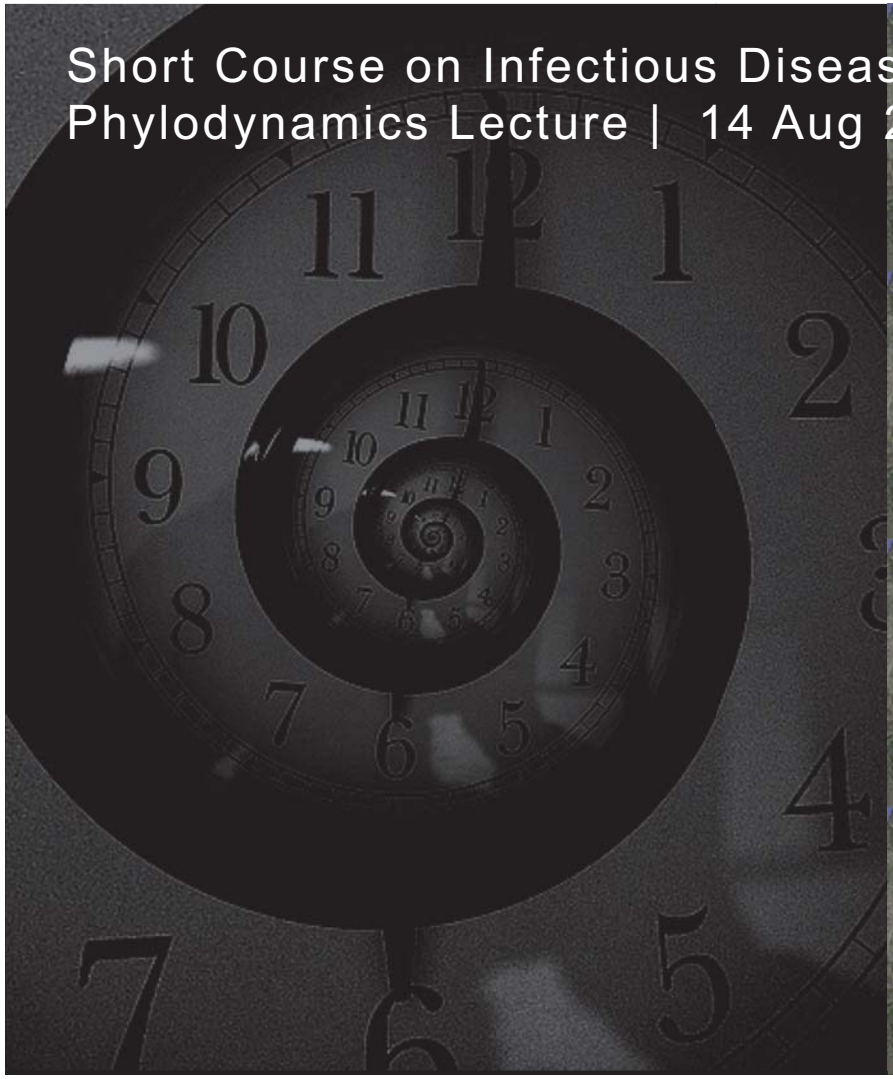
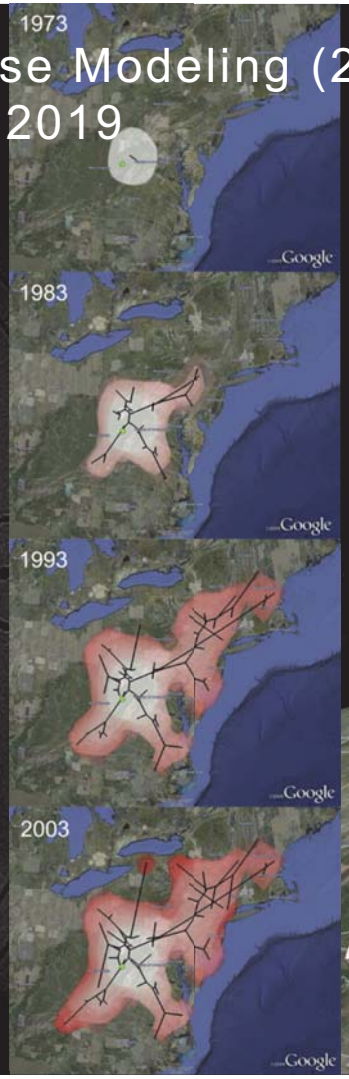
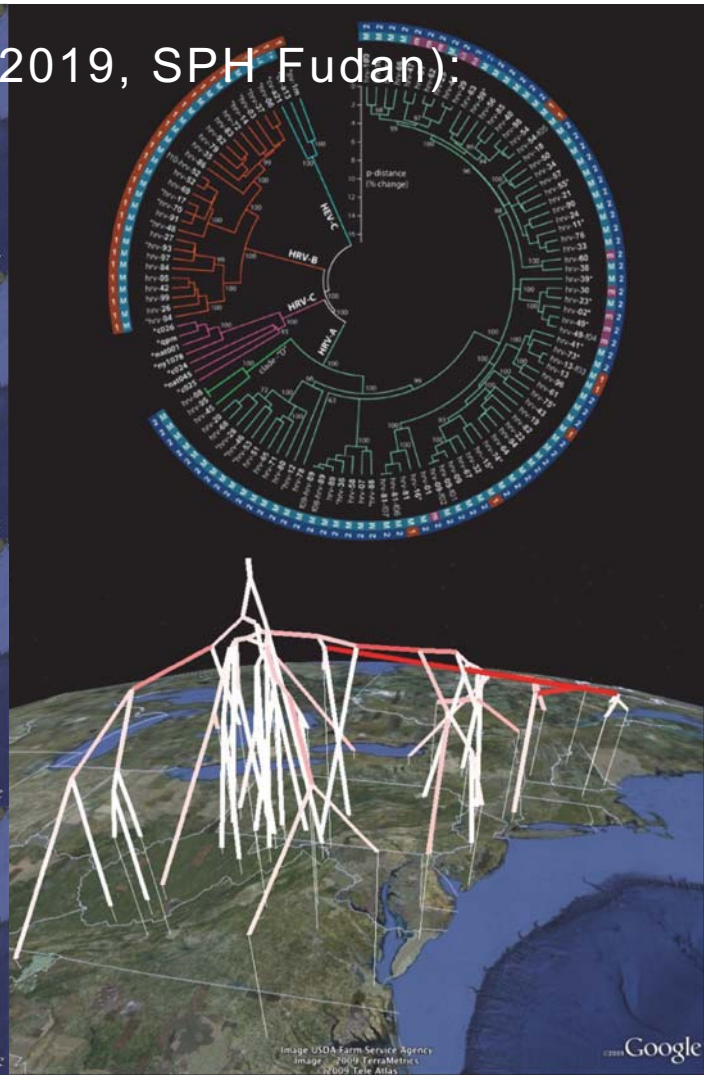
Short Course on Infectious Disease Modeling (2019, SPH Fudan):
Phylodynamics Lecture | 14 Aug 2019







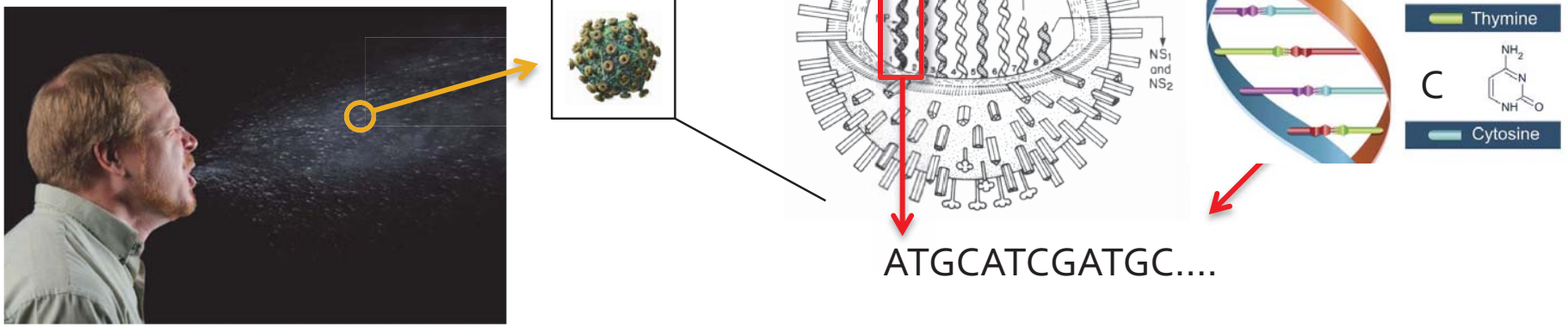
Phylodynamic Analysis

for studying infectious diseases

Tommy Lam
State Key Laboratory of Emerging Infectious Diseases
School of Public Health, HKU

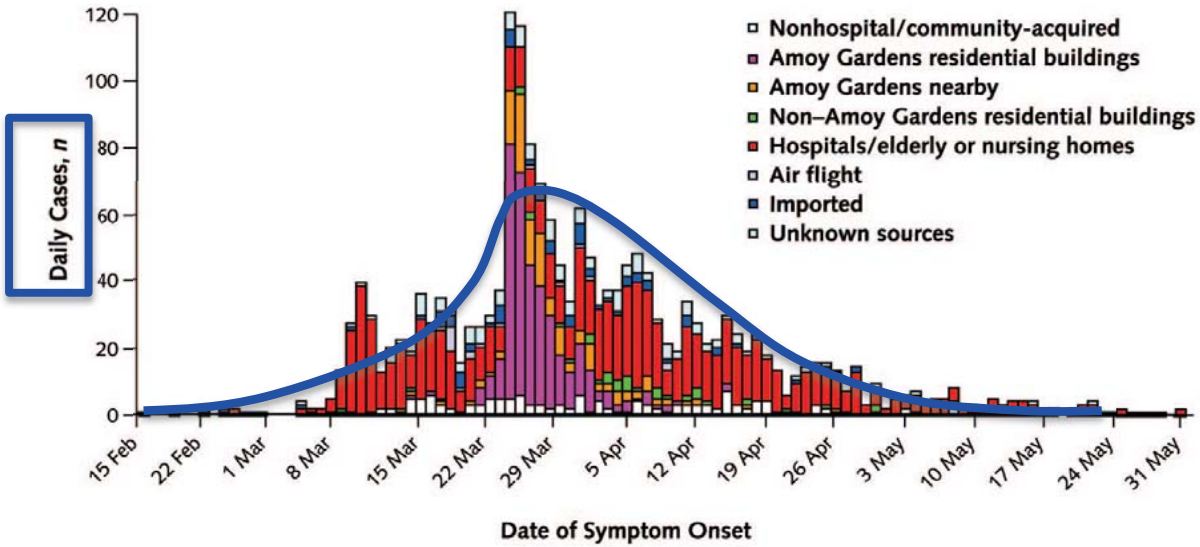
Genetic Epidemiology of Infectious Diseases

- **Genetic sequences** of the pathogens are often used as *molecular biomarkers* to study the molecular epidemiology.
  - **High precision** in obtaining the pathogen's identity
  - **Quantitative** transmission information is stored

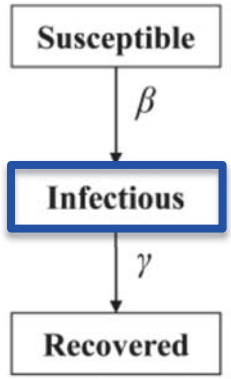


Studying ID using Incidence and Genetic Data

Annals of internal medicine 141, 662-673 (2004)



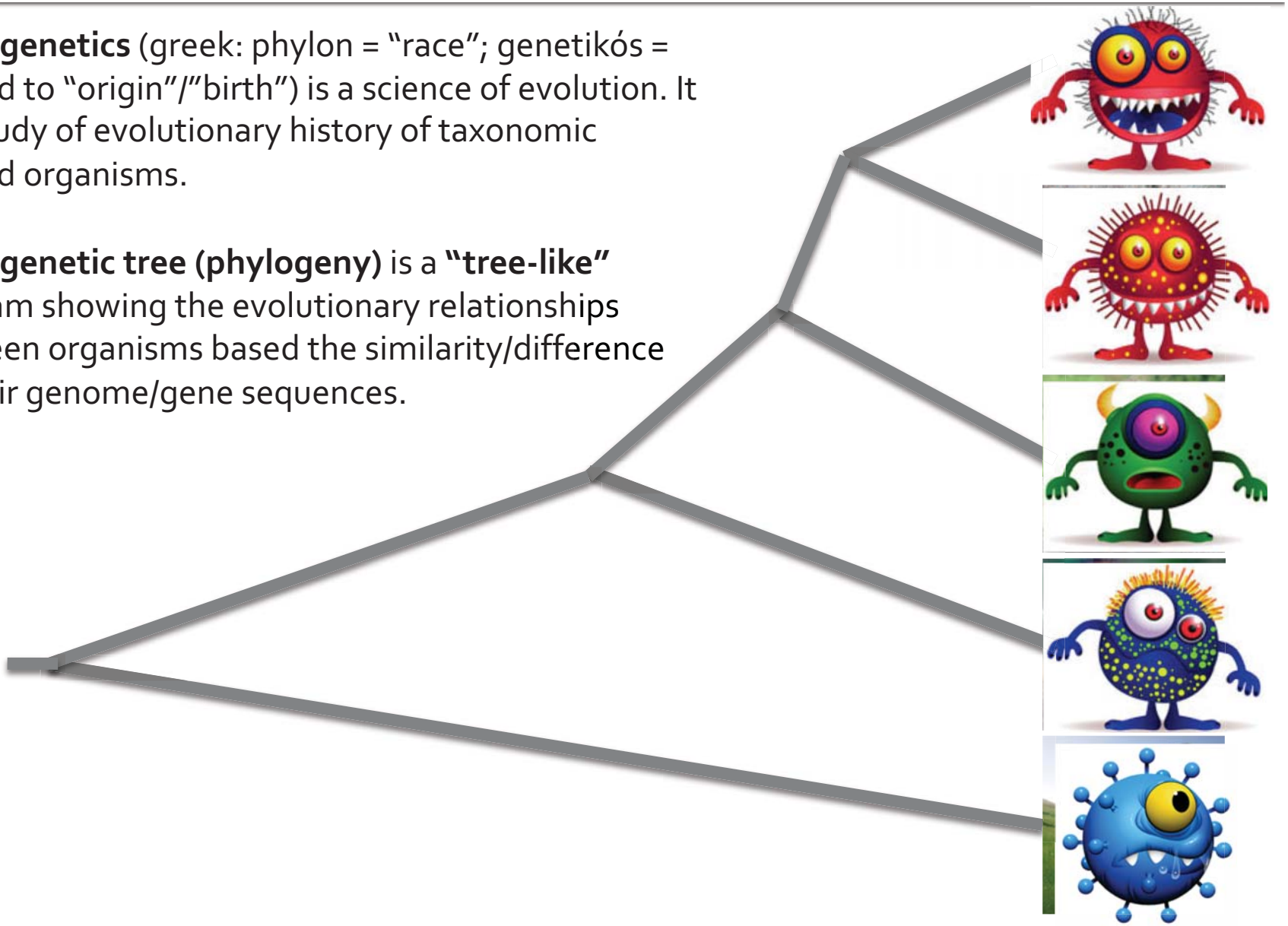
*Modeling the epidemic history via compartmental model of **host population** health status with incidence data*



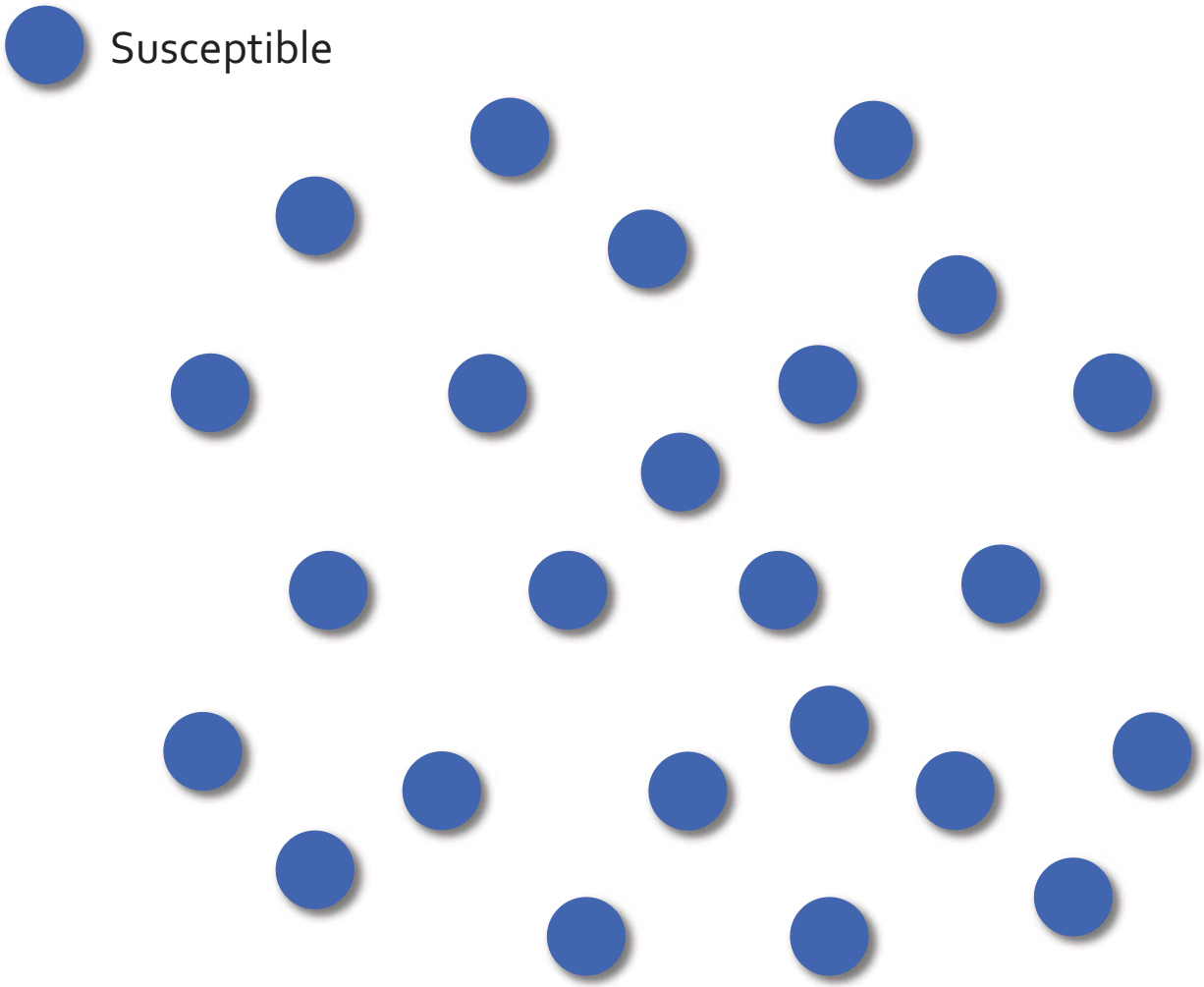
Terminology

- **Phylogenetics** (greek: phylon = "race"; genetikós = related to "origin"/"birth") is a science of evolution. It is a study of evolutionary history of taxonomic related organisms.

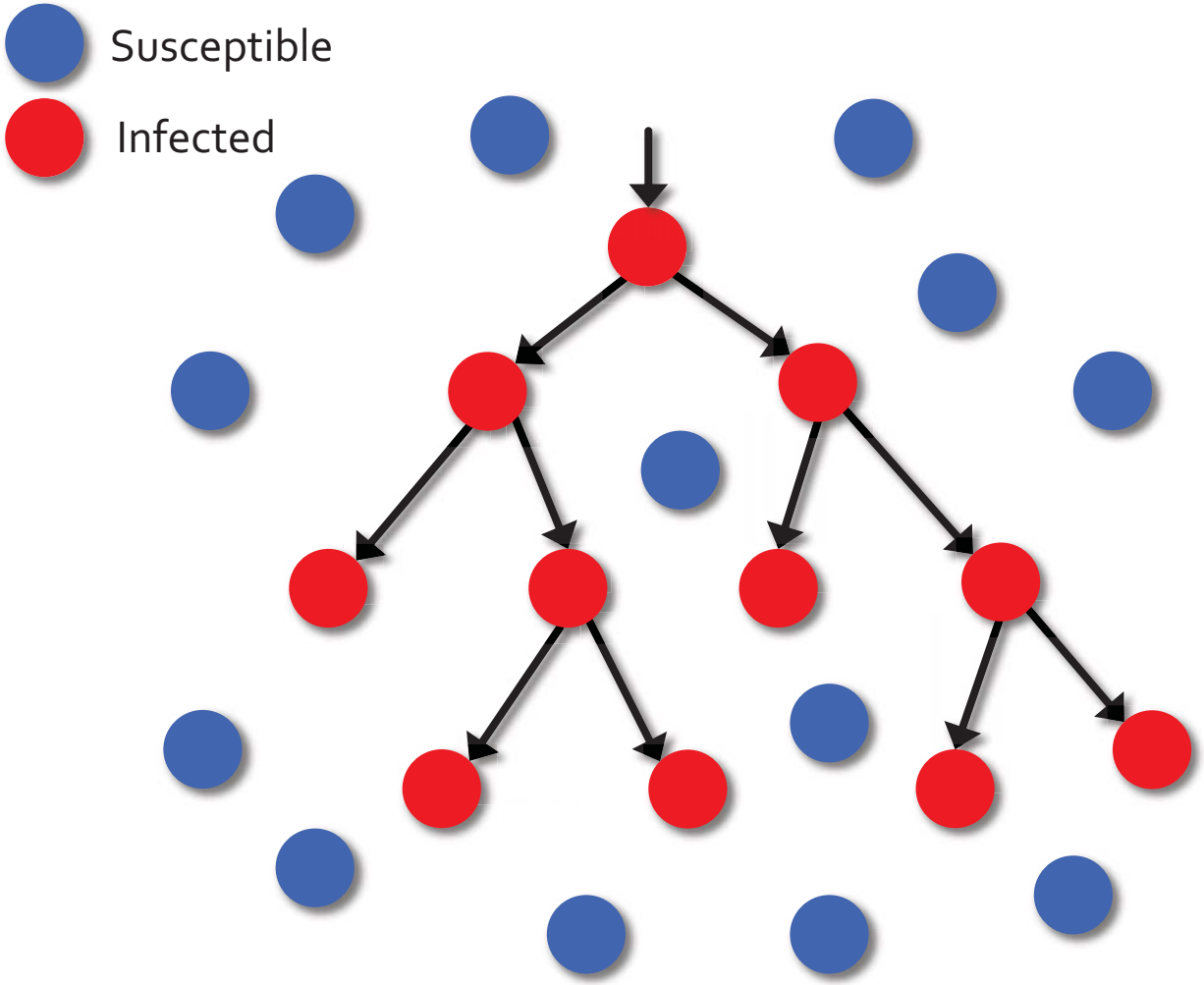
- **Phylogenetic tree (phylogeny)** is a **"tree-like"** diagram showing the evolutionary relationships between organisms based the similarity/difference of their genome/gene sequences.



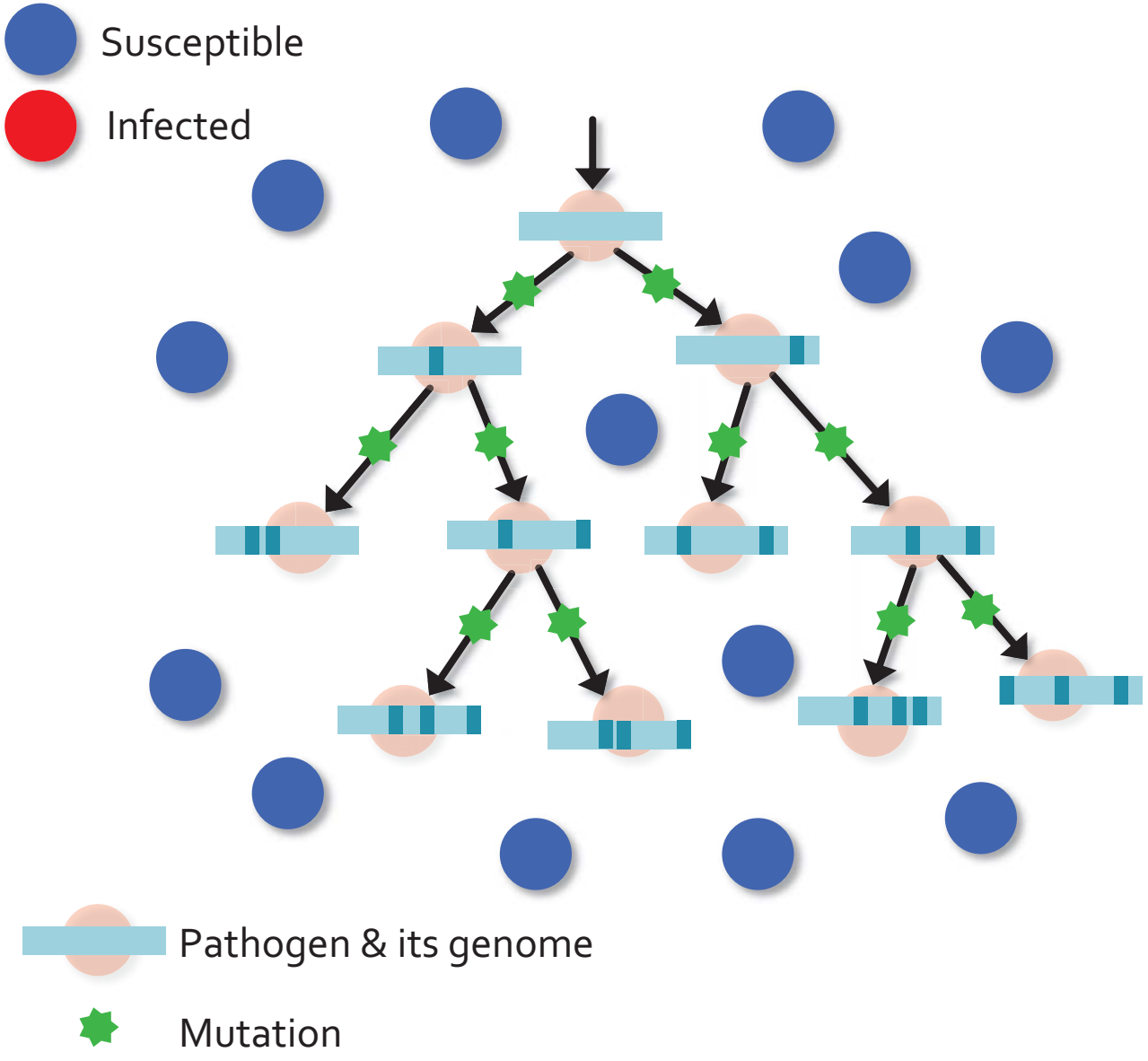
Transmission history is stored in pathogen genomes



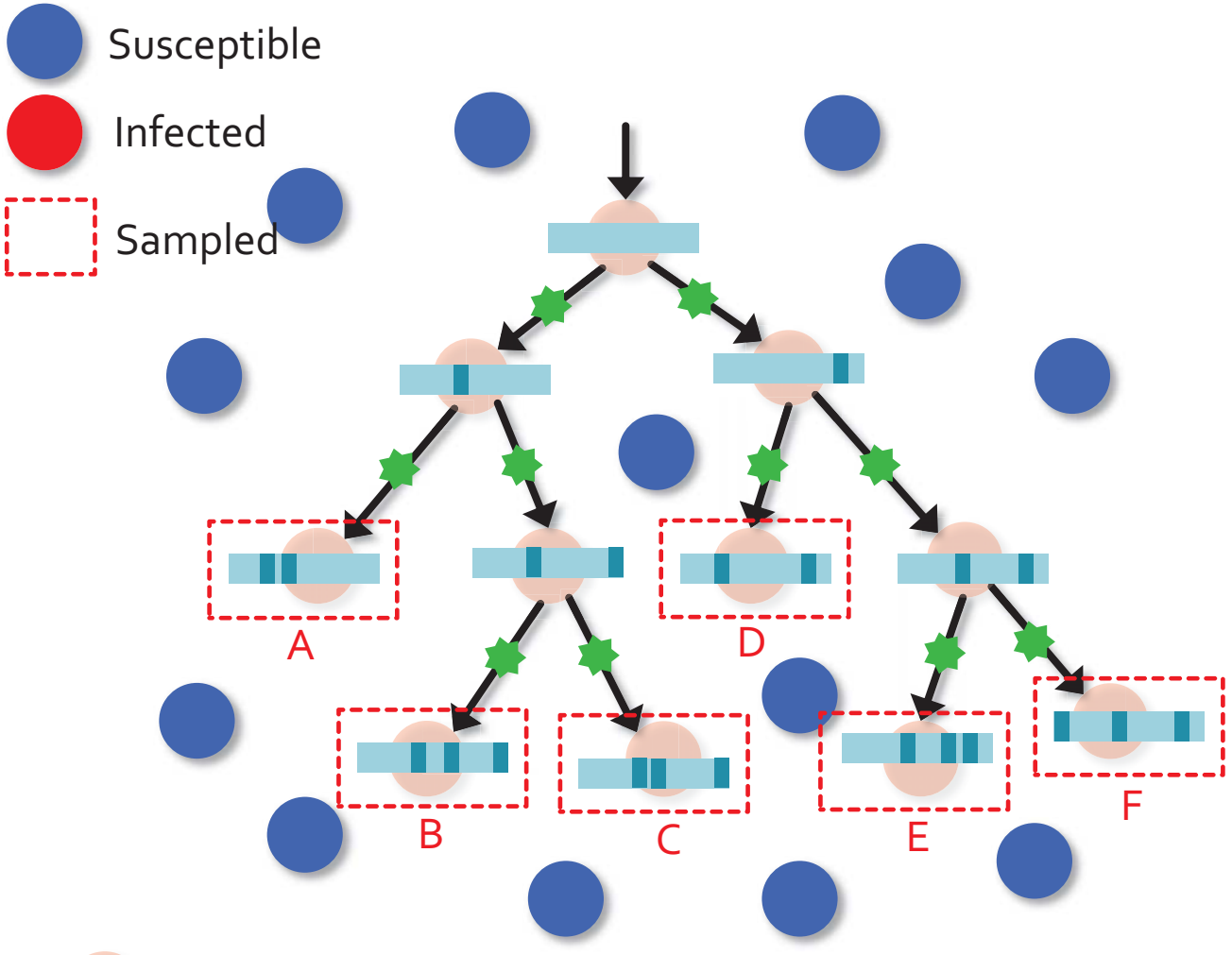
Transmission history is stored in pathogen genomes



Transmission history is stored in pathogen genomes

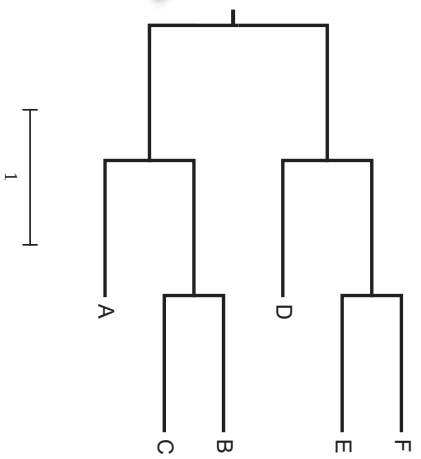


Transmission history is stored in pathogen genomes



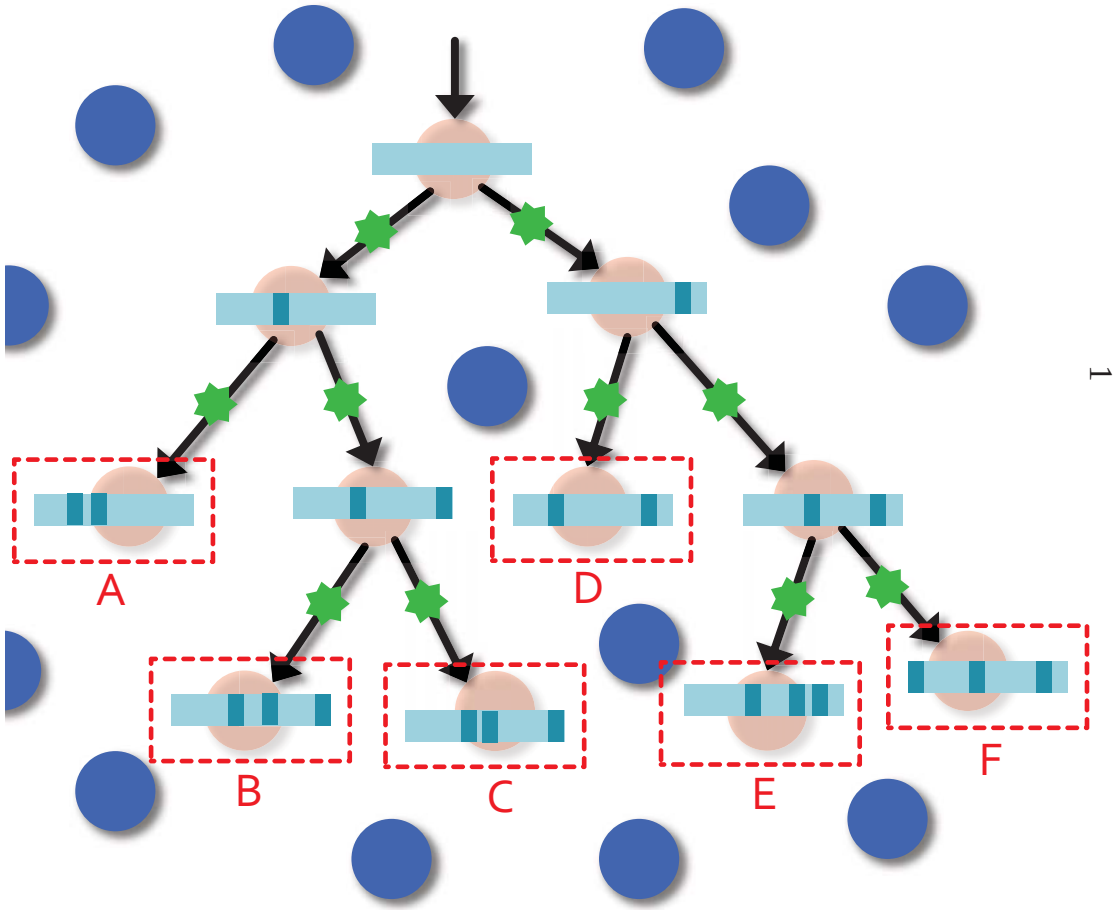
A| AATTAAAAAAA
B| ATATAAAAAAT
C| AAATTAAAAAT
D| AAAAATAATAA
E| AAAAAATTTAA
F| TAAAAATATAA

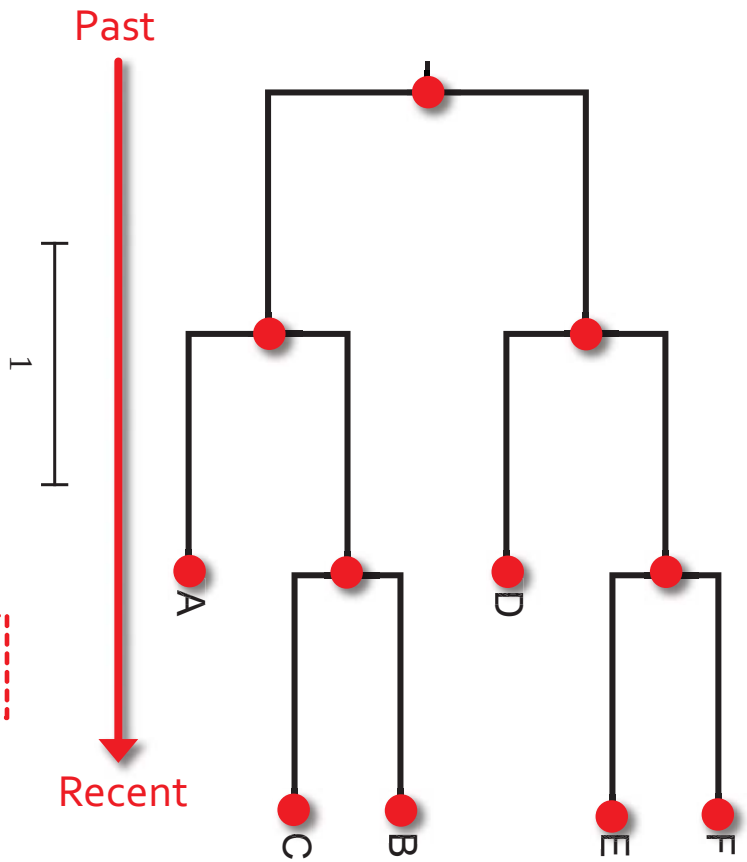
Build a phylogenetic tree



Transmission history ≈ Evolutionary history

Terminology in a phylogenetic tree

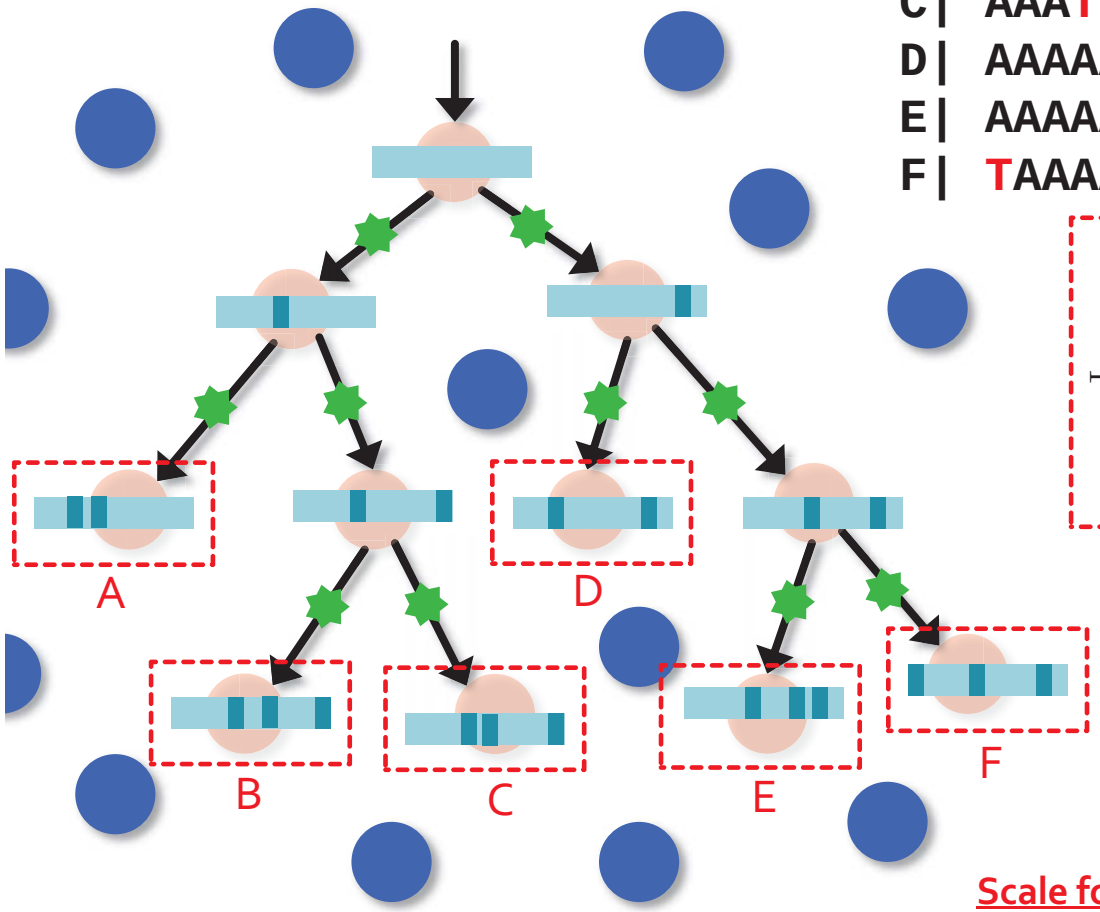




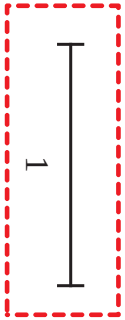
Balls = Node

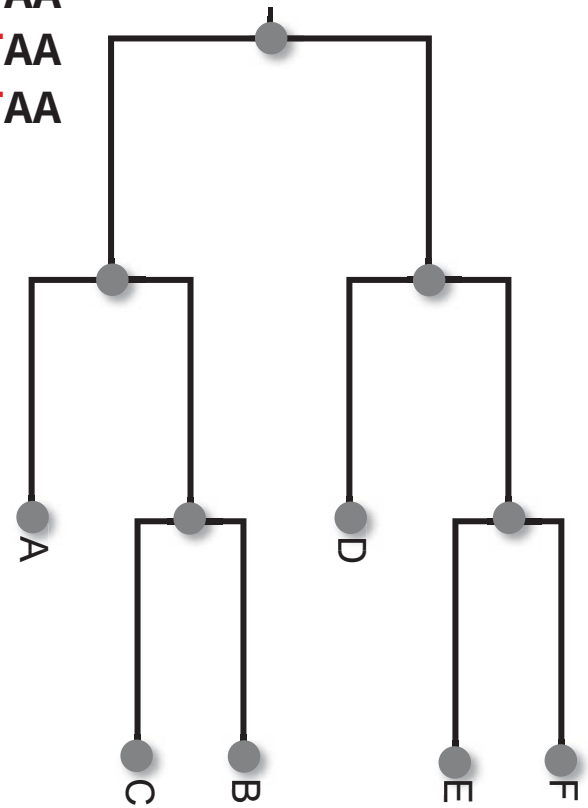
Sticks = Branch

Terminology in a phylogenetic tree




```
A|  AATTAAAAAAA
B|  ATATAAAAAAT
C|  AAATTAAAAAT
D|  AAAAATAATAA
E|  AAAAAATTTAA
F|  TAAAAATATAA
```

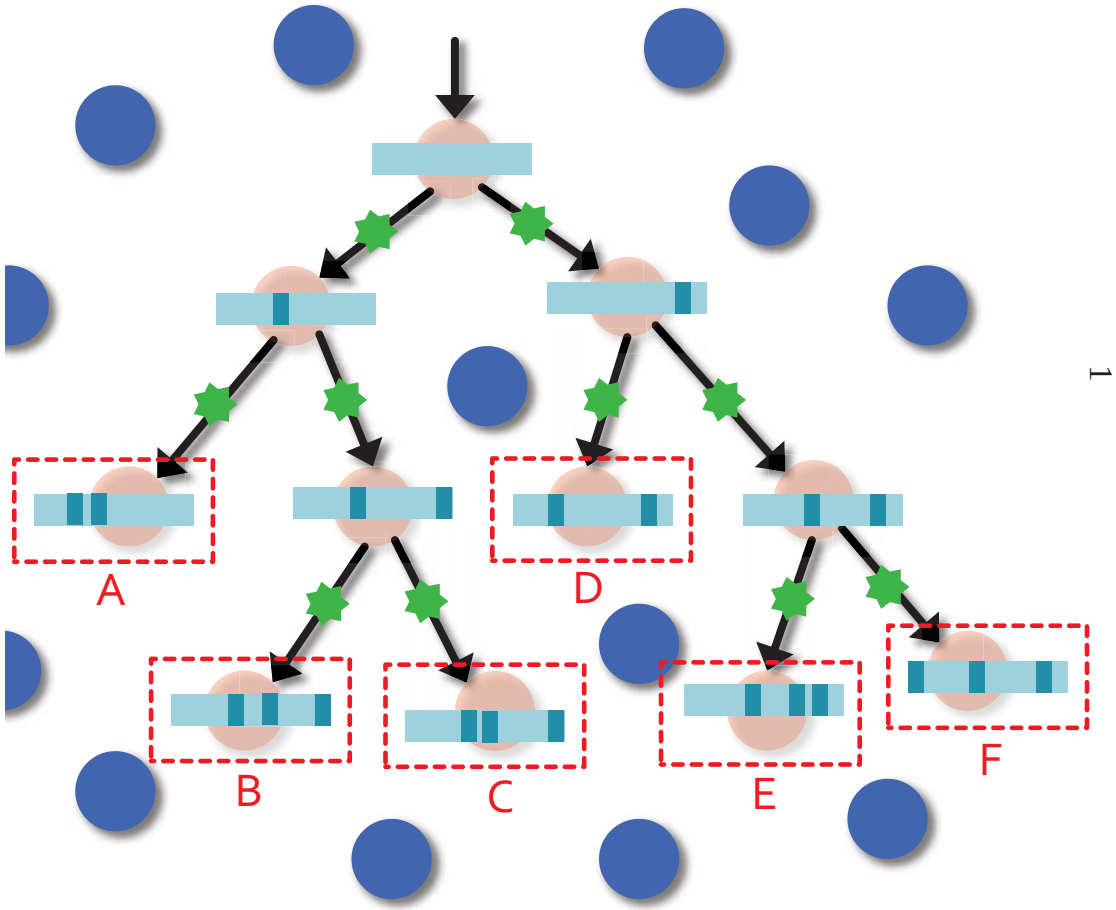


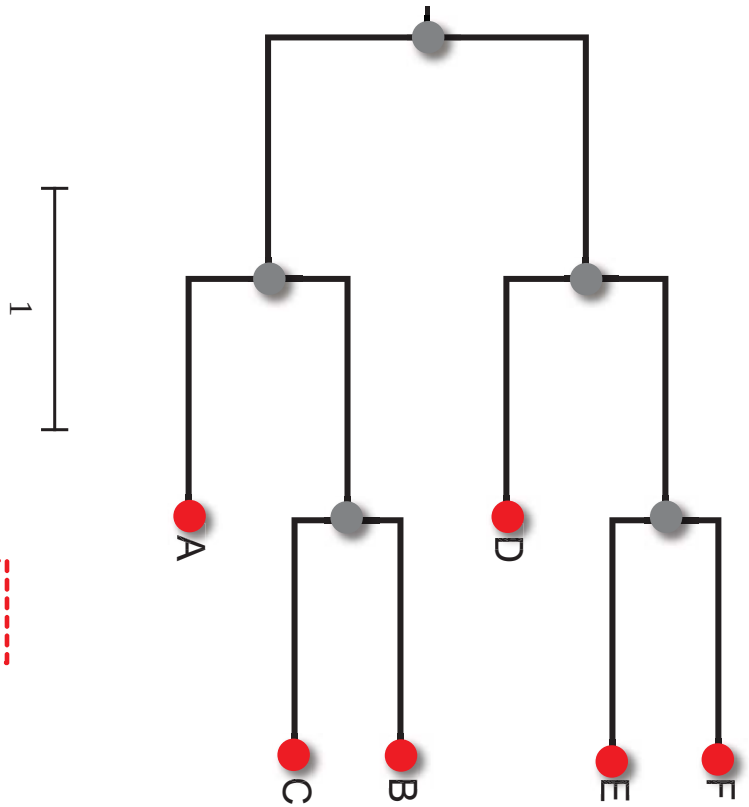


Scale for branch length

- absolute substitutions
- substitutions per total sites
- time

Terminology in a phylogenetic tree

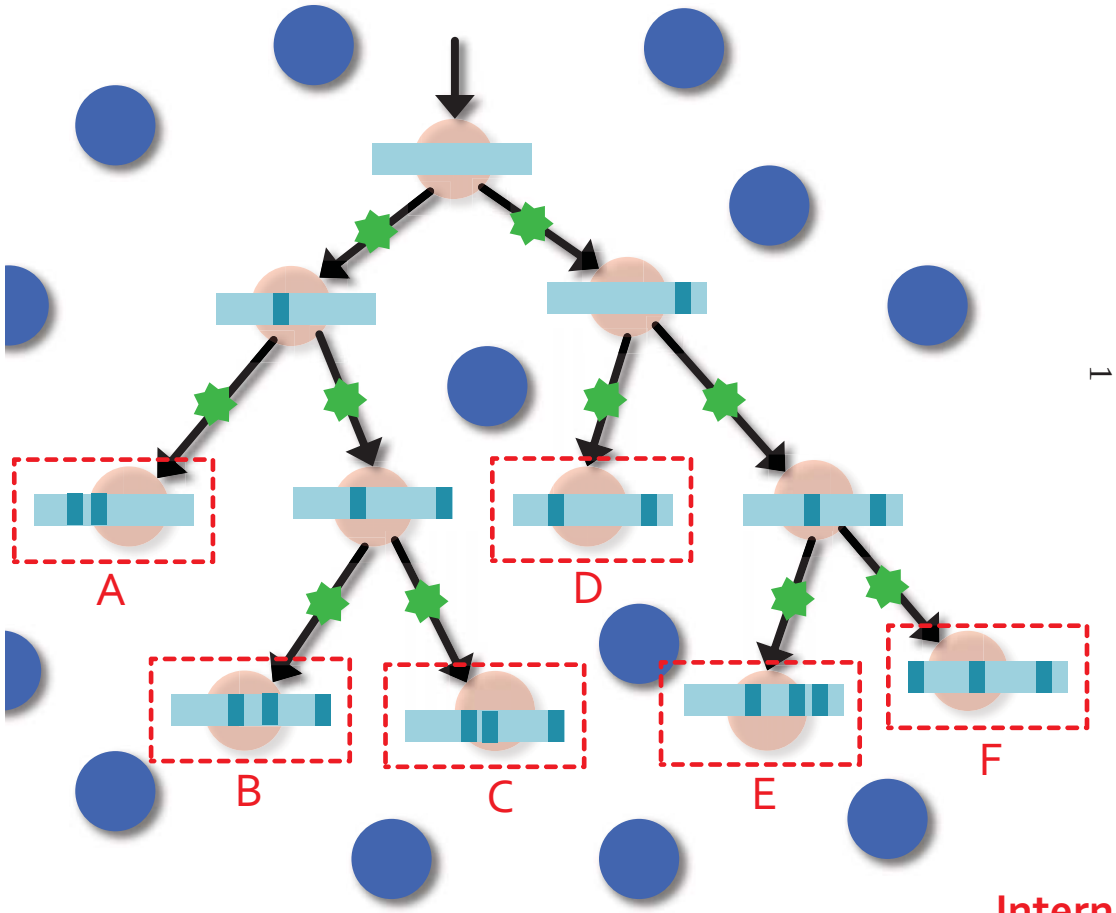


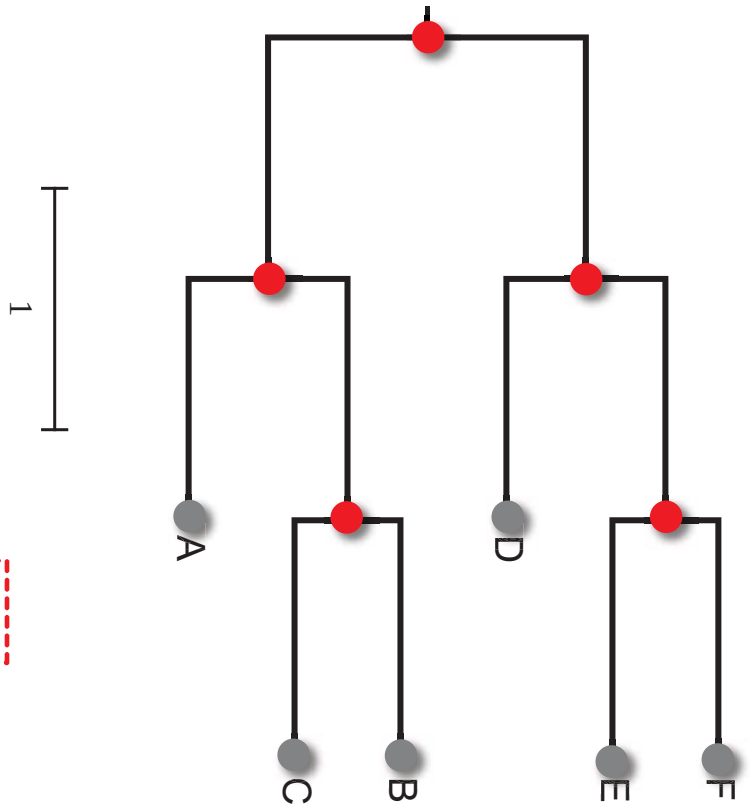


External nodes

- Evol-bio term: 'taxon' (plural 'taxa')
- Sequences obtained from samples

Terminology in a phylogenetic tree

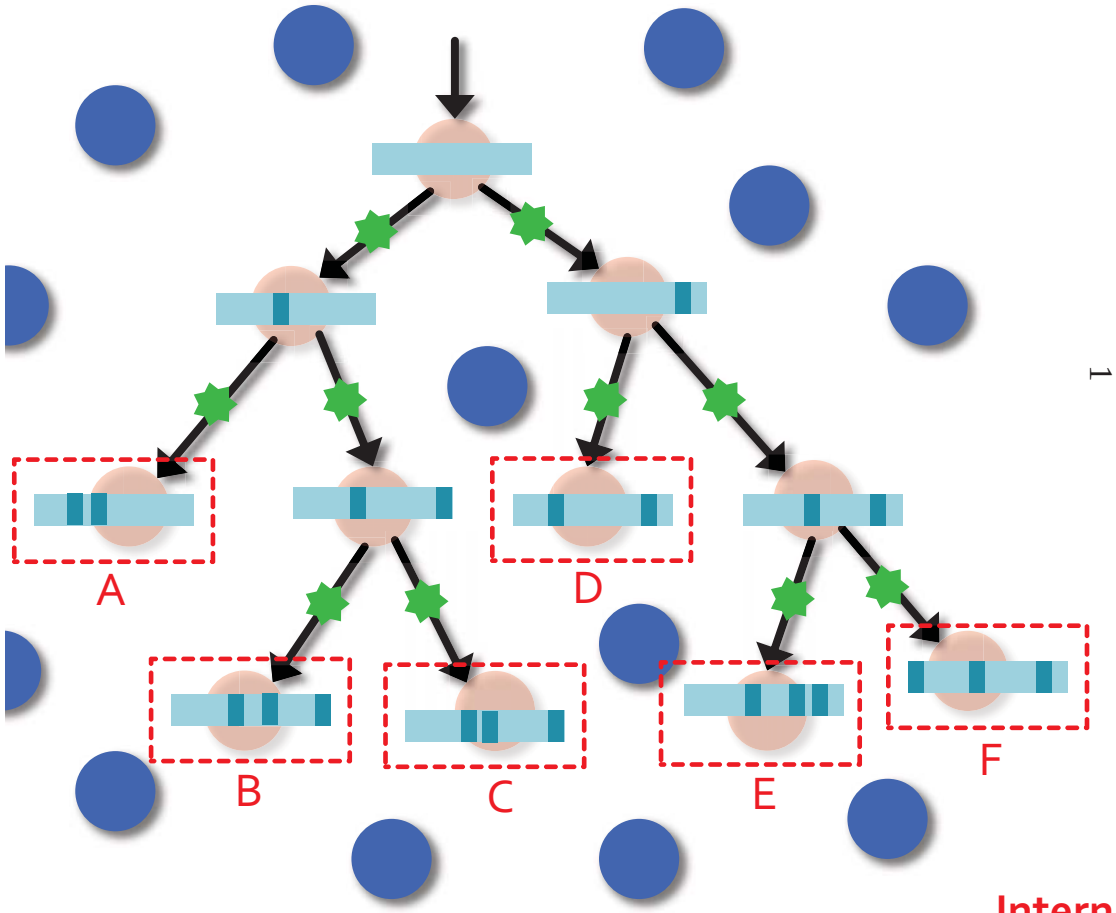


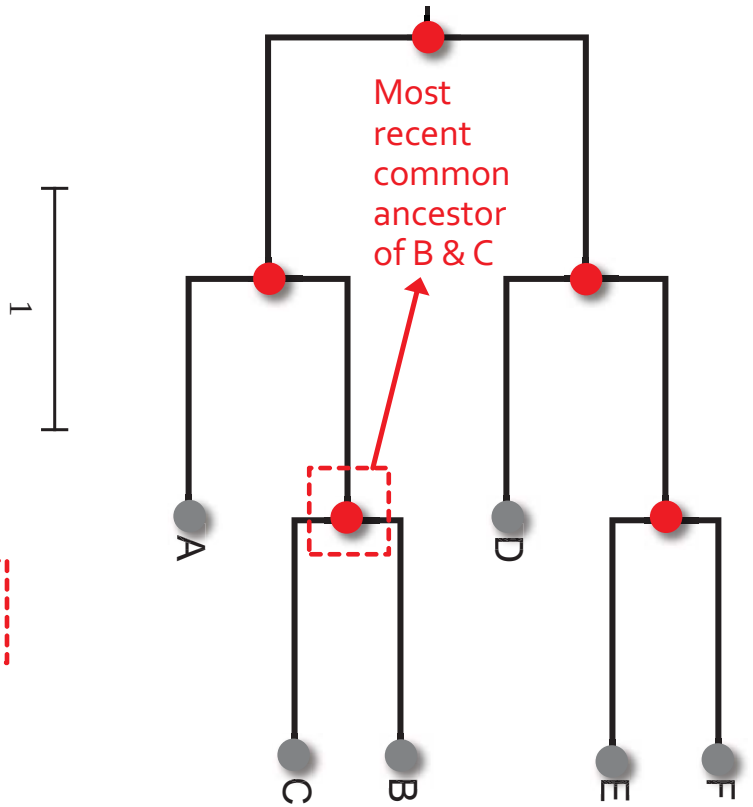


Internal nodes

Hypothetical ancestors of the samples or other nodes; Possibly un-sampled transmitters

Terminology in a phylogenetic tree

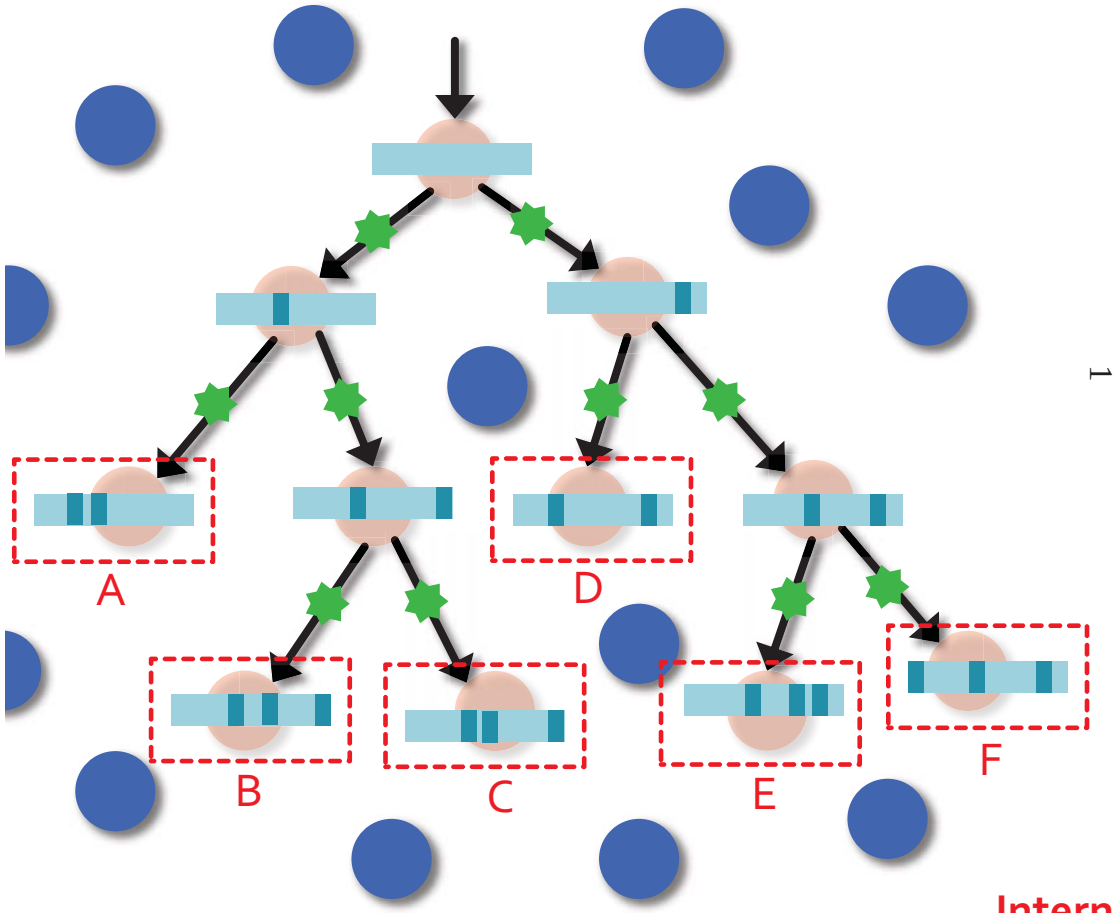


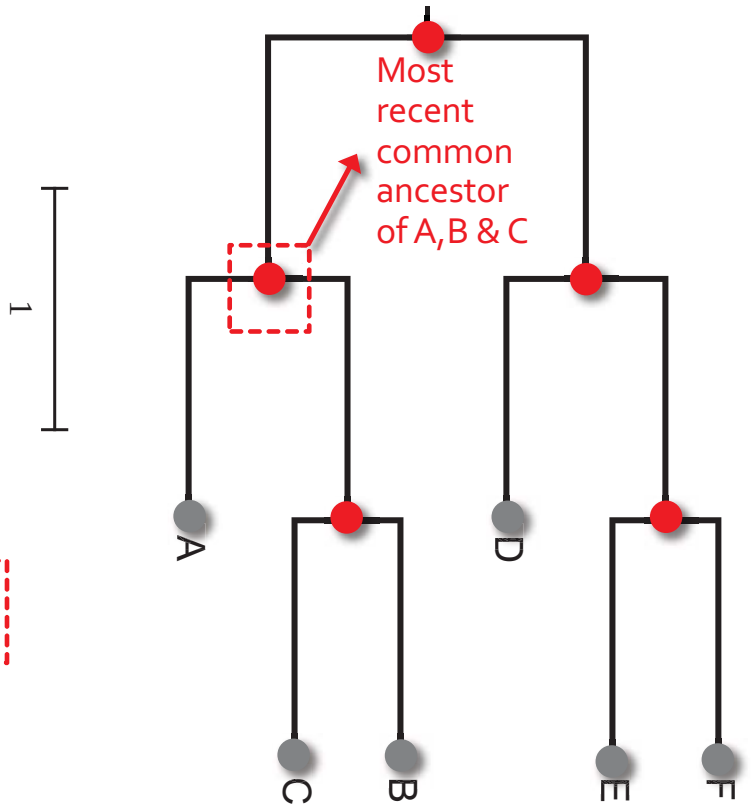


Internal nodes

Hypothetical ancestors of the samples or other nodes; Possibly un-sampled transmitters

Terminology in a phylogenetic tree





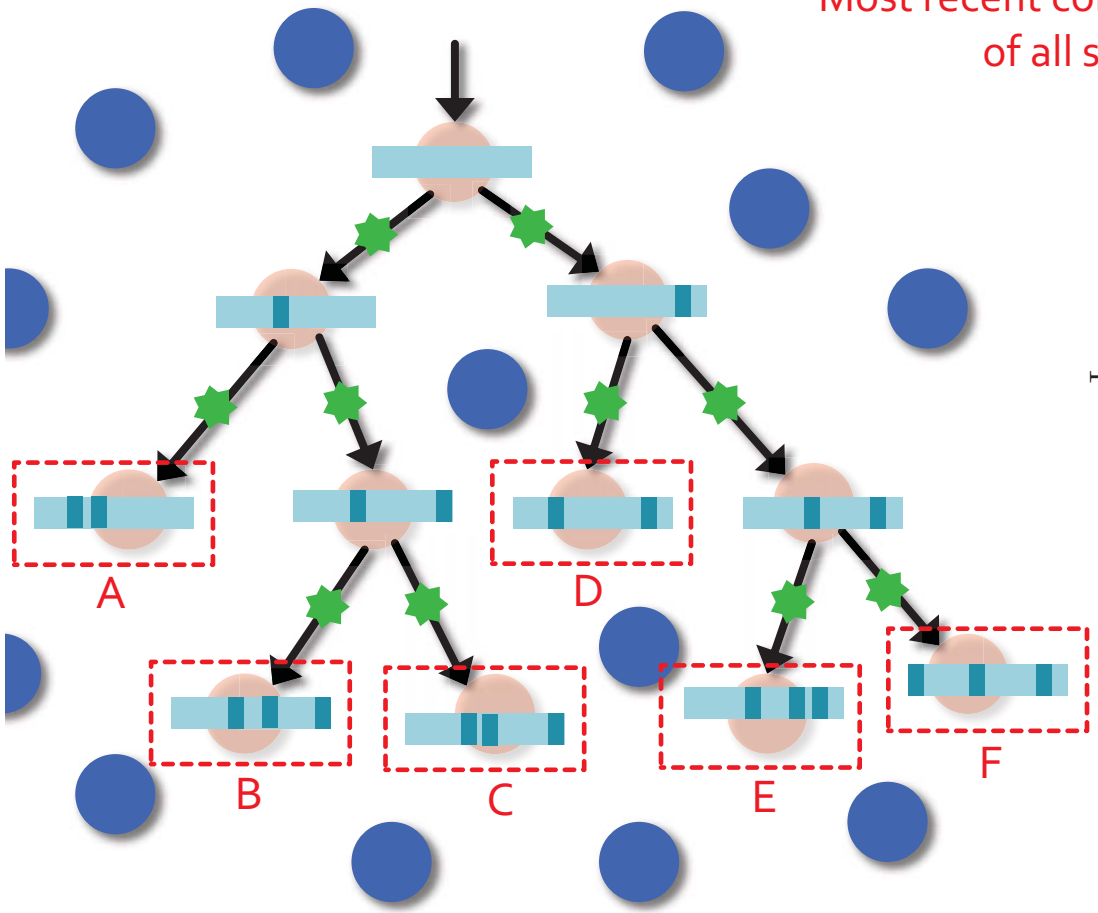
Internal nodes

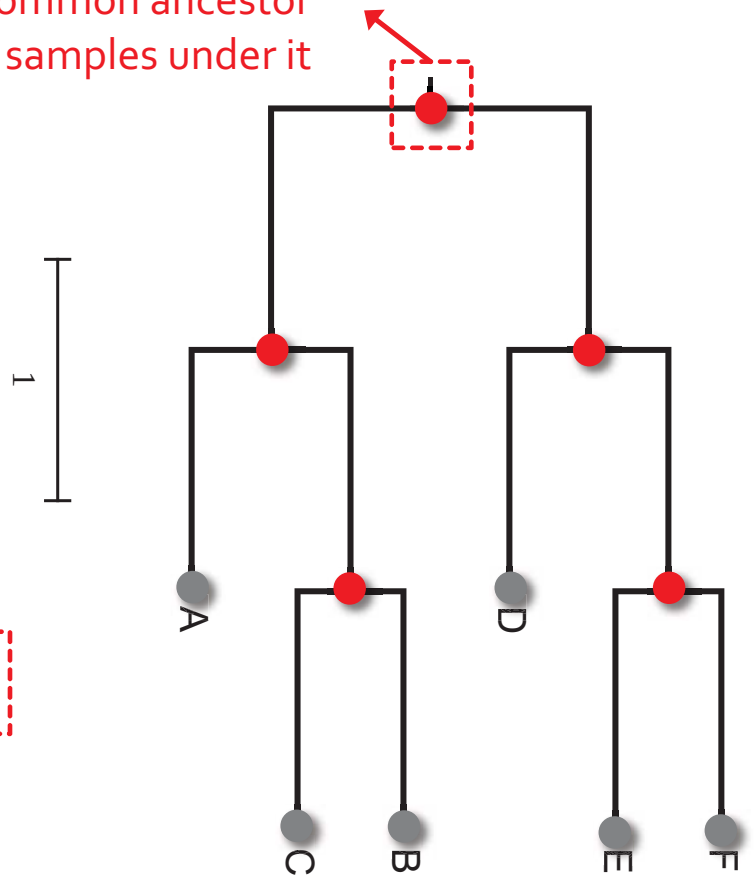
Hypothetical ancestors of the samples or other nodes; **Possibly un-sampled transmitters**

Terminology in a phylogenetic tree

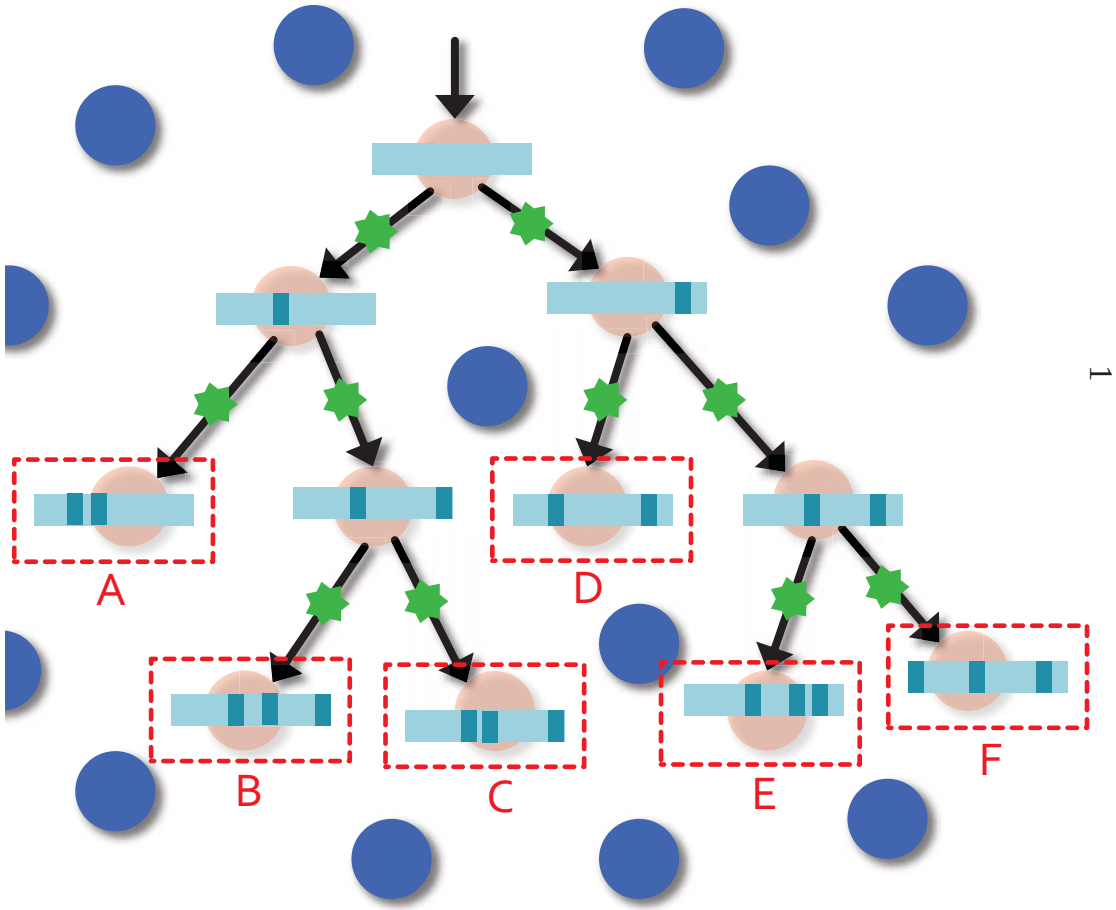
Root node

Most recent common ancestor of all samples under it

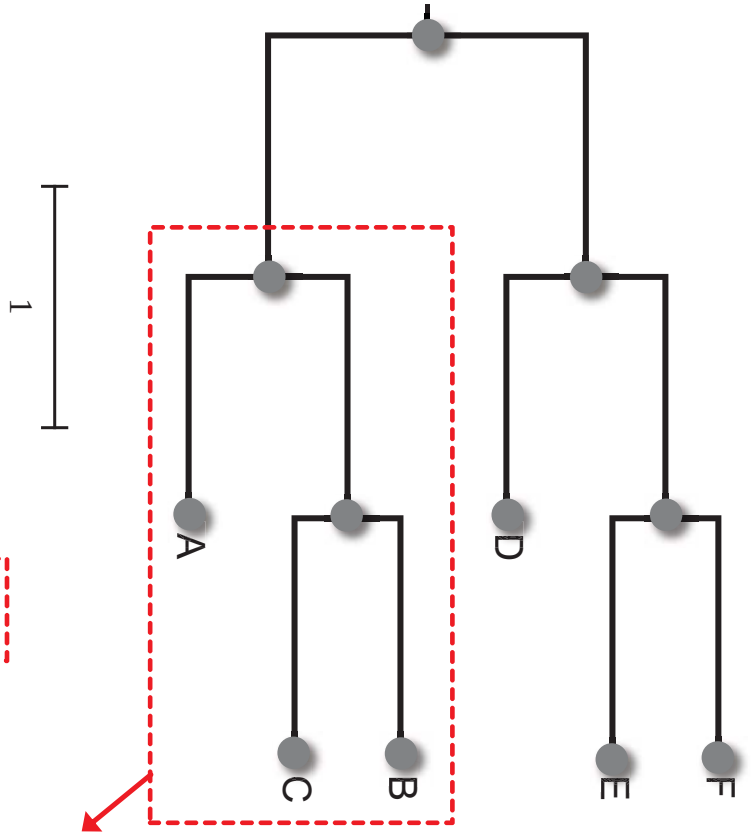




Terminology in a phylogenetic tree

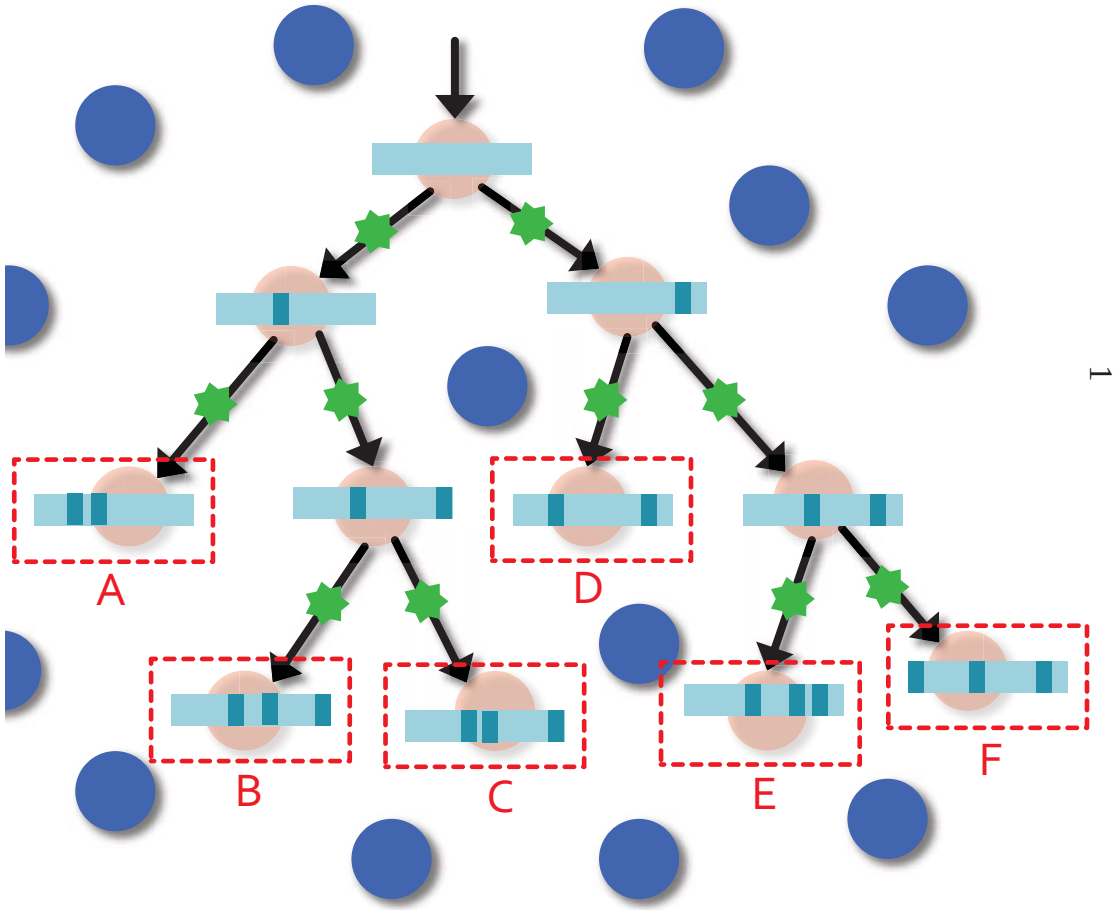


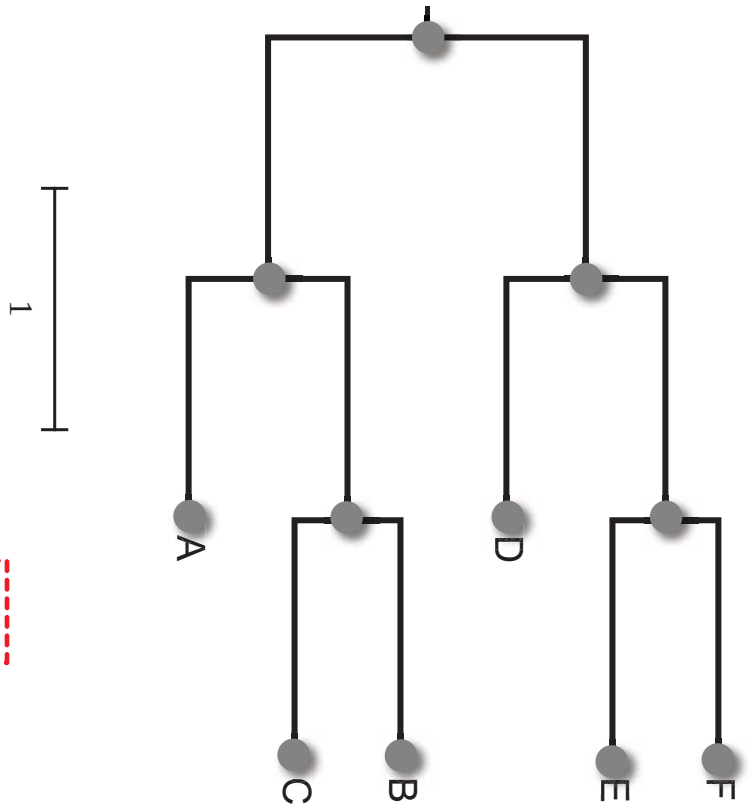
Topology
Branching order and pattern in the tree



Lineage
A cluster of taxa with more closely related evolutionary history

Terminology in a phylogenetic tree



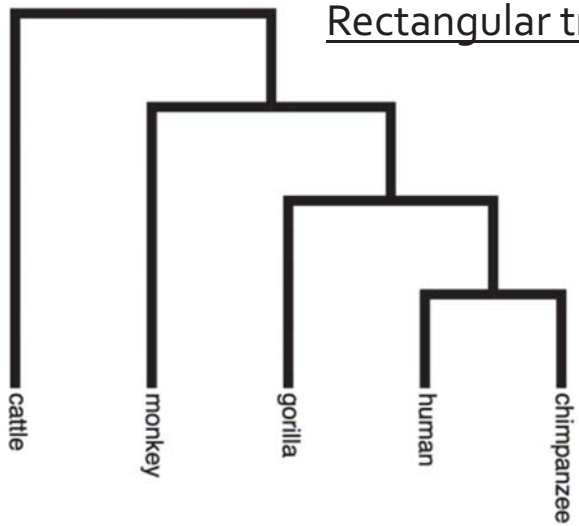


Node is rotatable

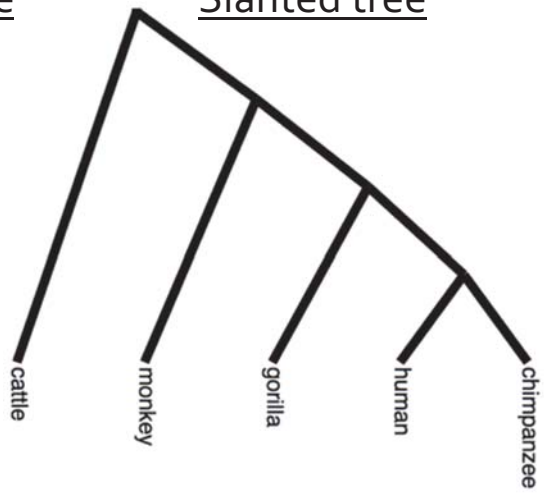
Node can be rotated without changing the biological meaning.

Different presentations of phylogeny

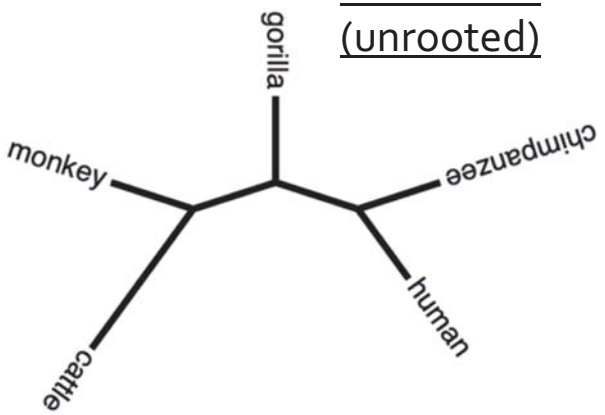
Rectangular tree

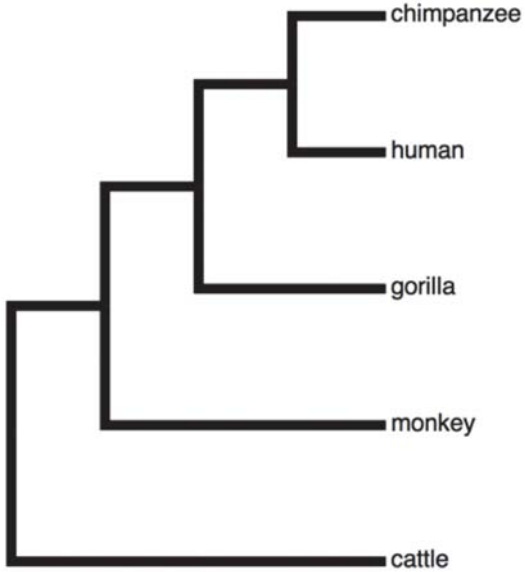


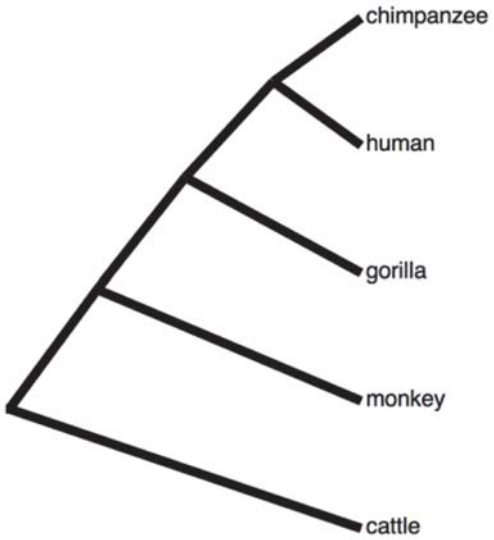
Slanted tree



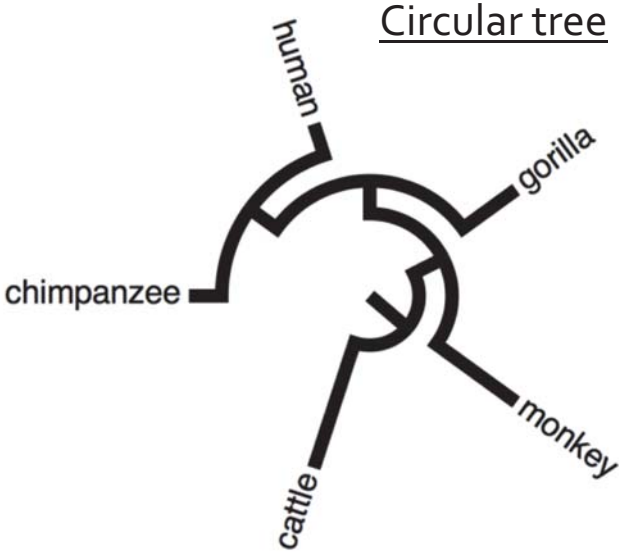
Radial tree (unrooted)







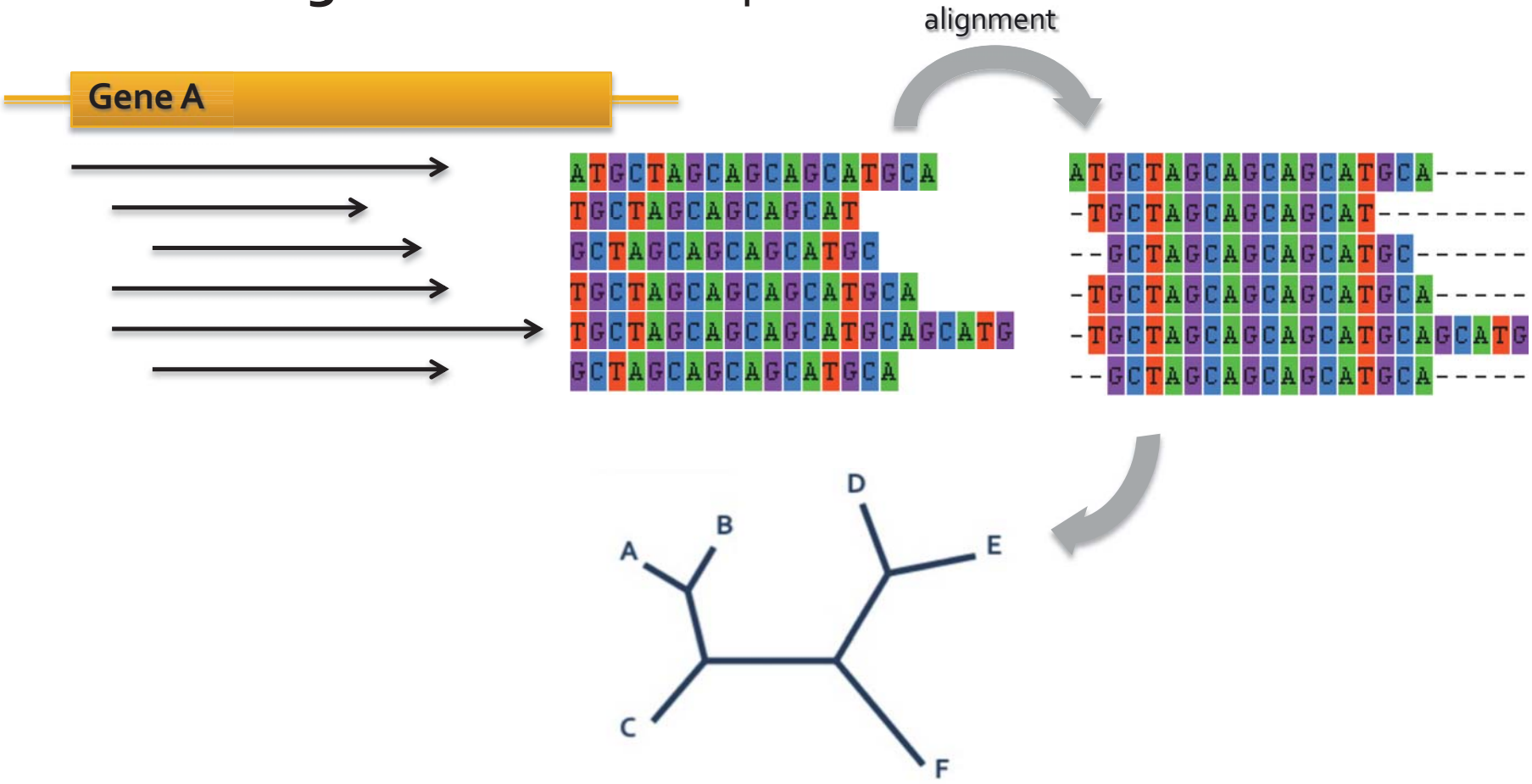
Circular tree



How to build a phylogenetic tree from the sequences of pathogen samples?

Raw sequences > Alignment > Tree

- Same/Similar region of the genome or gene is analyzed
- Need **alignment** of the sequences



Methods of Inferring Phylogeny

- **Distance based methods**
  - *UPGMA (Unweighted Pair Group Method with Arithmetic Mean)*

    Simple agglomerative hierarchical clustering, all tips same distance to the root
  - *Neighbor-joining*

    Divisive clustering (star-decomposition), into unrooted tree
  - *Minimum evolution*

    Tree with the smallest sum of branch lengths
- **Discrete-data based methods**
  - *Maximum parsimony*

    Searches for most parsimonious tree with the least evolutionary steps.
  - *Maximum likelihood*

    Searches for a tree (& substitution model) that may have the highest probability of observing the genetic sequence data.
  - *Bayesian inference*

    Generates posterior probability distributions for the tree model parameters, composed of tree & substitution models, based on the prior probabilities of the parameters and likelihood of the data.
- Different methods have pros and cons, in terms of accuracy, speed, memory-requirement, etc.

Measuring genetic distance

- **Example:**
  - **X: AATTTGTCCG**
    ↓↓
  - **Y: AATTTGTAAG**
- **Simplest genetic distance measure: P-distance**
  - P-distance (*p*) is the proportion of nucleotide sites at which two sequences being compared are different. It is obtained by dividing the number of nucleotide differences by the total number of nucleotides compared (i.e. length of alignment).
  - For the above example sequences:
  - 2 nucleotide difference (i.e. 2 substitutions),
  - 10 nucleotide sites in total
  - p_{XY} = 0.20 **substitutions/site**

Measuring genetic distance

- **Example:**
  - **X: AATTTGTCCG**

    ↓ ↓↓

    C T (multiple substitutions)

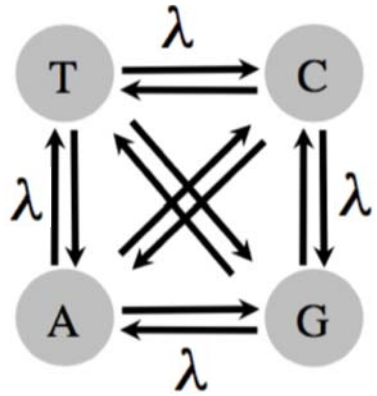
    ↓ ↓
  - **Y: AATTTGTAAG**
- Simple p-distance often under-estimates the number of substitutions because of multiple substitutions
- Multiple substitutions
  - Back substitutions, parallel substitutions, convergent substitution
  - E.g. actual substitutions = 5
  - Increased with the observed genetic difference between sequences
  - Can be accounted/corrected by using Continuous-Time-Markov-Chain (CTMC) model as the model of DNA evolution.

Jukes & Cantor(1969) one parameter model

- JC69 assumes equal substitution rate (λ) and equal nucleotide frequencies at equilibrium $\left(\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}\right)$

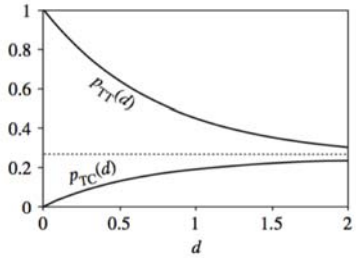
$$Q = \{q_{ij}\} = \begin{array}{c} \begin{matrix} \quad T & \quad C & \quad A & \quad G \end{matrix} \\ \begin{bmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{bmatrix} \end{array} \begin{matrix} T \\ C \\ A \\ G \end{matrix}$$

(equal instantaneous substitution rates among T, C, A, G)



$$P(t) = e^{Qt}$$

(transition probability at instantaneous rates of substitution after time *t*)



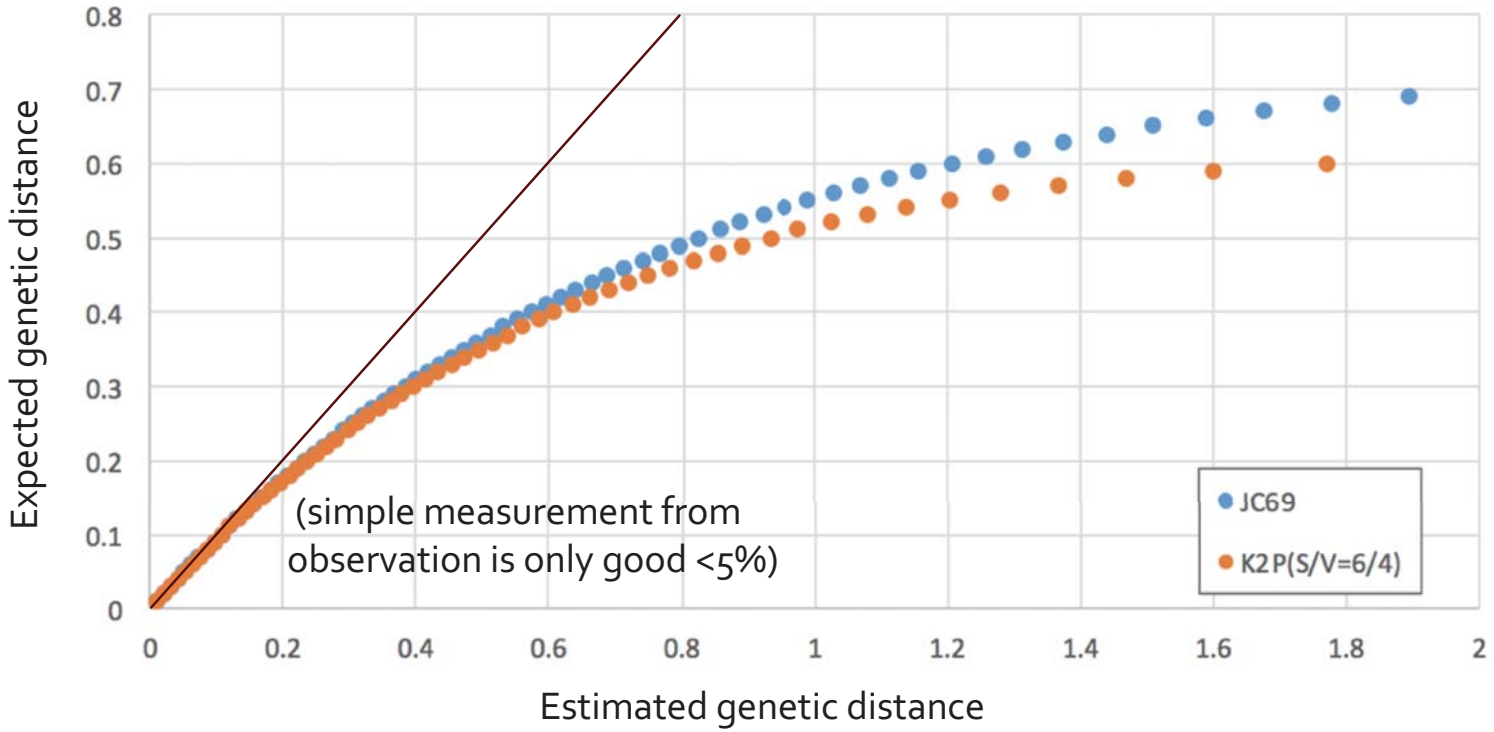
$$P(t) = e^{Qt} = \begin{bmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{bmatrix}, \quad \text{with} \begin{cases} p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}, & \text{(Probability that character is not changed)} \\ p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}. & \text{(Probability that character is changed)} \end{cases}$$

$$p = 3p_1(t) = \frac{3}{4} - \frac{3}{4}e^{-4\lambda t} = \frac{3}{4} - \frac{3}{4}e^{-4d/3}.$$

(p = the probability that the nucleotide in the descendant sequence is different from the nucleotide in the ancestral sequence is; and $d = 3\lambda t$)

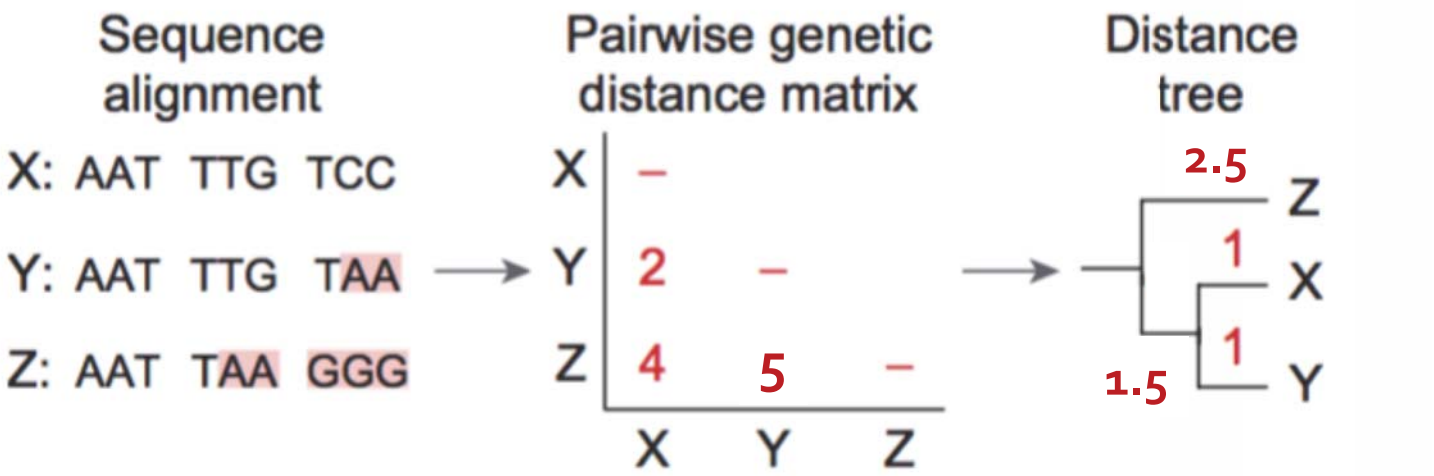
$$\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\hat{p}\right)$$

Observed distance and estimated distance



Distance based method – Overview

- Using the **genetic distances** measured between sequences to build the tree.
- Genetic distances can be determined by **p-distance** measurement or **substitution model** estimation.

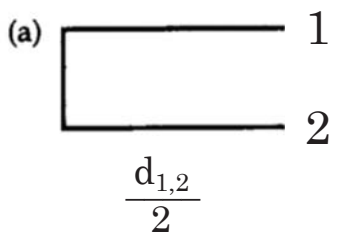


- The distance tree, with the shortest tree length, that reflect the genetic distances among the sequences

UPGMA *Unweighted Pair Group Method with Arithmetic Mean*

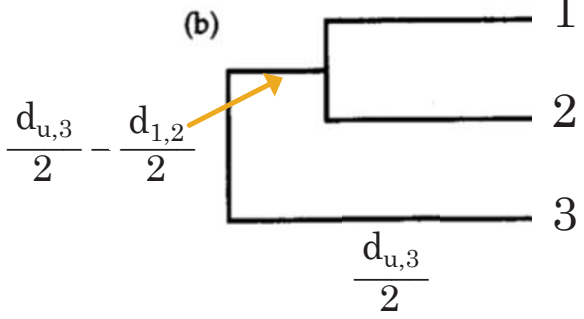
1. Choose the pair with the shortest distance to cluster, e.g. $d_{1,2}$; then the resultant branches share $d_{1,2}/2$

Taxon	1	2	3	4
2	d_{12}			
3	d_{13}	d_{23}		
4	d_{14}	d_{24}	d_{34}	
5	d_{15}	d_{25}	d_{35}	d_{45}



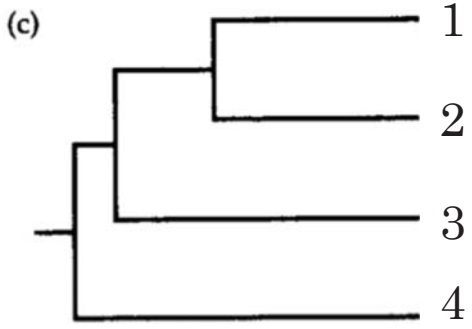
2. Recalculate the distance from the joined taxa (u=(1-2)) to other taxa (k) by $d_{u,k} = (d_{1,k} + d_{2,k})/2$
Then repeat step 1, e.g. $d_{u,3}$ is the shortest distance, so join taxon 3 with taxa 1 & 2.

Taxon	$u = (1\text{-}2)$	3	4
3	d_{u3}		
4	d_{u4}	d_{34}	
5	d_{u5}	d_{35}	d_{45}



Recalculate the distance e.g. $d_{v,4} = (d_{1,4}+d_{2,4}+d_{3,4})/3$

Taxon	$v = (1\text{-}2\text{-}3)$	4
4	d_{v4}	
5	d_{v5}	d_{45}



Maximum likelihood (ML)

- ML estimation is general statistical method for estimating unknown parameters of a probability model
- Likelihood (L) is defined as the probability of observing the data (**D**) when the model parameters (**M**) are given
- L = P(**D**|**M**)

- Simple coin-tossing example:

  HHTHHHHHTH (***D***)

  total 10 (***n***) toss, 8 (***k***) heads, 2 (***n***-***k***) tails

  What is the likelihood **L** = P(***D***|***p***,***n***)?

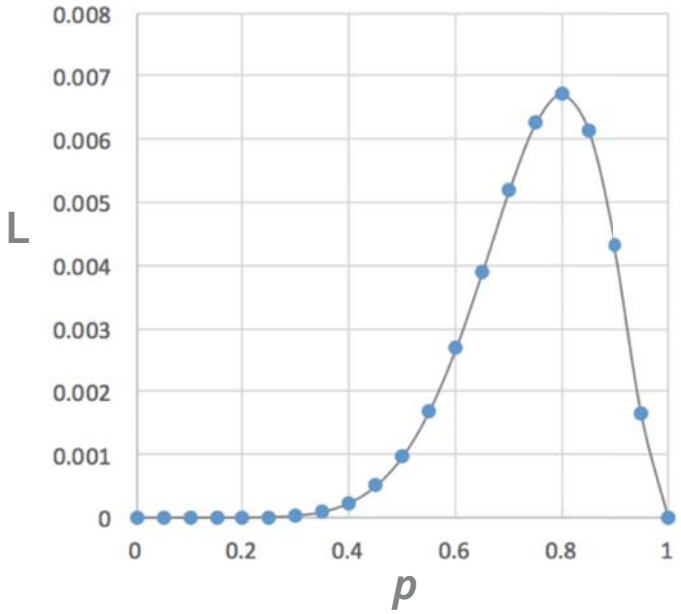
  p is the probability of get 'head' from a coin tossing.

  L = $p^k (1-p)^{n-k}$

  If a fair coin (***p***=0.5), **L** = 0.00098

  The maximum **L** is 0.00671 when ***p***=0.8

  Intutively, it's ***k***/***n***



Maximum likelihood (ML) for phylogeny

- To find a phylogeny that has maximized L = P(D|Θ)

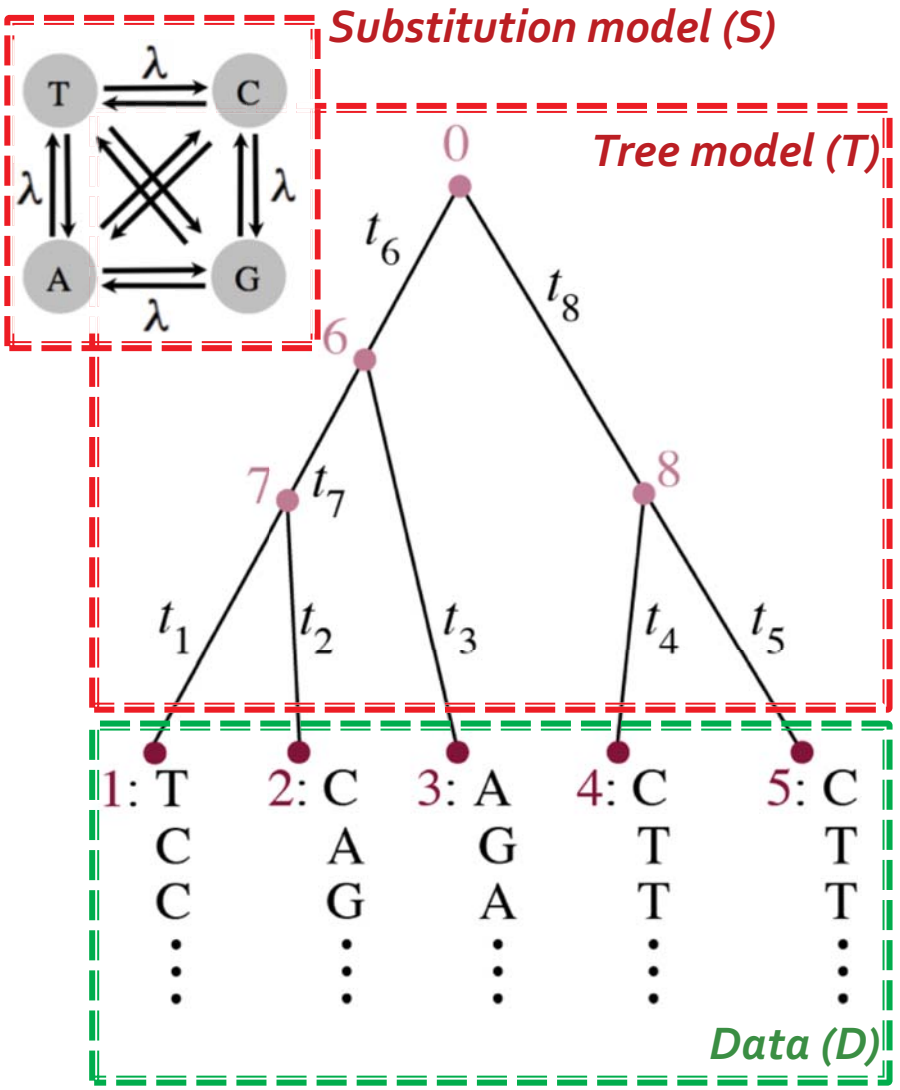
  Θ = T+S

  = $\{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, \lambda\}$

  T is the tree model with branch length = t_i.

  S is the DNA substitution model; Using JC69 model as an example, λ is the substitution rate.

- Efficiently **evaluate** the **likelihood L** of given Θ.
- Smartly search across the **tree and parameter space** to identify Θ that **maximize** L.



Evaluating likelihood of a phylogeny

- Assumption of **independent evolution among sites**

$$\ell = \log(L) = \sum_{h=1}^{n} \log\{f(\mathbf{x}_h|\theta)\}$$

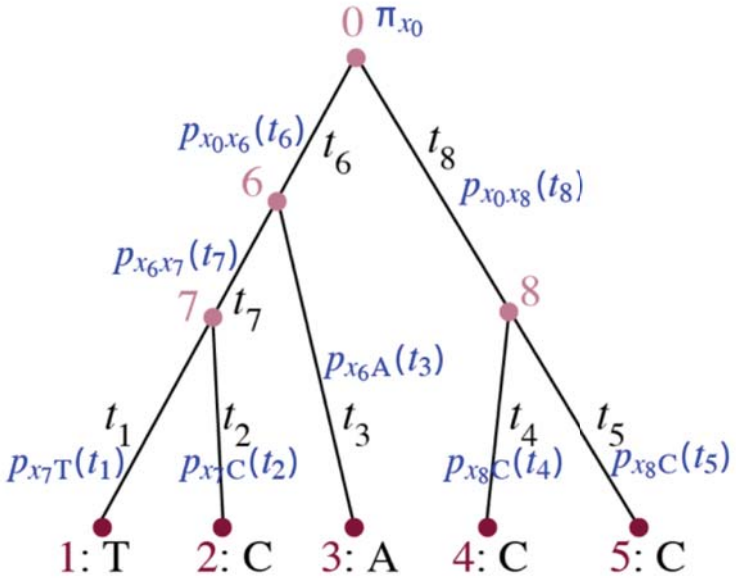
(n = number of sites)

The probability of the whole data set is the product of the probabilities of data at individual sites. Equivalently the log likelihood is a sum over sites in the sequence.

(any combination of ancestral nucleotides $x_0x_6x_7x_8$)

$$f(\mathbf{x}_h|\theta) = \sum_{x_0}\sum_{x_6}\sum_{x_7}\sum_{x_8}[\pi_{x_0}p_{x_0x_6}(t_6)p_{x_6x_7}(t_7)p_{x_7\mathrm{T}}(t_1)p_{x_7\mathrm{C}}(t_2) \times p_{x_6\mathrm{A}}(t_3)p_{x_0x_8}(t_8)p_{x_8\mathrm{C}}(t_4)p_{x_8\mathrm{C}}(t_5)].$$

So, now focus on one site only, X_h. $\mathbf{X}_i$ is the state at ancestral node i. $f(\mathbf{X}_h)$ is the sum over all possible nucleotide combinations (A,T,C,G) for the ancestors.

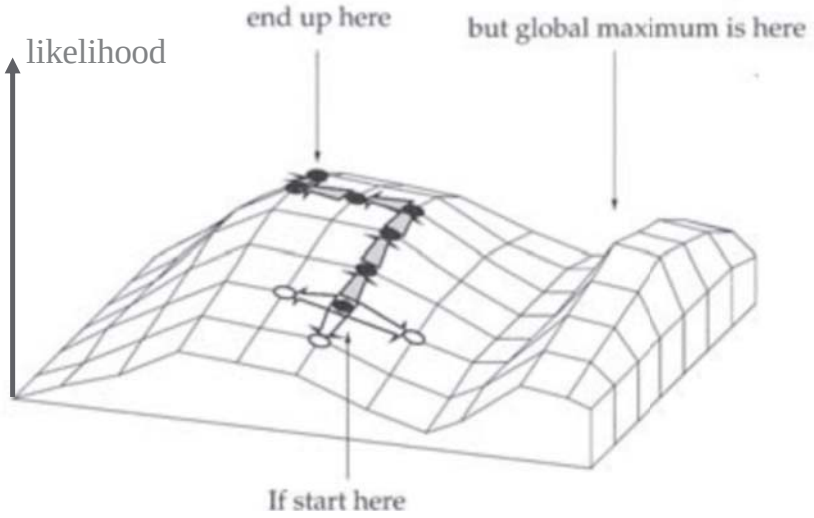


Brute-force computation takes $\mathbf{4}^{s-1} * (2s-2)$ calculations. **s** is the number of sequences. Here total = **2048**. In fact, many calculations are repeats.

Felsenstein used *pruning* algorithm (1973, 1981) that calculate successively probabilities (partial likelihoods) of data on many subtrees. It will reduce the computation steps to 4(5**s**-8) = **68** in this example.

Search in tree and parameter space

- Parameter space
  - **Branch lengths** in the tree (e.g. t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8) and **parameters** (e.g. λ) in the **substitution model**
  - **Numerical optimization** on a fixed tree
  - **Multivariate** algorithms such as BFGS
- Tree space
  - Huge, grow **factorial** with **number of sequences (*n*)** $N_{unrooted} = \frac{(2n-5)!}{2^{n-3}(n-3)!}$
  - E.g. 5 sequences = 15 possible topologies
  - E.g. 14 sequence > 3×10^{11} possible topologies
  - Heuristic approach
    - Branch swapping
      - Nearest-Neighbor Interchange (NNI)
      - Tree bisection and reconnection (TBR)
      - Subtree Pruning and Re-grafting (SPR)



Use of phylogeny for studying infectious disease

- **Qualitative interpretations**
  - Based on the **clustering**, **tree topology/shape**, **branch lengths** to give **qualitative** interpretation on the disease origin and transmission
- **Quantitative inferences**
  - Co-estimating **epidemic parameters** (e.g. *epidemic starting time*, *growth rate*) with the phylogeny
- **Hypothesis testing**
  - **Testing** the consistency between the hypothesized transmission history and the inferred phylogeny