-3 is placed at position one in bootstrap sequence and next five randomly chosen sites: 2, 1, 1, 5, 4 are placed in next five position.

Krishna *et. al*, 2021. *Bioinformatics*.

Evolutionary dynamics

- 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 evolutionary viral analysis

- Increasing availability and quality of viral genome sequences.
- Growth in computer processing power.
  - GPU enabled servers
- Development of sophisticated statistical methods.
  - Bayesian Evolutionary Analysis by Sampling Trees (BEAST)

Phylodynamic modeling applications

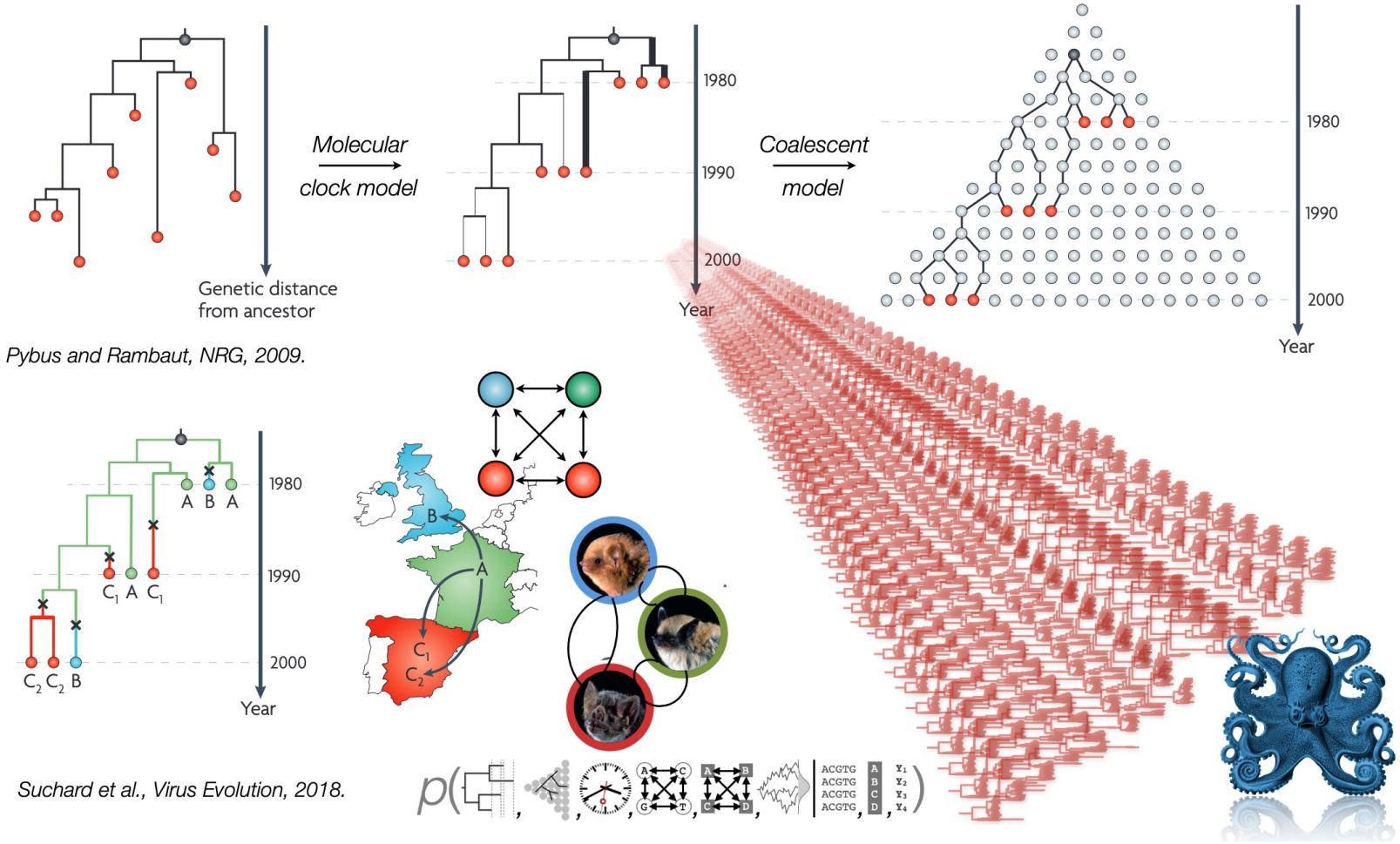
- Molecular clock: Reconstructing evolutionary change on a natural timescale of months or years
  - Date epidemiologically important events such as zoonotic transmissions.
  - Allows pathogen evolution to be directly compared with known surveillance data.

- 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

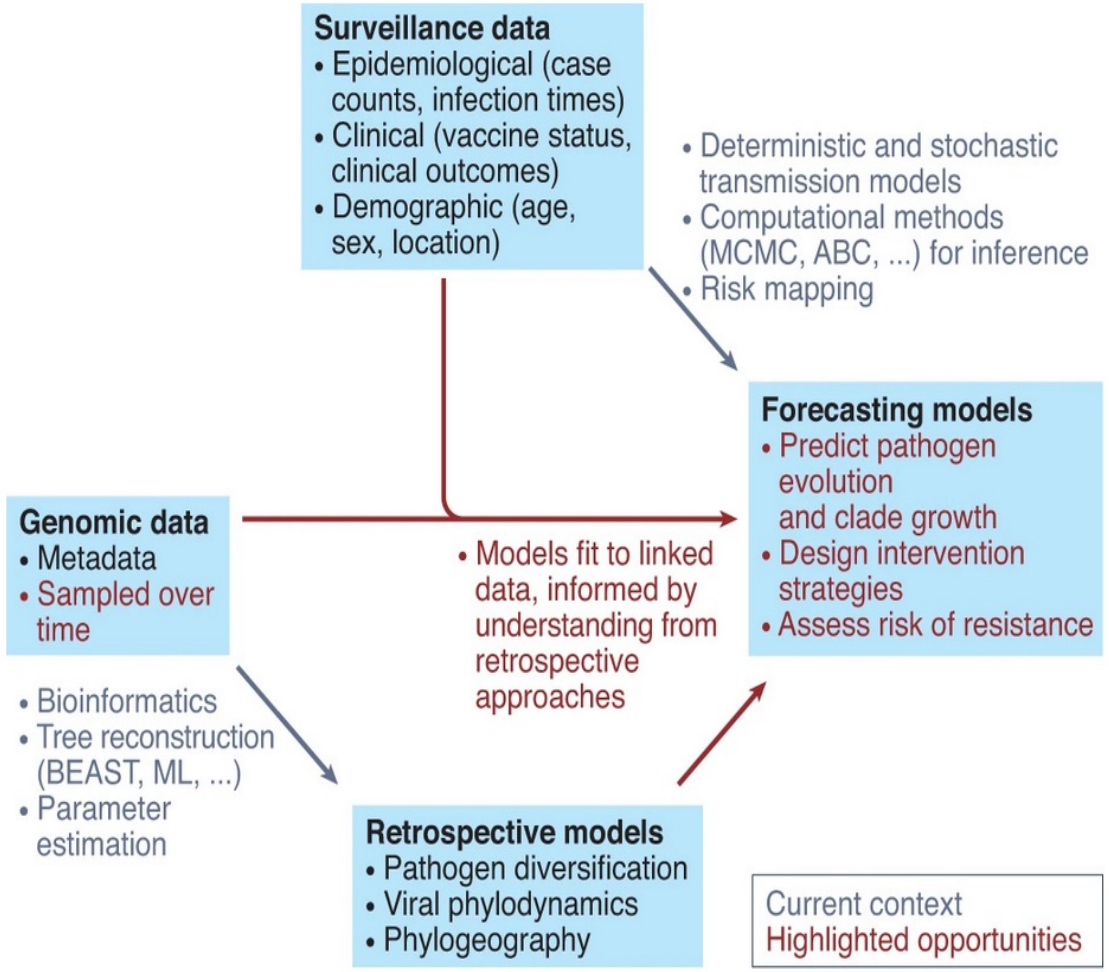
Phylodynamic modeling

Phylodynamics: union of immunodynamics, epidemiology and evolutionary biology. This captures both the evolutionary and epidemiological information from pathogens.

Phylodynamics relies on tools such as Bayesian evolutionary analysis sampling trees (BEAST), in which genomic sequence data are used to build a time-labelled phylogenetic tree using a specific evolutionary process as a guide.



Future considerations and general conclusions



Stockdale, J.E., Liu, P & Colijn, C *et al. Nature comunications* **7,** 1736–1743 (2022)
doi:https://doi.org/10.1038/s41564-022-01233-6


ILRI
INTERNATIONAL
LIVESTOCK RESEARCH
INSTITUTE
CGIAR


ASLM
AFRICAN SOCIETY FOR LABORATORY MEDICINE


AFRICA CDC
Centres for Disease Control and Prevention
Safeguarding Africa's Health