Northwestern Medicine®

Center for Pathogen Genomics and Microbial Evolution

Robert J. Havey Institute for Global Health

# Introductory Phylogenetic Analysis

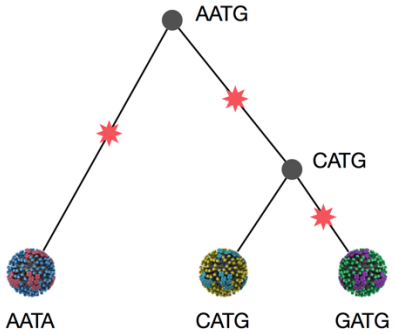
## IMS Workshop

Ramon Lorenzo-Redondo, Ph.D.

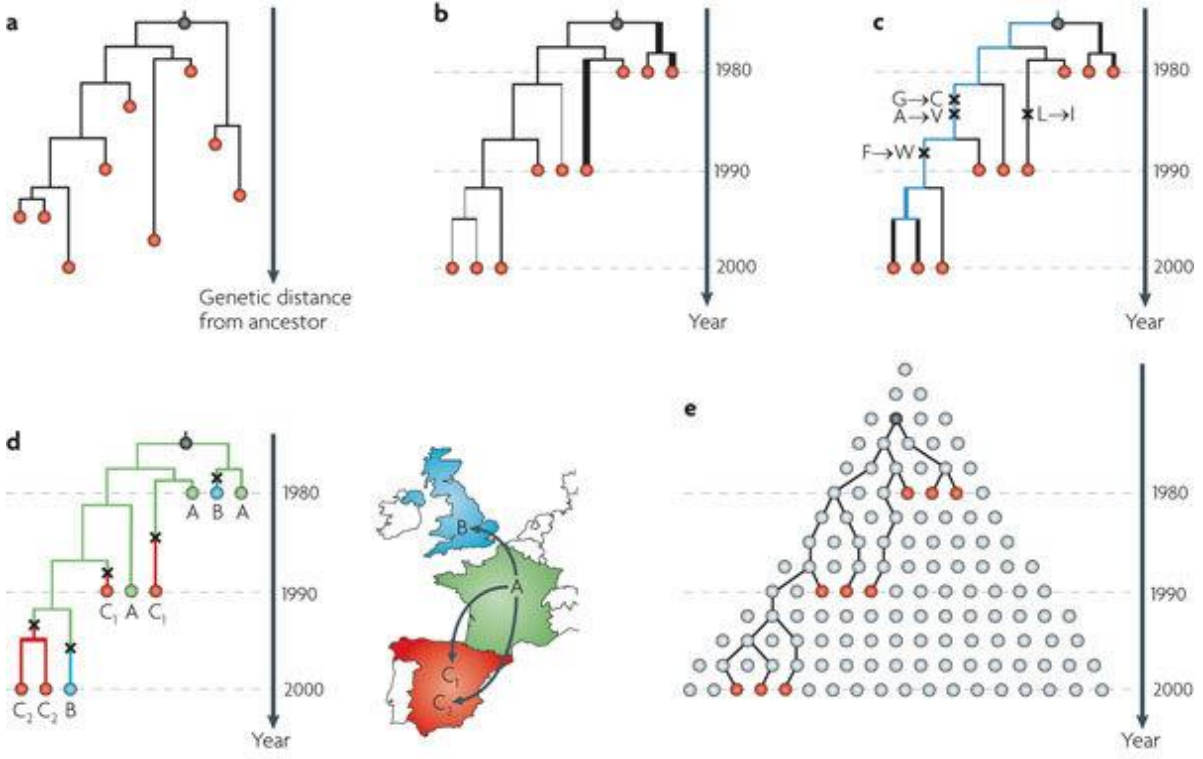
Assistant Professor of Medicine, Division of Infectious Diseases

Bioinformatics Director, Center for Pathogen Genomics and Microbial Evolution (CPGME)

# Genome Sequencing can provide insights into the molecular epidemiology of pathogens



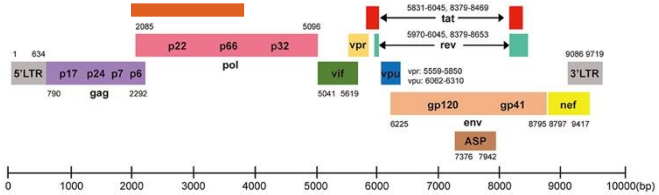

Bedford, T.



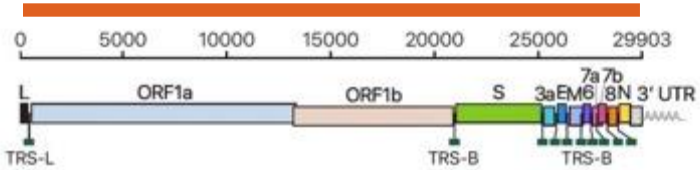

**Pybus, O., Rambaut, A.** *Nat Rev Genet* **10,** 540–550 (2009)

Northwestern Medicine®

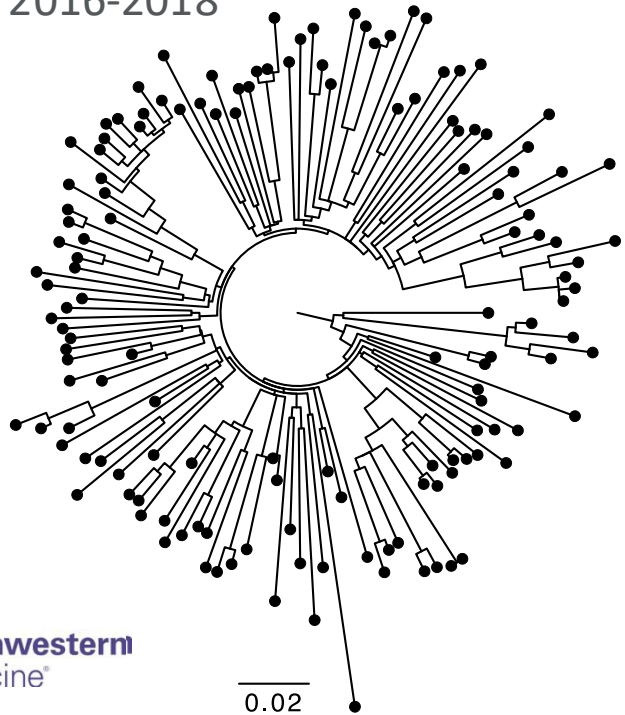
# The context is important (e.g., HIV-1 vs SARS-CoV-2)




HIV-1 Chicago 2016-2018

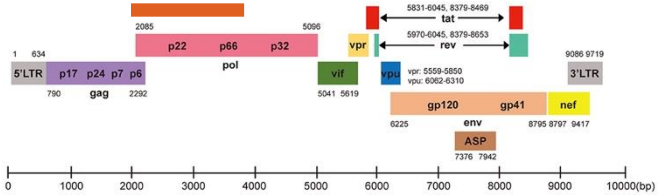


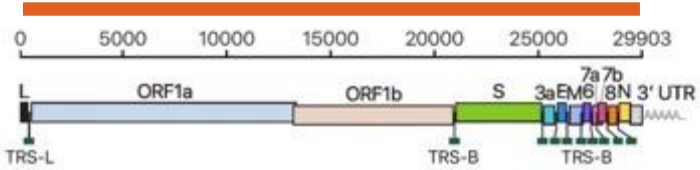

SARS-CoV-2 Chicago 2020-2021



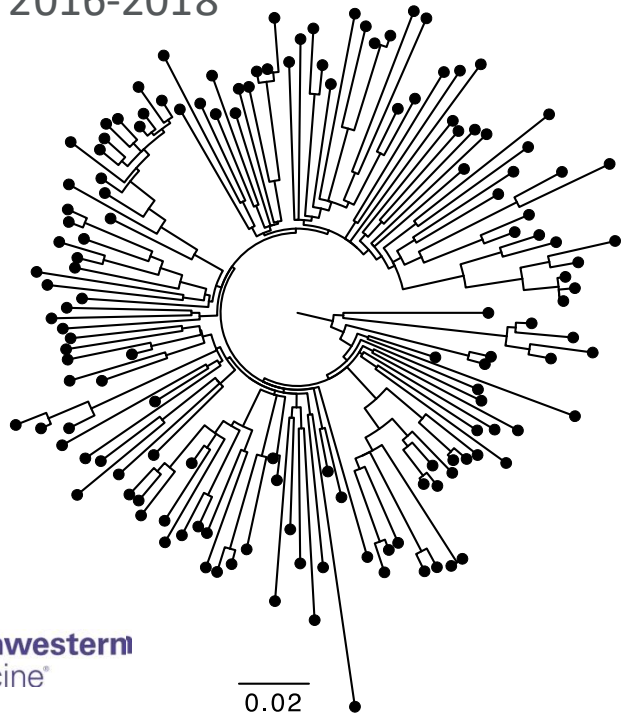

Northwestern Medicine®

# The context is important (e.g., HIV-1 vs SARS-CoV-2)

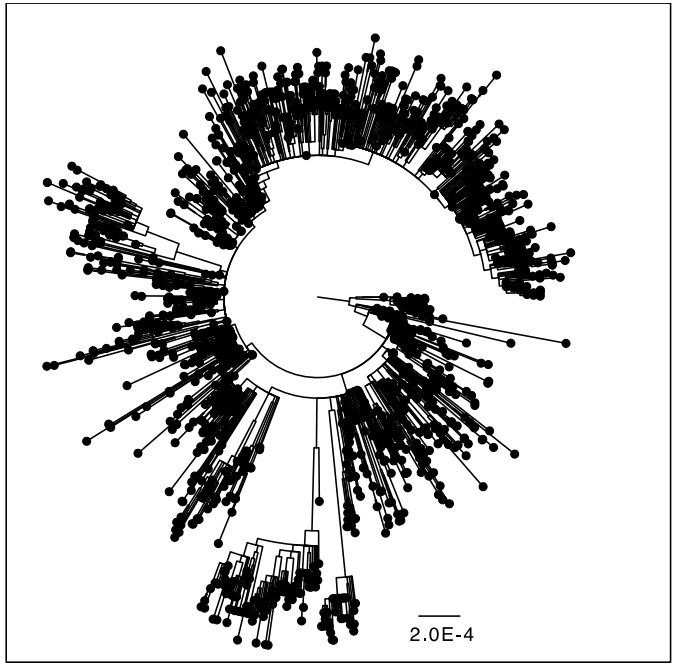



HIV-1 Chicago 2016-2018

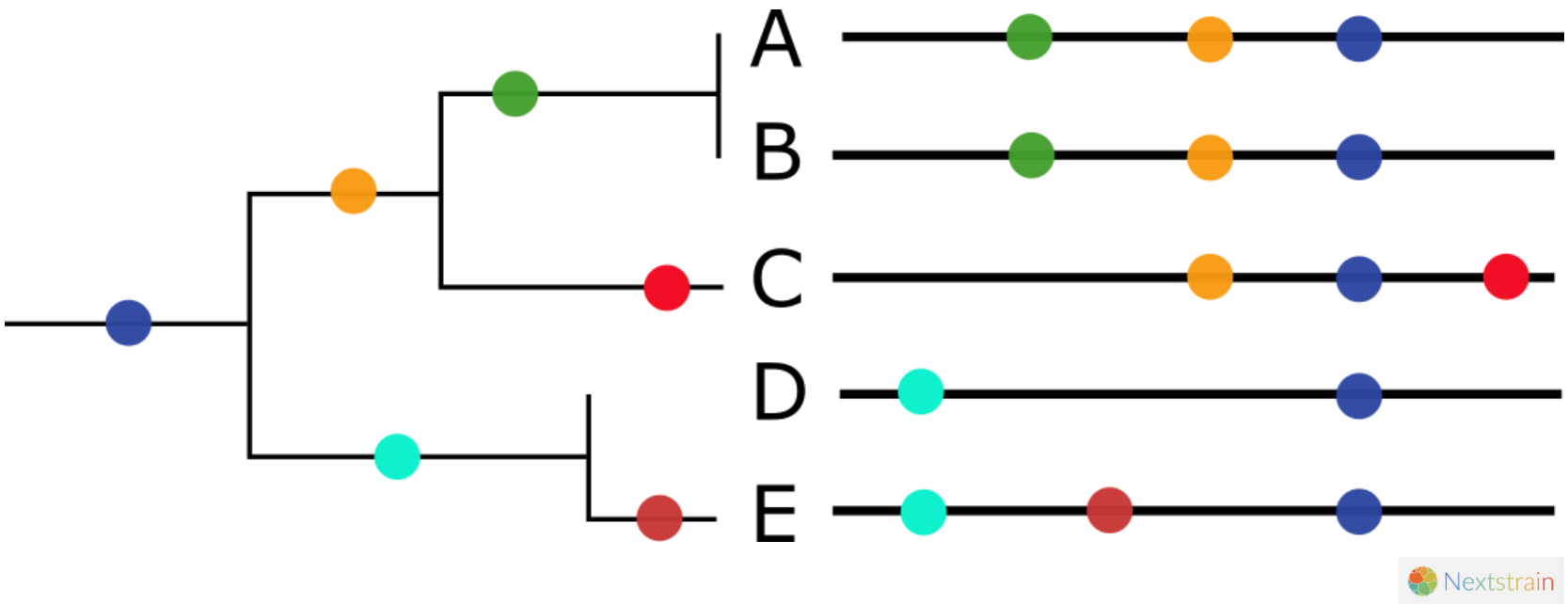



SARS-CoV-2 Chicago 2020-2021



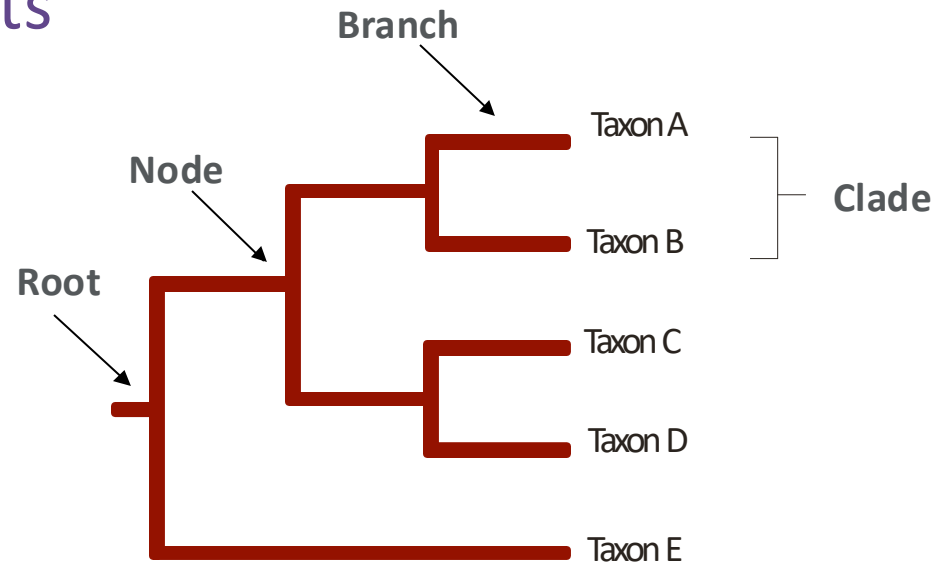

Northwestern Medicine®

# What does a phylogenetic tree represent?

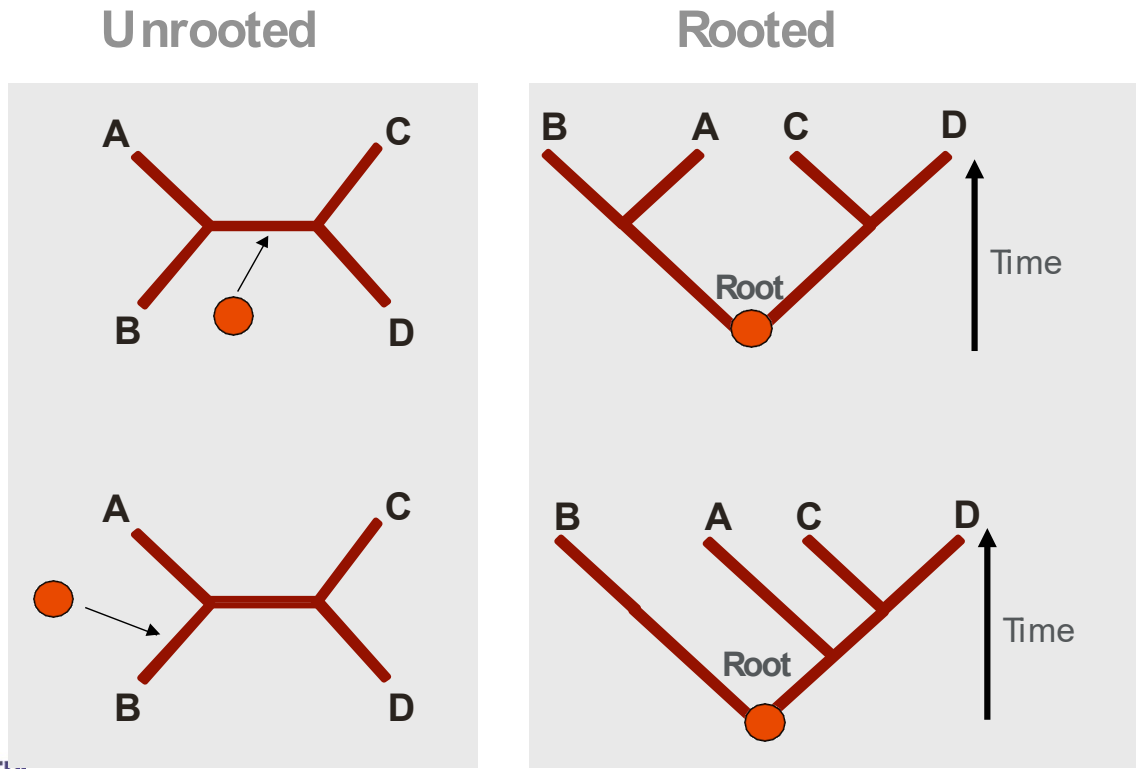


# Phylogenetic Tree Components

- **Taxon**: elements whose relationships we are studying. Can be species, groups, genes, alleles.
- **Node**: branch ramification point (likely an ancestral taxon).
- **Branch**: defines relationships between taxa according to descending ancestors
- **Topology:** branching pattern
- **Branch length**: represents number of changes or change probability. **Root**: most recent common ancestor (MRCA).
- **Clade**: group of taxa that includes the common ancestor and all descendants.



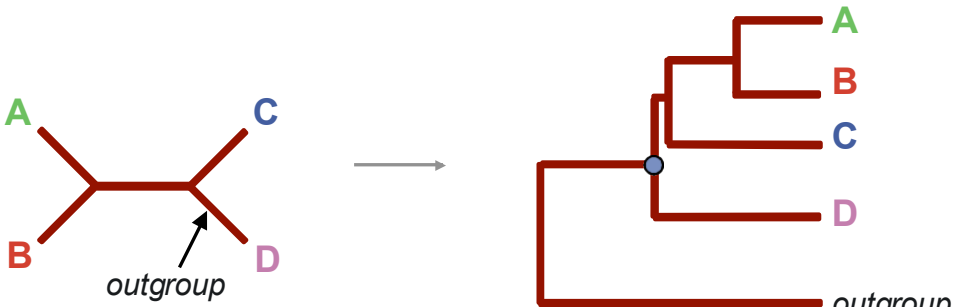
# Types of Trees



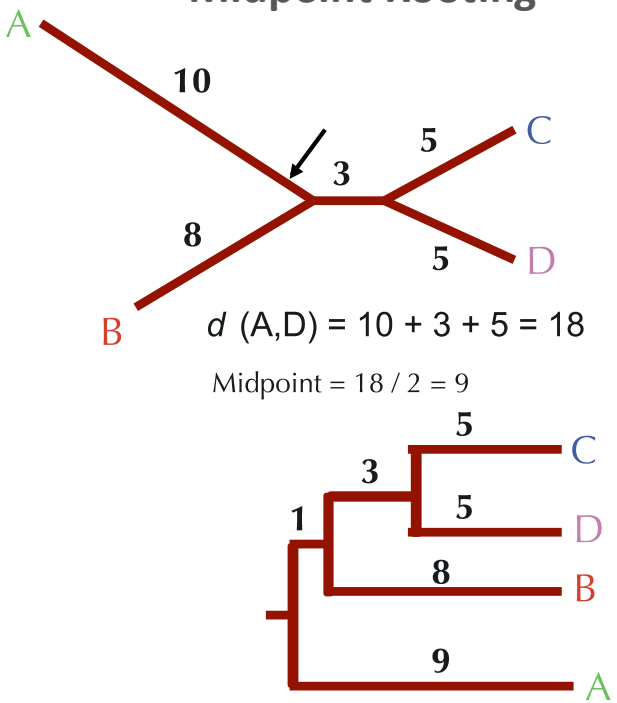

Northwestern Medicine®

# Rooting methods

## Outgroup Rooting




## Midpoint Rooting



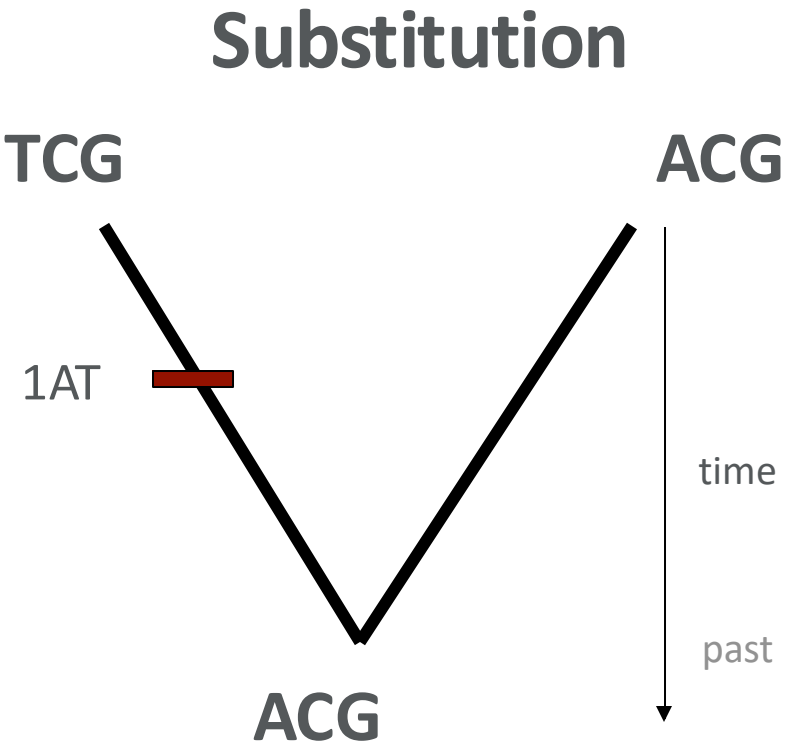
Northwestern Medicine®

# Alignment

# Alignment

```
TCG
| |
ACG
```

- 1 differences
- 2 matches

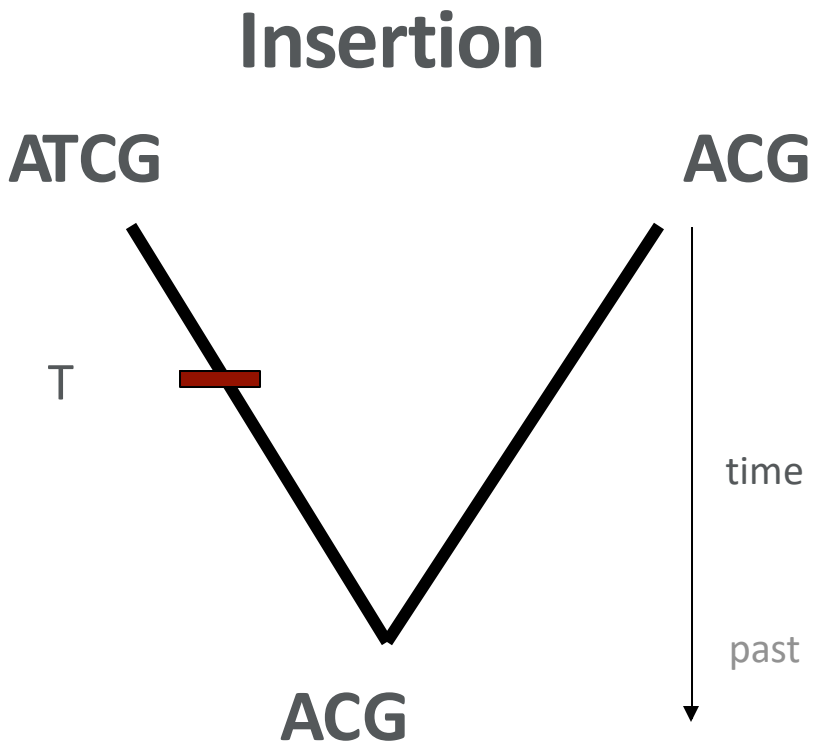



Northwestern Medicine®

# Alignment

```
ATCG
|  ||
A-CG
```

- 0 differences
- 3 matches
- 1 gap

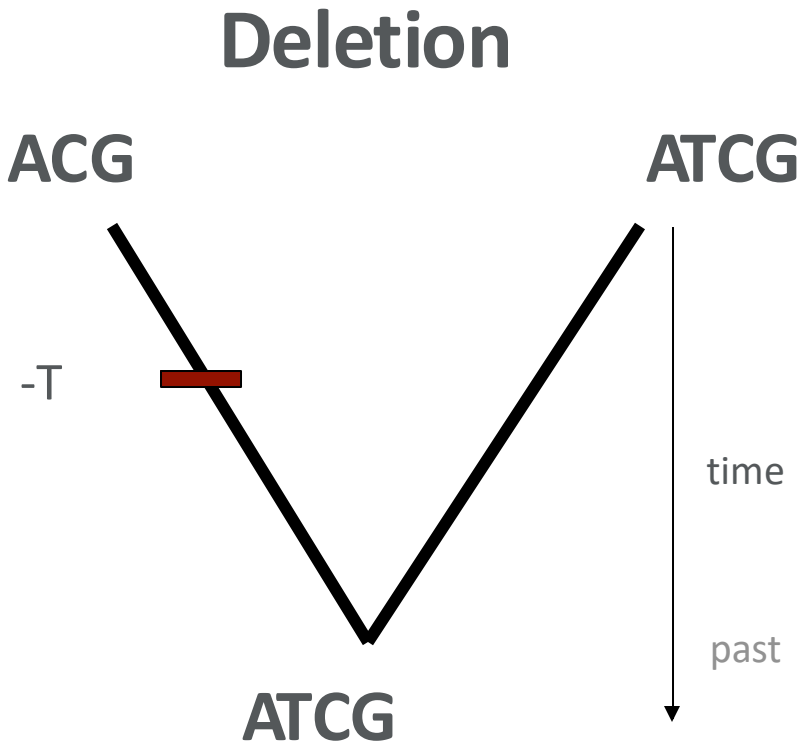



Northwestern Medicine®

# Alignment

```
ATCG
|  ||
A-CG
```

- 0 differences
- 3 matches
- 1 gap




Northwestern Medicine®

# Optimal Alignment

- Optimal alignment is the one that minimizes differences and gaps

```
(I)     T C A G - A C G - A T T G        0 differences
        | |   |   |   |     |   |        7 matches
        T C - G G A - G C - T - G        6 gaps

(II)    T C A G A C G A T T G            5 differences
        | | * | | * * * *                4 matches
        T C G G A G C T G - -            1 gap

(III)   T C A G - A C G A T T G          2 differences
        | |   |   |   | * | *            6 matches
        T C - G G A - G C T G -          4 gaps
```

- …but depends on the cost of events

# Alignment Penalties

- To compare gaps and mismatches:
  - Gap penalty
  - Mismatch penalty

- Dissimilarity Index:

  $y_i$ = number of changes type *i*

  $m_i$ = penalty of changes type *i*

  $z_k$ = number of gaps length *k*

  $w_k$ = length *k* gap penalty

$$D = \sum m_i y_i + \sum w_k z_k$$

# Nucleotide substitution penalties

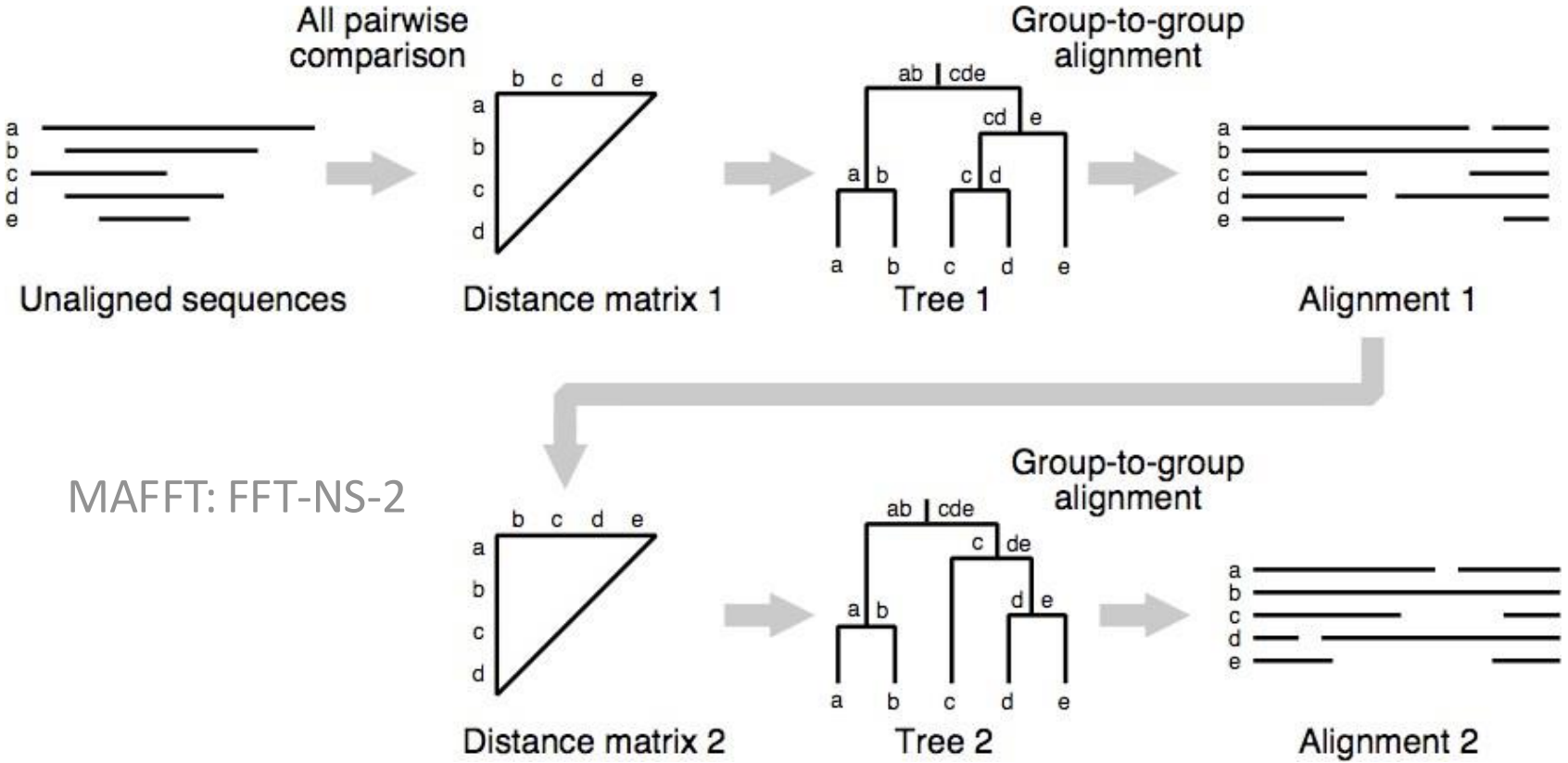
There can be different costs for different nucleotide substitutions, for example:

| | A | C | G | T |
|---|---|---|---|---|
| A | 0 | 1 | 1 | 1 |
| C | | 0 | 1 | 1 |
| G | | | 0 | 1 |
| T | | | | 0 |

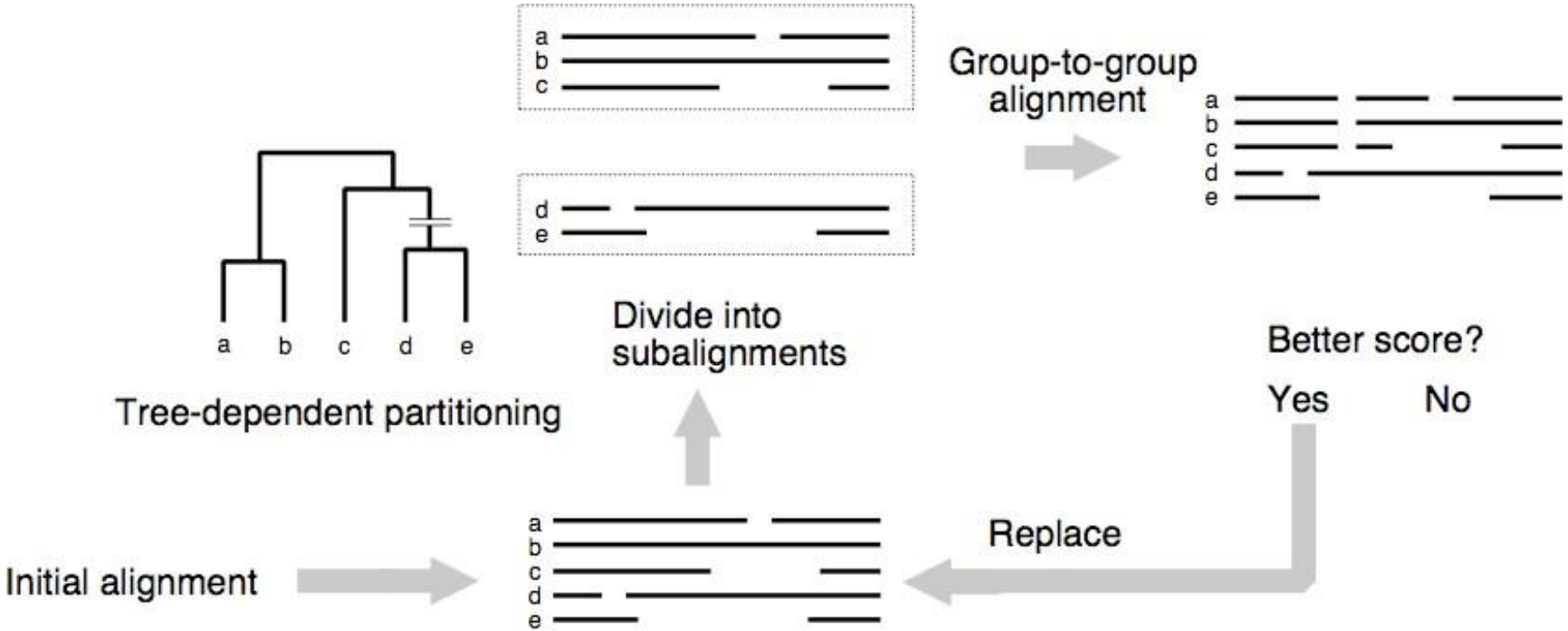
| | A | C | G | T | A |
|---|---|---|---|---|---|
| C | 0 | 2 | 1 | 2 | |
| G | | 0 | 2 | 1 | |
| T | | | 0 | 2 | |
| | | | | 0 | |

Northwestern Medicine®

# Progressive Alignment with MAFFT



# Iterative Refinement in MAFFT



Northwestern Medicine®

Northwestern Medicine®

# Tree Building

# Phylogenetic Tree Inference

**Table 1.5** Classification of phylogenetic analysis methods and their strategies

| | Optimality search criterion | Clustering |
|---|---|---|
| **Character state** | Maximum parsimony (MP) | |
| | Maximum likelihood (ML) | |
| | Bayesian inference | |
| **Distance matrix** | Fitch–Margoliash | UPGMA |
| | | Neighbor-joining (NJ) |

# Phylogenetic Tree Inference

## Types of data used in phylogenetic inference:

**Character-based methods:** Use the aligned characters, such as DNA or protein sequences, directly during tree inference.

| Taxa | Characters |
|---|---|
| Species A | ATGGCTATTCTTATAGTACG |
| Species B | ATCGCTAGTCTTATATTACA |
| Species C | TTCACTAGACCTGTGGTCCA |
| Species D | TTGACCAGACCTGTGGTCCG |
| Species E | TTGACCAGTTCTCTAGTTCG |

**Distance-based methods:** Transform the sequence data into pairwise distances (dissimilarities), and then use the matrix during tree building.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| Species A | ---- | 0.20 | 0.50 | 0.45 | 0.40 |
| Species B | 0.23 | ---- | 0.40 | 0.55 | 0.50 |
| Species C | 0.87 | 0.59 | ---- | 0.15 | 0.40 |
| Species D | 0.73 | 1.12 | 0.17 | ---- | 0.25 |
| Species E | 0.59 | 0.89 | 0.61 | 0.31 | ---- |

Example 1: Uncorrected "p" distance (=observed percent sequence difference)

Example 2: Kimura 2-parameter distance (estimate of the true number of substitutions between taxa)

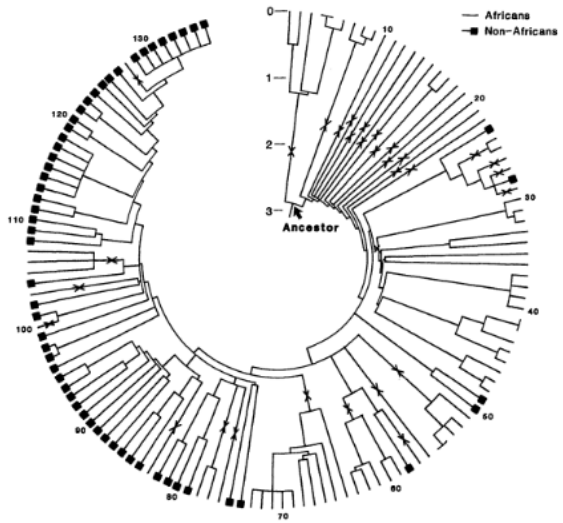
# How many trees are there?

For $n$ species there are

$$T_n = 1 \times 3 \times 5 \times \cdots \times (2n-3) = \frac{(2n-3)!}{(n-2)!2^{n-2}}$$

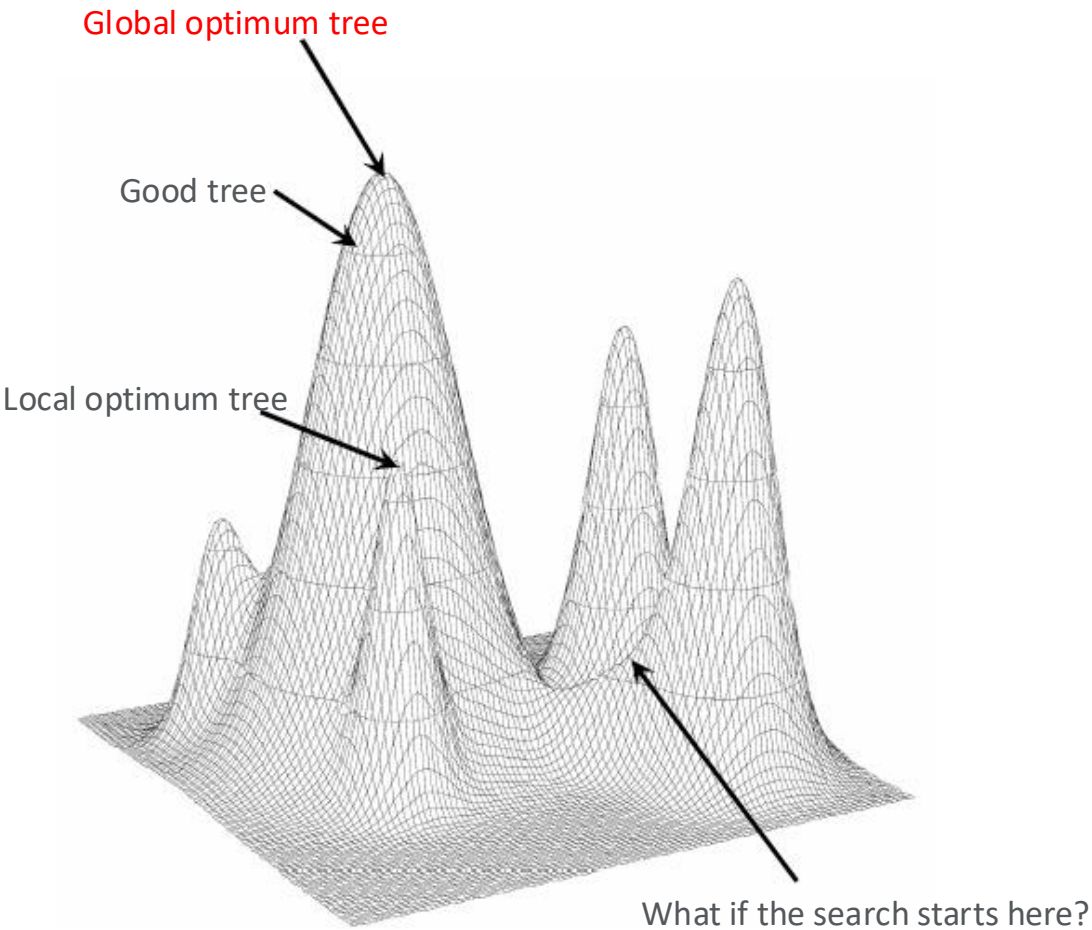
rooted, tip-labelled binary trees:

| $n$ | #trees | |
|---|---|---|
| 4 | 15 | enumerable by hand |
| 5 | 105 | enumerable by hand on a rainy day |
| 6 | 945 | enumerable by computer |
| 7 | 10395 | still searchable very quickly on computer |
| 8 | 135135 | about the number of hairs on your head |
| 9 | 2027025 | greater than the population of Auckland |
| 10 | 34459425 | $\approx$ upper limit for exhaustive search |
| 20 | $8.20 \times 10^{21}$ | $\approx$ upper limit of branch-and-bound searching |
| 48 | $3.21 \times 10^{70}$ | $\approx$ the number of particles in the Universe |
| 136 | $2.11 \times 10^{267}$ | number of trees to choose from in the "Out of Africa" data (Vigilant *et al.* 1991) |



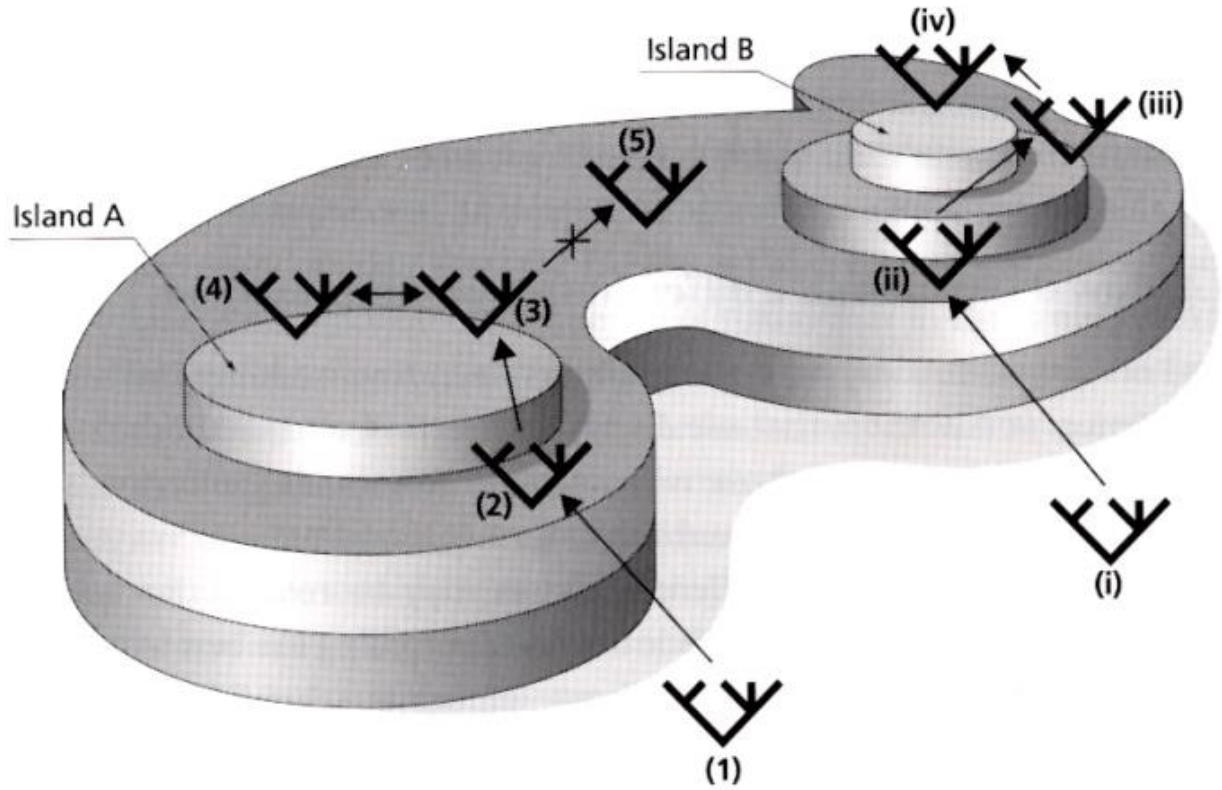

Northwestern Medicine

# Trees Landscape




Northwestern Medicine®

# Searching Movements




Northwestern Medicine®

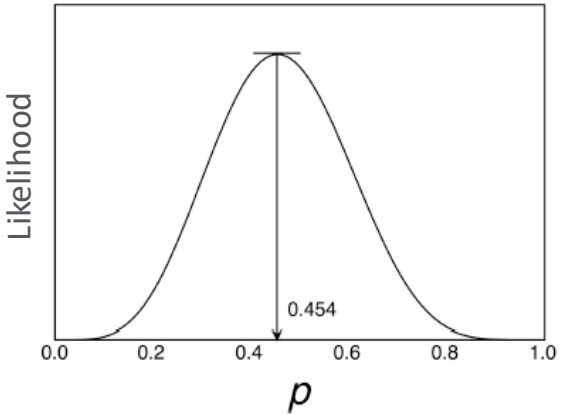
# MAXIMUM LIKELIHOOD (ML)

$L = P(D|H) =$ Probability of the data given a hypothesis

If we throw a coin 11 times and we obtain:


What's the expected probability *p* to obtain  when we throw the coin?

$$L = P(D \mid p) = pp(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p) = p^5(1-p)^6$$




$$\frac{dL}{dp} = \left(\frac{5}{p} - \frac{6}{1-p}\right) p^5(1-p)^6 = 0$$

$$5 - 11p = 0$$

$$\hat{p} = \frac{5}{11} = 0.454$$

$$\ln L = 5\ln p + 6\ln(1-p)$$

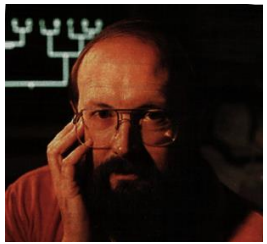
$$\frac{d(\ln L)}{dp} = \frac{5}{p} - \frac{6}{1-p} = 0$$

$$5 - 11p = 0$$

$$\hat{p} = \frac{5}{11} = 0.454$$

# ML in phylogenies

- It evaluates a hypothesis about evolutionary history in terms of the probability that the proposed model and the hypothesized history would give rise to the observed data set.
- The supposition is that a history with a higher probability of reaching the observed state is preferred to a history with a lower probability.
- The method searches for the tree with the **highest probability or likelihood**.



Joe Felsenstein

Northwestern Medicine®

# ML in phylogenies

The Likelihood (L) is proportional to the probability of the data (D) given an evolutionary model (M), a vector θ of K parameters of the evolutionary model, topology τ and a vector ν of tree lengths:

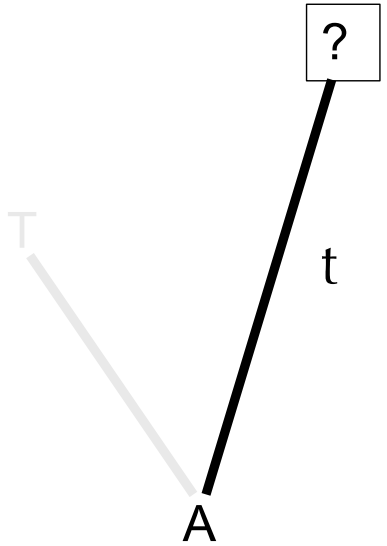
$$L = P(D|M,\theta,\tau,\nu)$$

# Substitution Rate

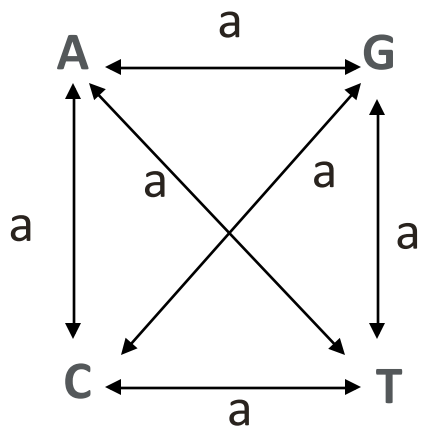
- Substitution probability along a branch of length ***t*** (μ × time)

$$P_t = e^{Qt}$$

$$P_t = \begin{bmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{bmatrix}$$

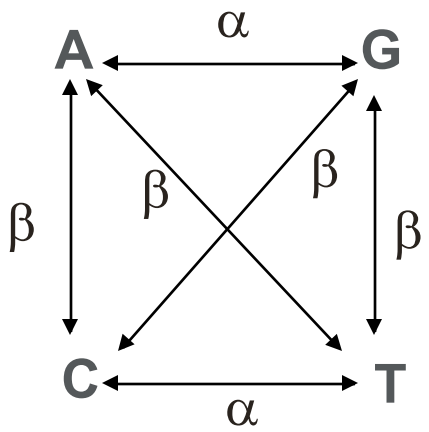


# Jukes and Cantor (JC69)




$$\mathbf{P}_t = \begin{bmatrix} . & \alpha & \alpha & \alpha \\ \alpha & . & \alpha & \alpha \\ \alpha & \alpha & . & \alpha \\ \alpha & \alpha & \alpha & . \end{bmatrix}, \qquad \mathbf{f} = [\tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}]$$

# Kimura 2 parameters (K80)

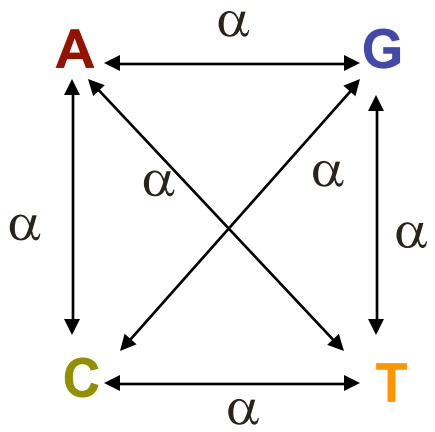



$$\mathbf{P}_t = \begin{bmatrix} . & \beta & \alpha & \beta \\ \beta & . & \beta & \alpha \\ \alpha & \beta & . & \beta \\ \beta & \alpha & \beta & . \end{bmatrix}, \qquad \mathbf{f} = [\tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}].$$

$$ti : tv = \kappa$$

# Evolutionary Models

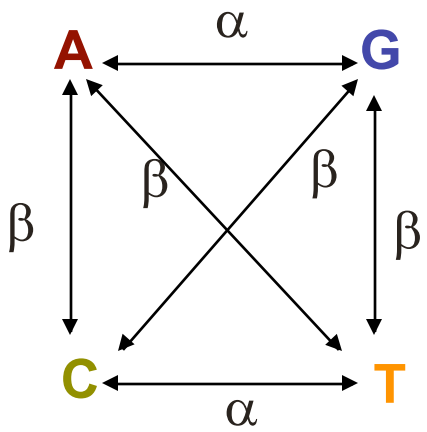
## Felsenstein 1981 (F81)




$$\mathbf{P}_t = \begin{bmatrix} . & \pi_C\alpha & \pi_G\alpha & \pi_T\alpha \\ \pi_A\alpha & . & \pi_G\alpha & \pi_T\alpha \\ \pi_A\alpha & \pi_C\alpha & . & \pi_T\alpha \\ \pi_A\alpha & \pi_C\alpha & \pi_G\alpha & . \end{bmatrix}, \qquad \mathbf{f} = [\pi_A\ \pi_C\ \pi_G\ \pi_T]$$

# Evolutionary Models

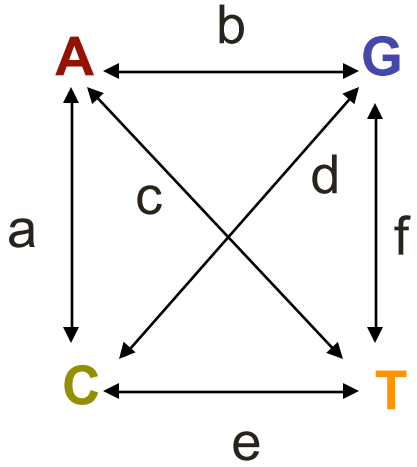
## Hasegawa-Kishino-Yano (HKY85)




$$\mathbf{P}_t = \begin{bmatrix} . & \pi_C\beta & \pi_G\alpha & \pi_T\beta \\ \pi_A\beta & . & \pi_G\beta & \pi_T\alpha \\ \pi_A\alpha & \pi_C\beta & . & \pi_T\beta \\ \pi_A\beta & \pi_C\alpha & \pi_G\beta & . \end{bmatrix}, \qquad \mathbf{f} = [\pi_A\ \pi_C\ \pi_G\ \pi_T]$$

# Evolutionary Models

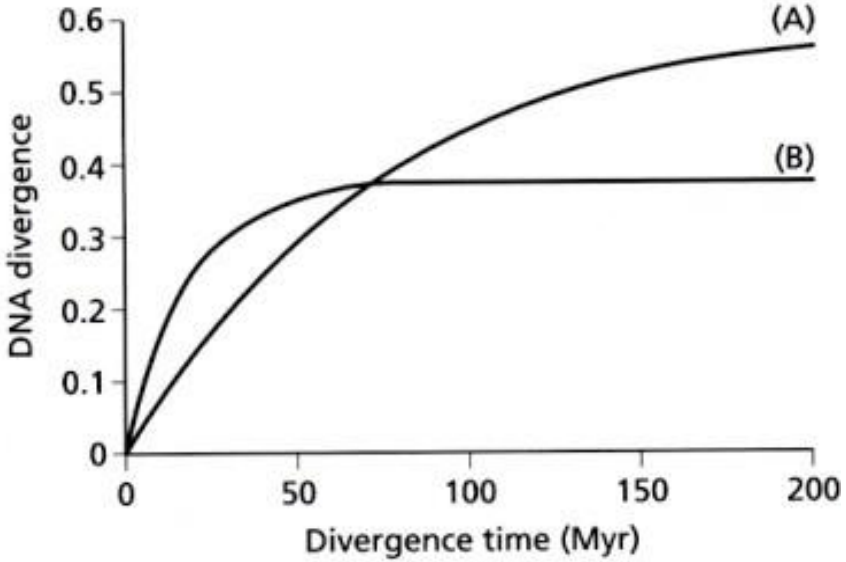
## General time reversible (GTR or REV)



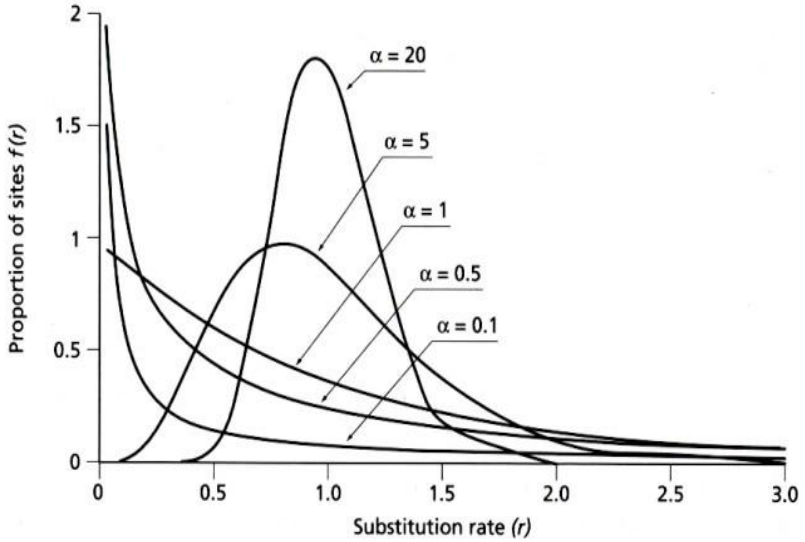

$$\mathbf{P}_t = \begin{bmatrix} . & \pi_C a & \pi_G b & \pi_T c \\ \pi_A a & . & \pi_G d & \pi_T e \\ \pi_A b & \pi_C d & . & \pi_T f \\ \pi_A c & \pi_C e & \pi_G f & . \end{bmatrix}, \qquad \mathbf{f} = [\pi_A \ \pi_C \ \pi_G \ \pi_T]$$

# Additional parameters

- **Proportion of invariant sites (*p-inv*)**

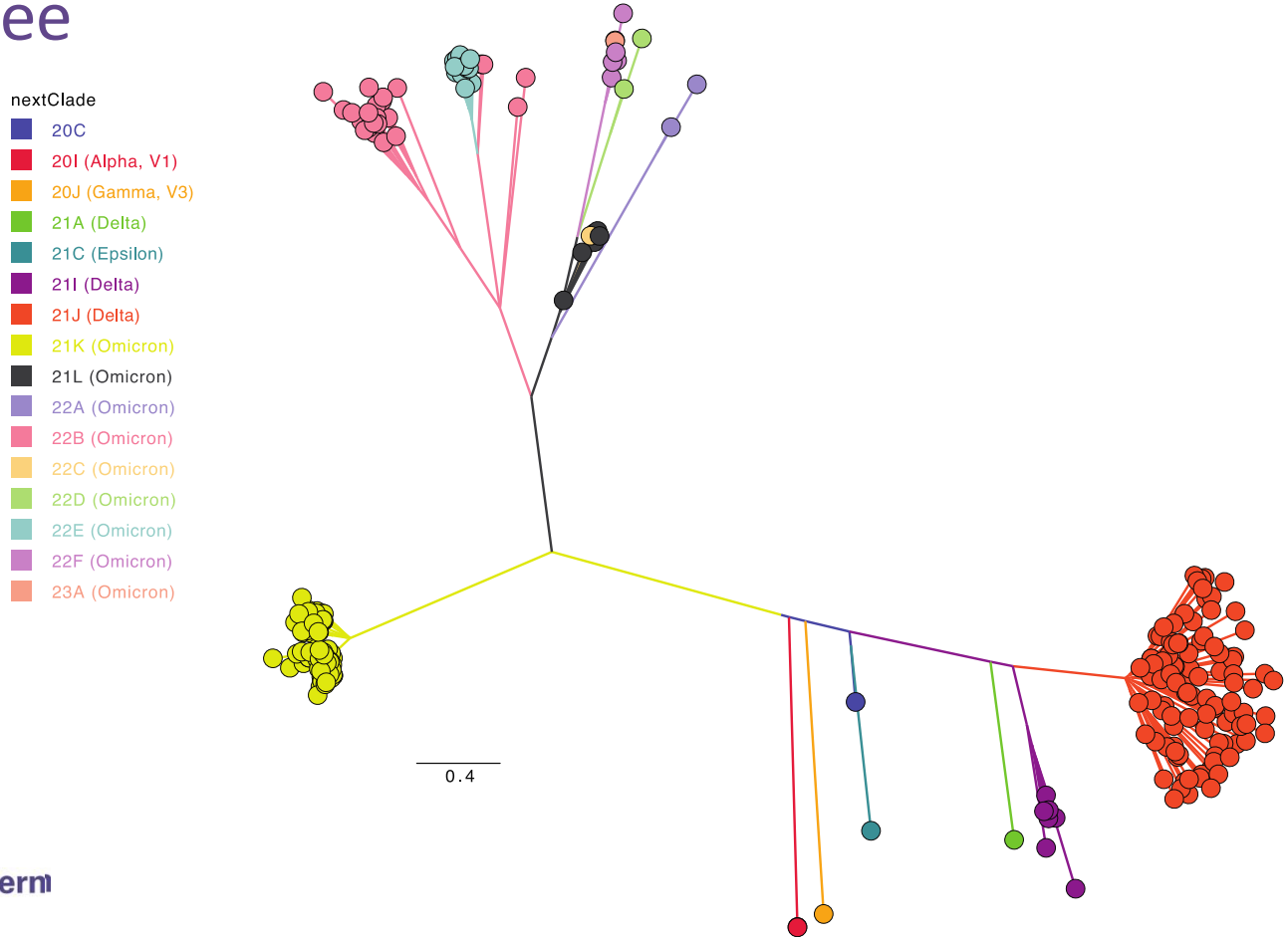



- **Gamma distribution ($\Gamma$) (alpha)**



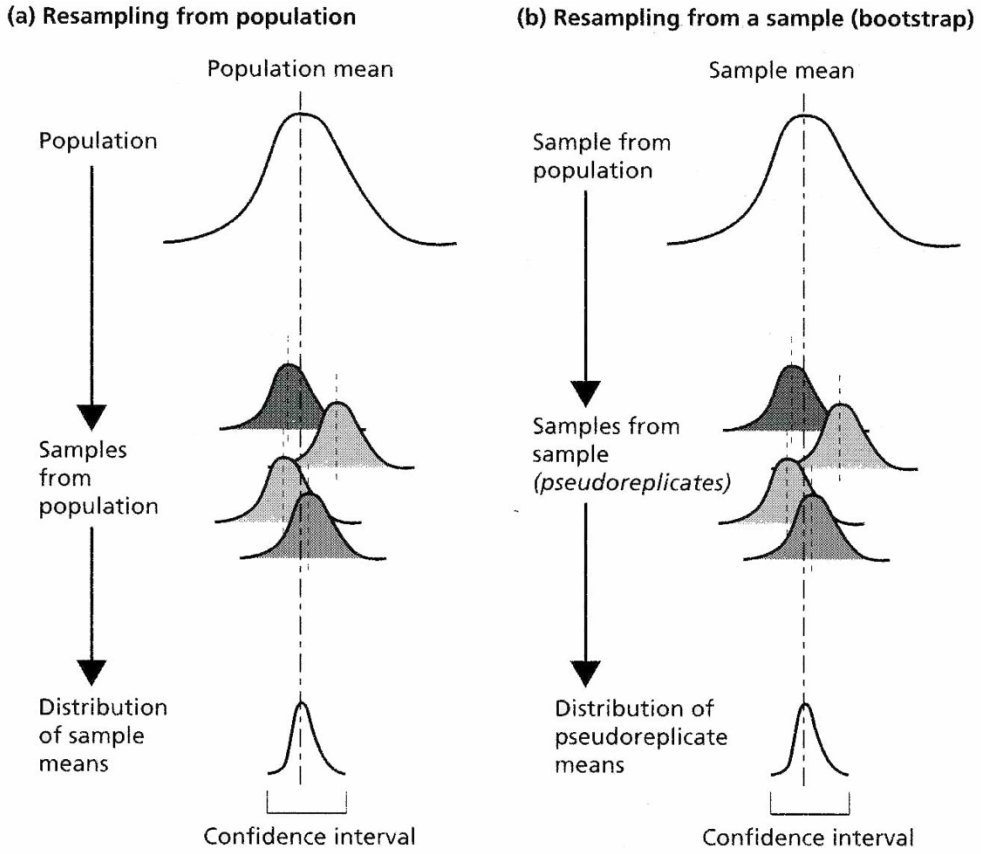

Northwestern Medicine®

# ML Tree




Northwestern Medicine®

# Bootstrap




*Baron Munchausen tells, "In pursuit of a hare, I wanted to set my horse over a swamp … [I jumped] … too short and fell into the mud not far from the other bank up to my neck. Here I would have died infallibly if the strength of my arm had not pulled me out again by my own plait of hair, together with the horse, which I locked firmly between my knees.* ***Rudolf Erich Raspe***

Northw
Medicir

# Bootstrapping

- Used to generate the pool of plausible trees in ML

- Resamples CHARACTERS

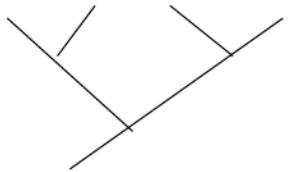
Original Matrix

| | | 1. | 2. | 3. | 4. | 5. | 6. |
|---|---|---|---|---|---|---|---|
| 1 | Lizard | A | A | C | C | G | T |
| 2 | Frog | A | A | C | C | G | T |
| 3 | Fish | G | A | G | C | T | T |
| 4 | Dog | A | A | G | C | G | T |

Bootstrapped Matrix 1

| | | 3. | 2. | 3. | 6. | 5. | 6. |
|---|---|---|---|---|---|---|---|
| 1 | Lizard | C | A | C | T | G | T |
| 2 | Frog | C | A | C | T | G | T |
| 3 | Fish | G | A | G | T | T | T |
| 4 | Dog | G | A | G | T | G | T |

Bootstrap tree 1



# Bootstrapping

- Used to generate the pool of plausible trees in ML

- Resamples CHARACTERS

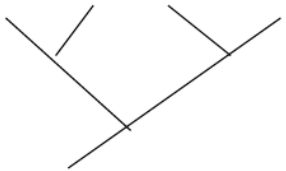
Original Matrix

| | | 1. | 2. | 3. | 4. | 5. | 6. |
|---|---|---|---|---|---|---|---|
| 1 | Lizard | A | A | C | C | G | T |
| 2 | Frog | A | A | C | C | G | T |
| 3 | Fish | G | A | G | C | T | T |
| 4 | Dog | A | A | G | C | G | T |

Bootstrapped Matrix 2

| | | 3. | 3. | 2. | 6. | 6. | 5. |
|---|---|---|---|---|---|---|---|
| 1 | Lizard | A | C | A | T | T | G |
| 2 | Frog | A | C | A | T | T | G |
| 3 | Fish | A | G | A | T | T | T |
| 4 | Dog | A | G | A | T | T | G |

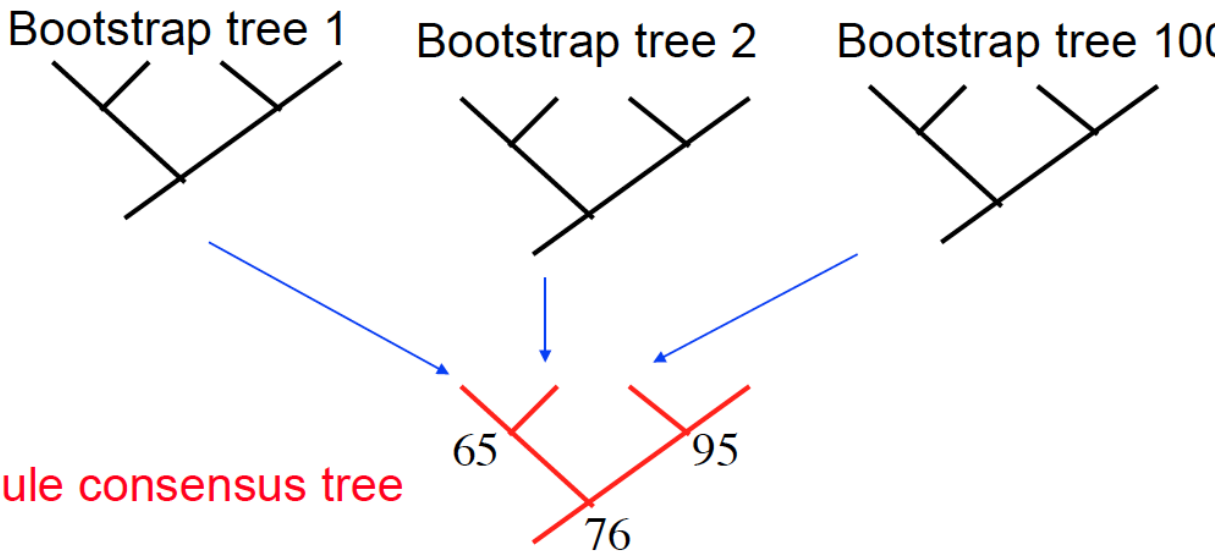
Bootstrap tree 2



# Bootstrapping

- Used to generate the pool of plausible trees in ML

- Resamples CHARACTERS

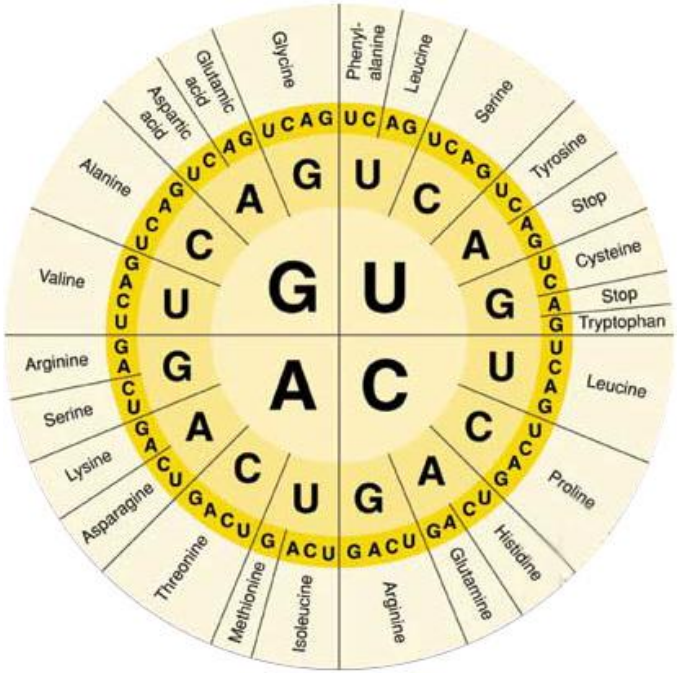


Northwestern Medicine®

Northwestern Medicine®

# Selection Analysis

# Selection




## Molecular signatures of selection

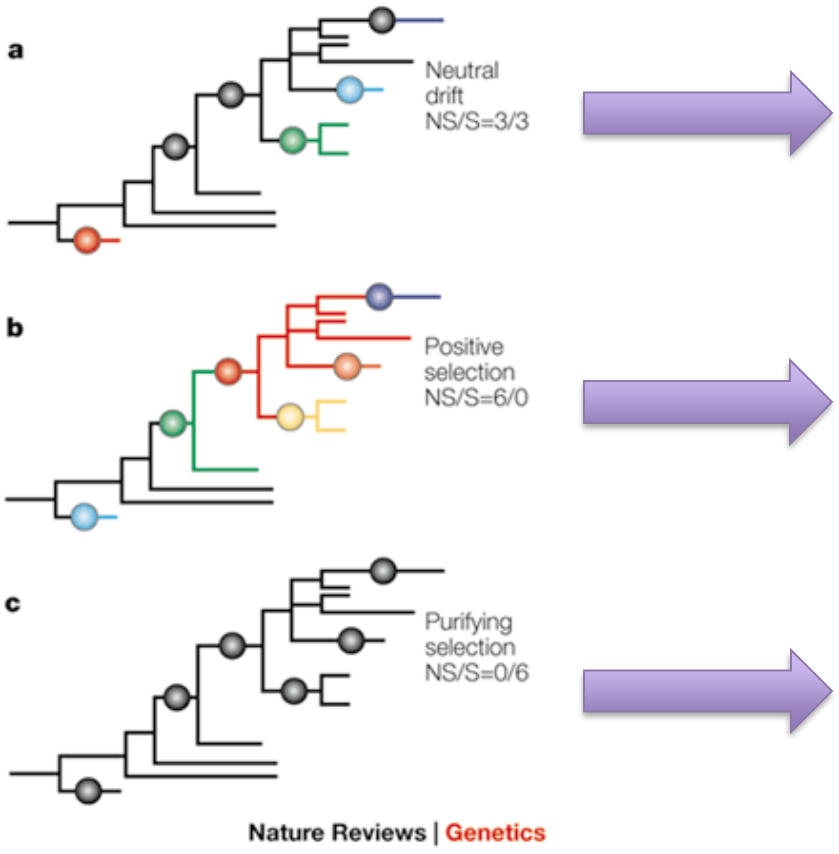
- Because synonymous substitutions do not alter the protein, we often posit that they are neutral
- The **rate** of accumulation of synonymous substitutions (**dS**) gives the neutral background
- We can compare the **rate** of accumulation of non-synonymous substitutions (**dN**), which alter the protein sequence, to classify the nature of the evolutionary process

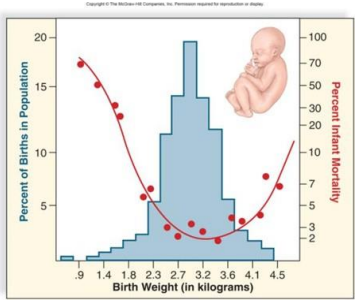
$$dS \sim \frac{\text{number of fixed synonymous mutations}}{\text{proportion of random mutations that are synonymous}}$$

$$dN \sim \frac{\text{number of fixed non-synonymous mutations}}{\text{proportion of random mutations that are non-synonymous}}$$

Northwestern Medicine®

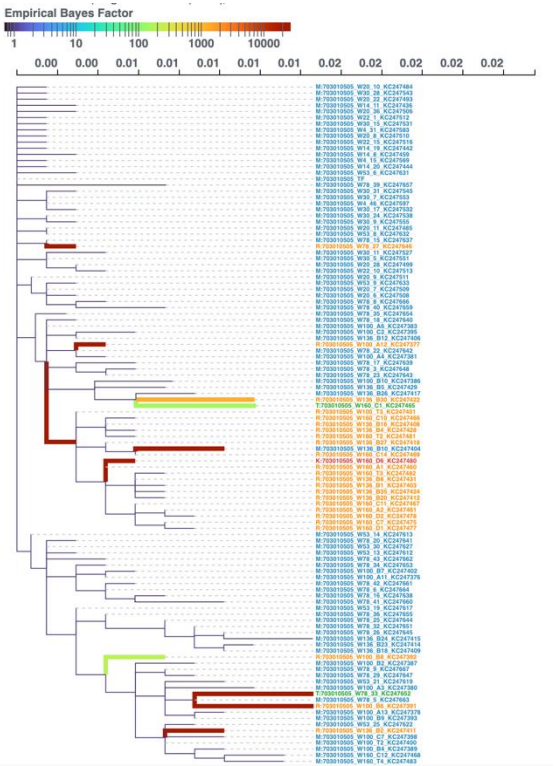
# Selective Processes



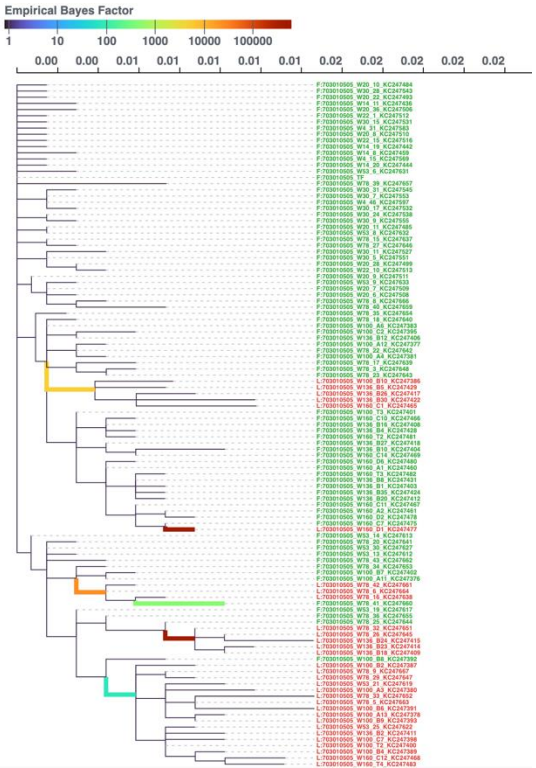



# Episodic diversifying selection results

Codon 4




Codon 21

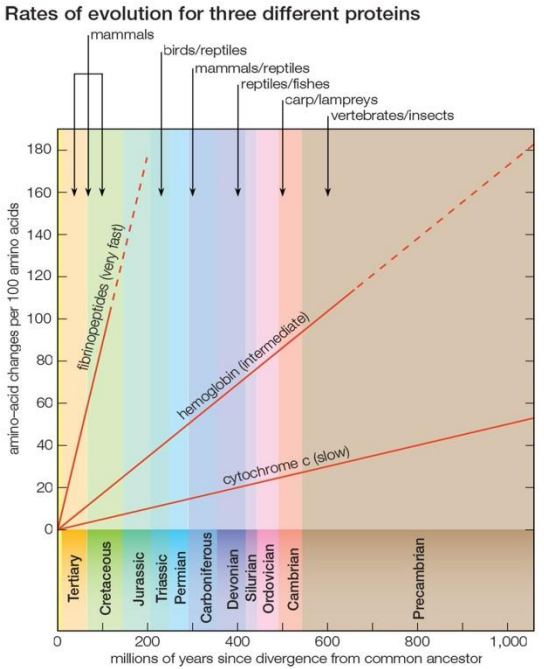


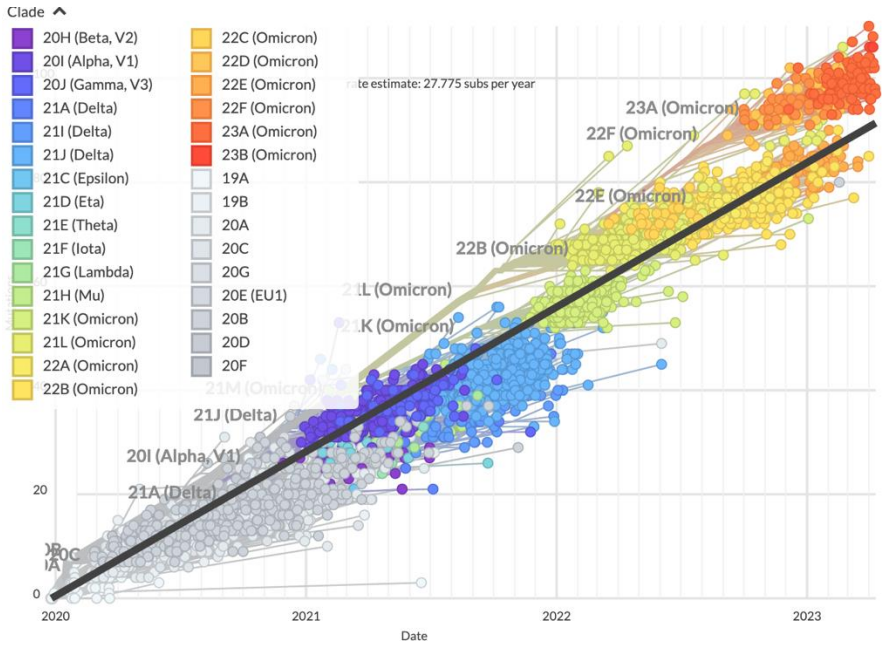
Northwestern Medicine®

# Phylodynamics

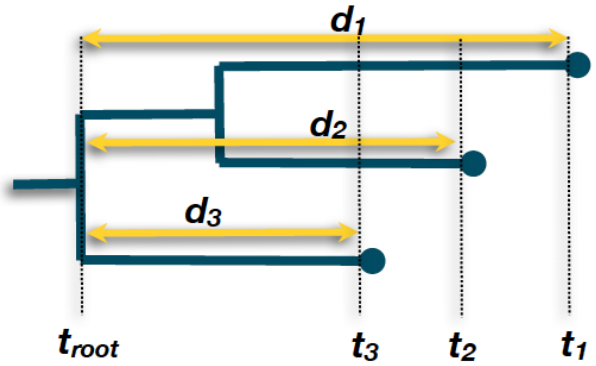
# Molecular clock

- Proposed by Zuckerkandl and Pauling (1965) from hemoglobin data.
- Sequences accumulate changes at a constant rate.
- There's a linear relationship between molecular and temporal.



# Evolutionary Rates




$$\mu = d_i / (t_i - t_{root})$$

- can be rearranged:

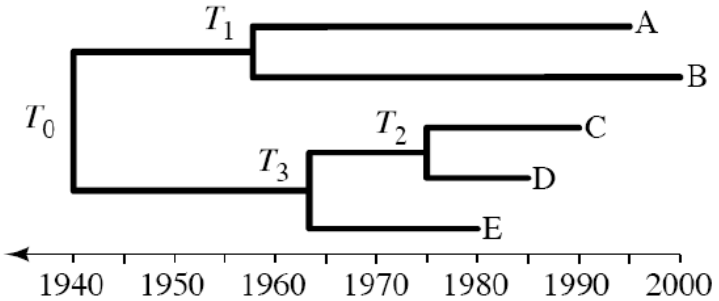
$$d_i = \mu\,(t_i - t_{root})$$

$$E[d_i] = \mu \cdot t_i - \mu \cdot t_{root}$$

gradient is: $\mu$

y-intercept is: $-\mu \cdot t_{root}$

x-intercept is: $t_{root}$

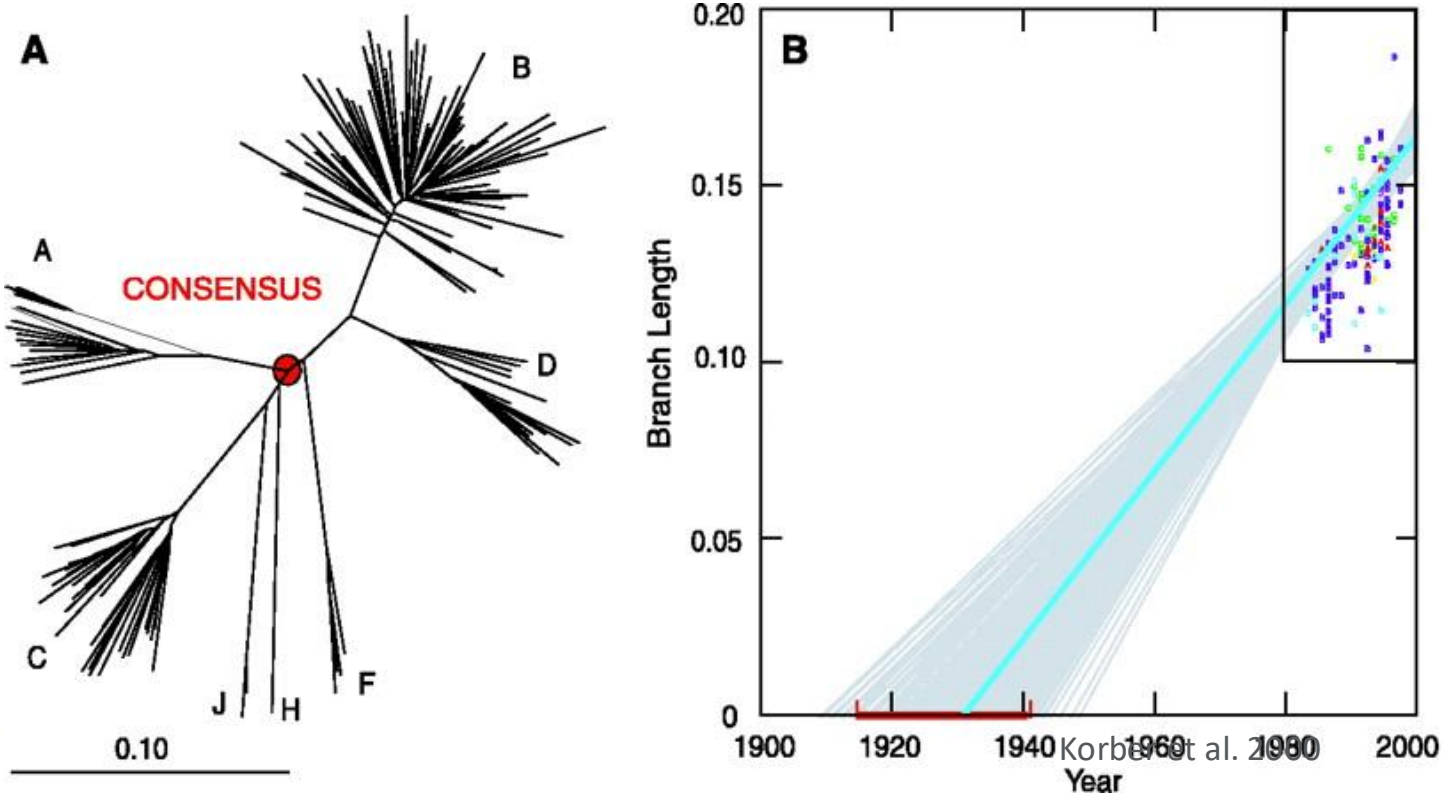



Drummond A.J. & Rambaut A. BMC Evolutionary Biology 2007; 7:214

Northwestern Medicine

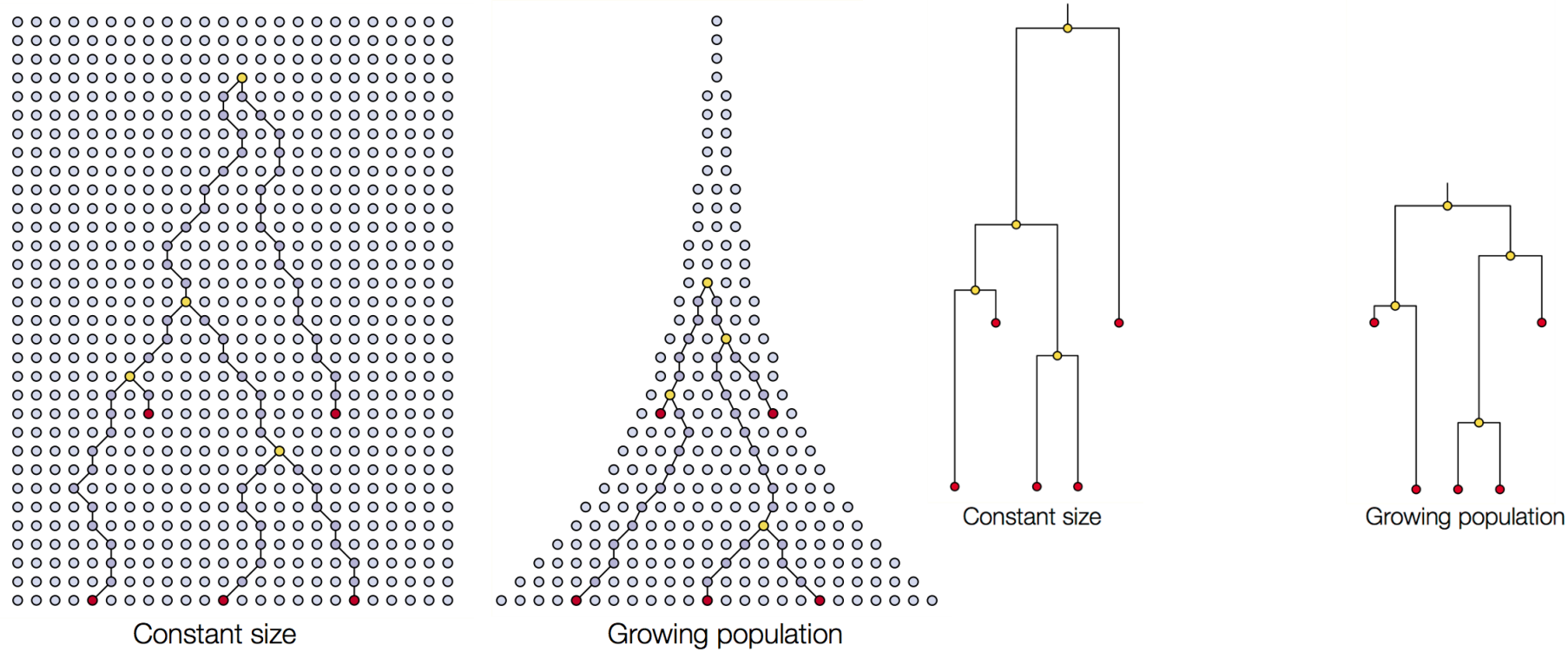
# Time to Most Recent Common Ancestor (TMRCA)

HIV-1 group M origin estimated at around 1930




Korber et al. 2000

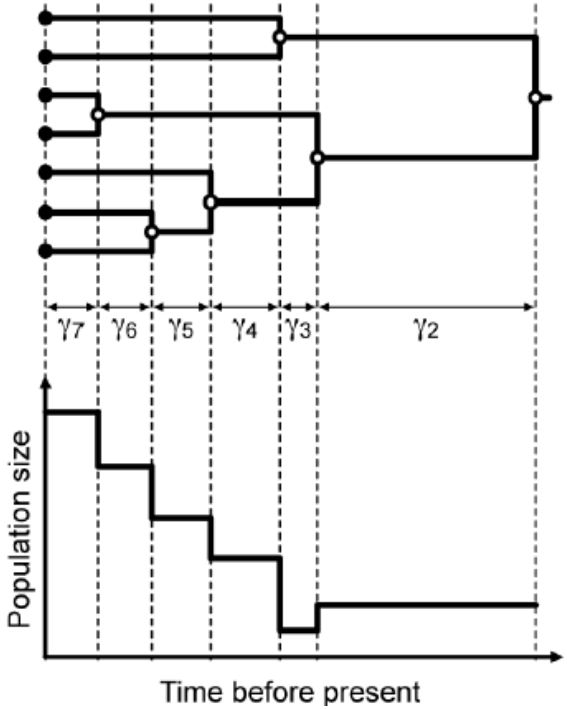
# Changing population size alters coalescent rate




Genealogy is generated by a coalescent process

# Estimation of demographic history

## Classic skyline

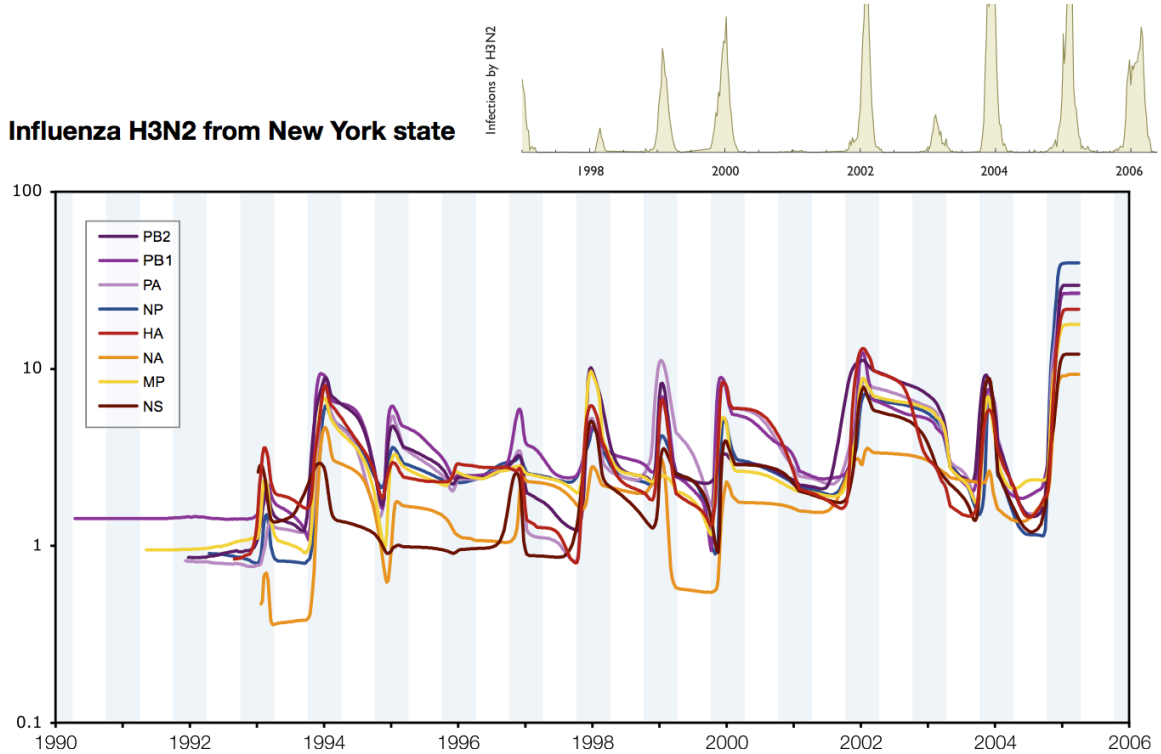



1. Obtain estimate of genealogy
2. Divide into coalescent intervals
3. Estimate population size for each coalescent interval by:

$$N_i = \gamma_i \, i(i\text{-}1)/2$$

Northwestern Medicine®

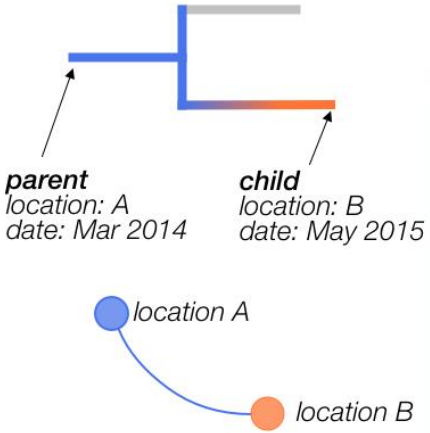
# Skyline model shows seasonality in flu

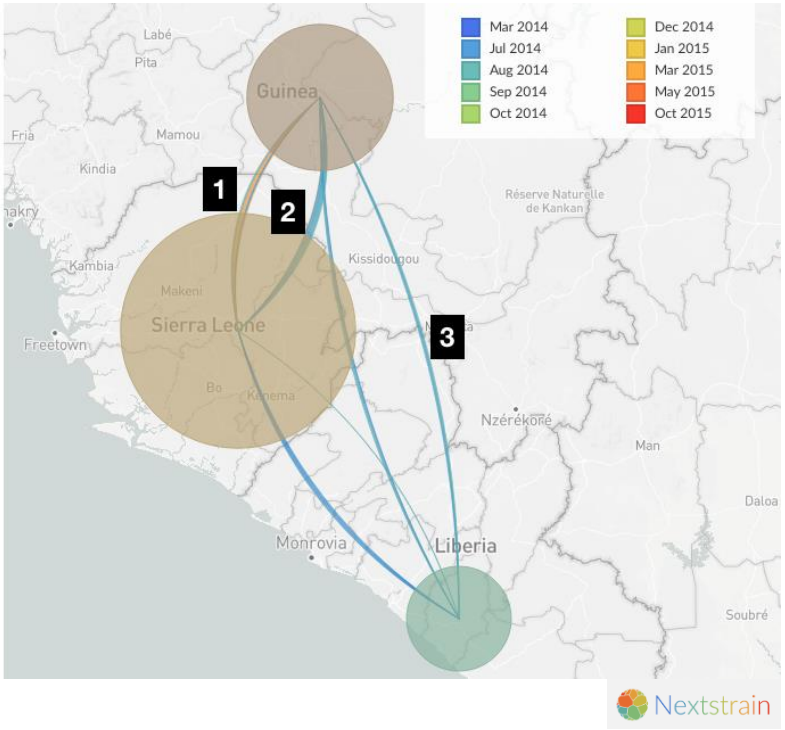



Rambaut et al. 2008

Northwestern Medicine®

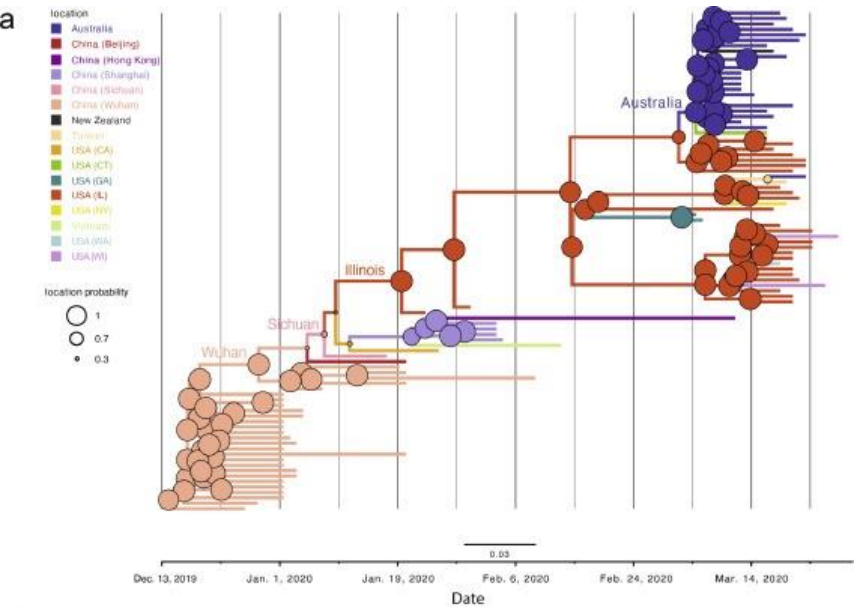
# Phylogeography

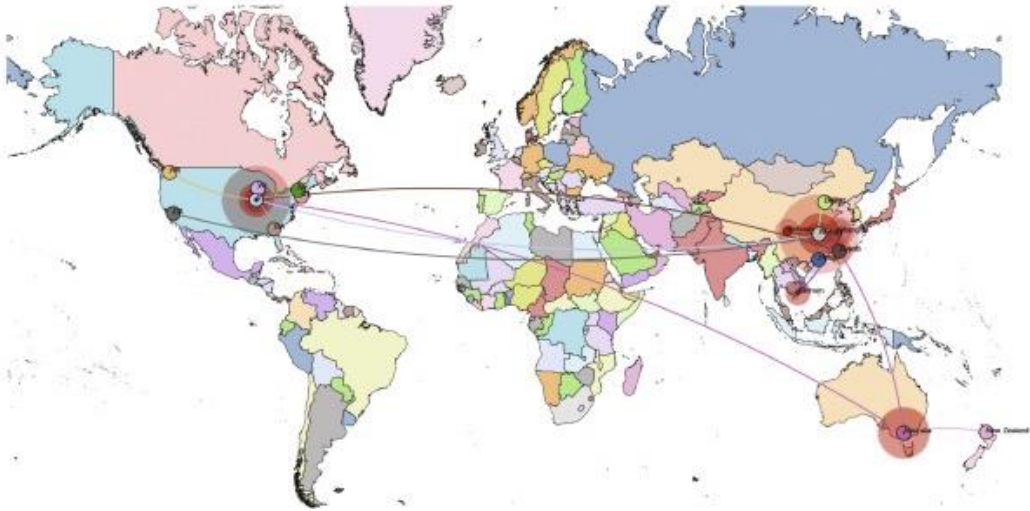



Northwestern Medicine

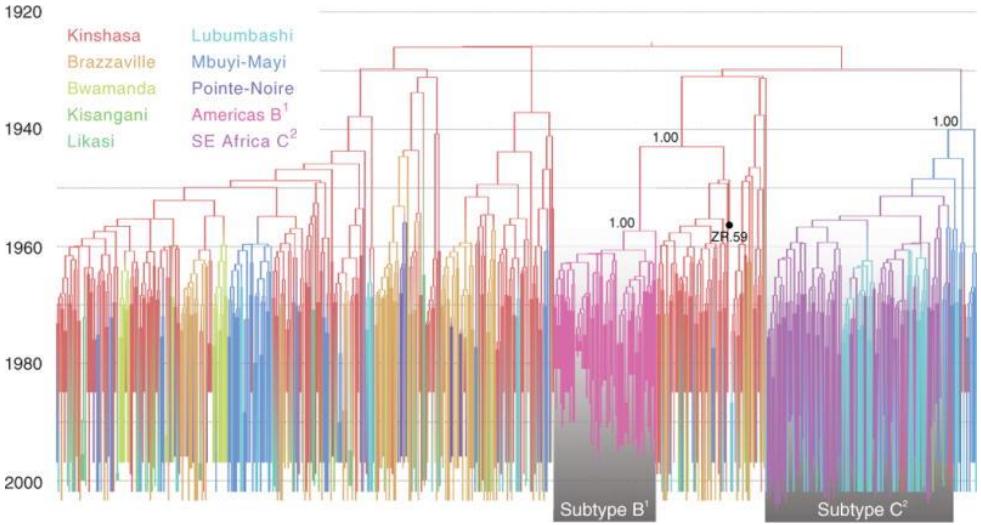
# Phylogeography

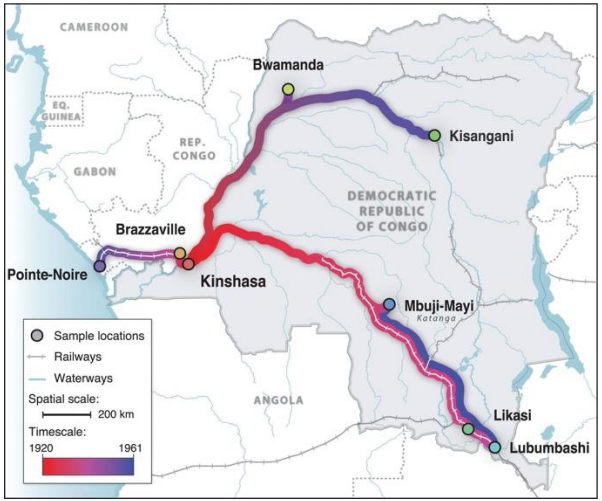


Lorenzo-Redondo et al. 2020

# Phylogeographic analyses




Faria et al. 2014

Northwestern Medicine®

# Phylodynamics

HEALTH • CORONAVIRUS

## BA.4 and BA.5, two new Omicron variants sweeping South Africa, detected in U.S.

Northwestern Medicine®