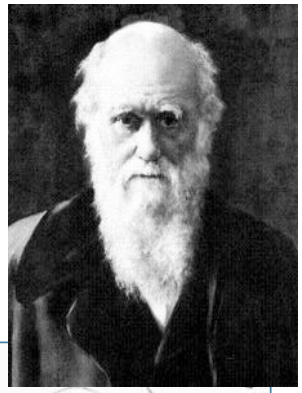
# The basics of phylogenomic trees

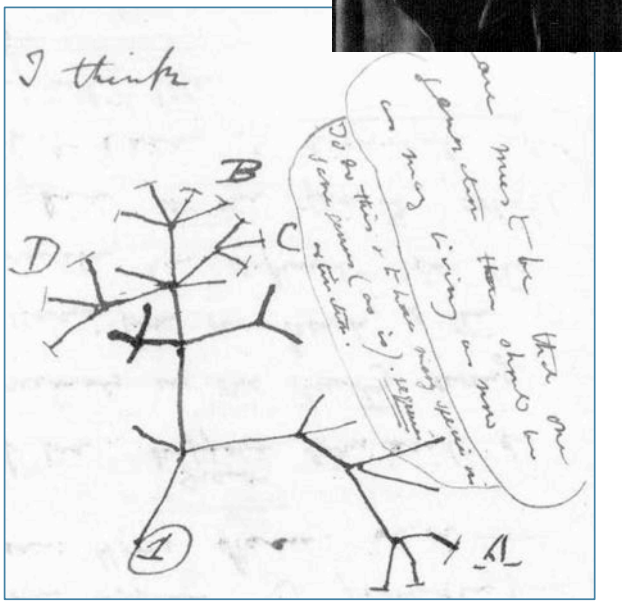

Roman Biek
(Roman.Biek@glasgow.ac.uk)


Institute of Biodiversity, Animal Health & Comparative Medicine

# A quick primer on phylogenetics

- Building trees from genetic sequence data: Reconstructing the ancestral relationships among taxa
- Taxa can be species, individuals or particular genes
- Tree is only an estimate => “truth” usually unknown

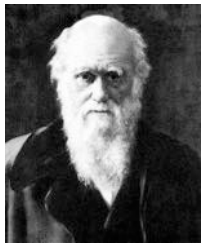




# A simple four taxa example

Who is our closest relative?





?

?

?

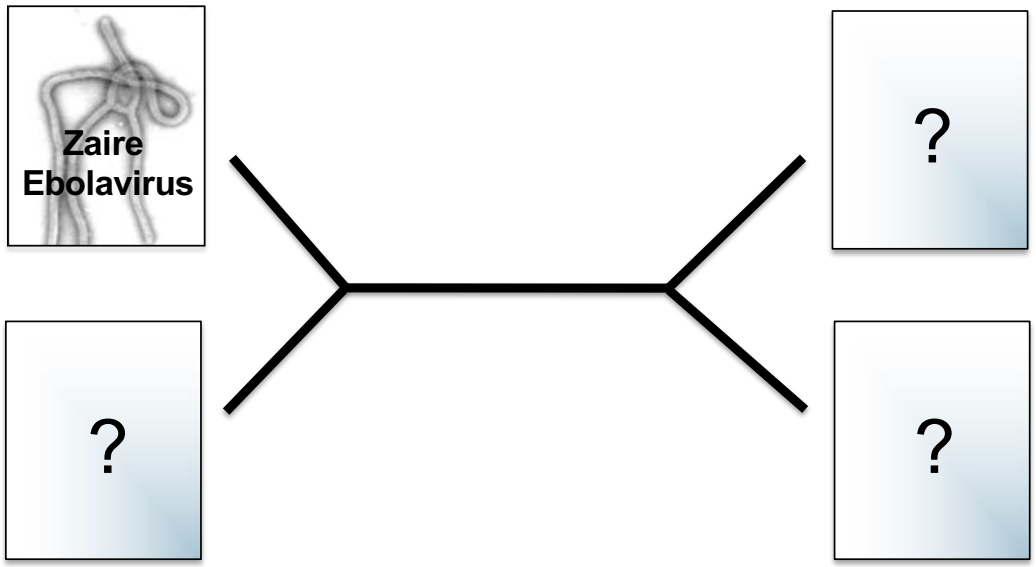
# A simple four taxa example

Which virus is closest related
to Zaire Ebolavirus?

Reston Ebolavirus

Tai Forest Ebolavirus

Sudan Ebolavirus

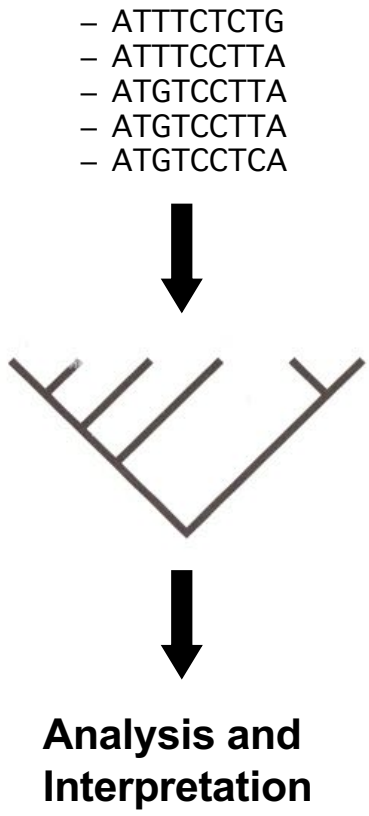


# The overall aim

Measure variation at the molecular level

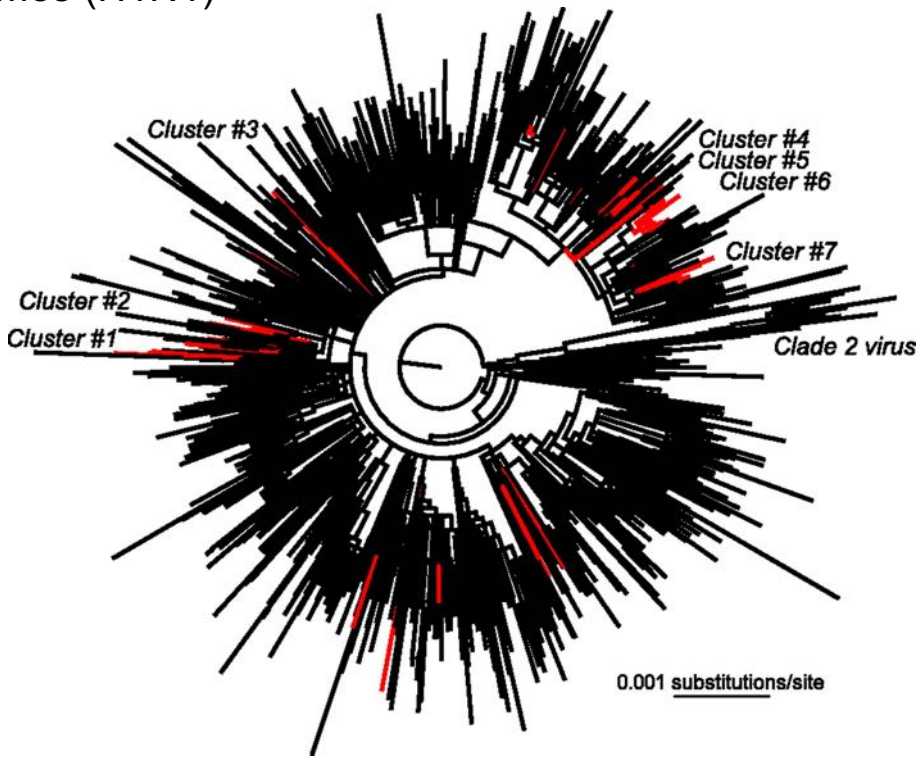
Develop models that fit the observed patterns

Infer process from patterns



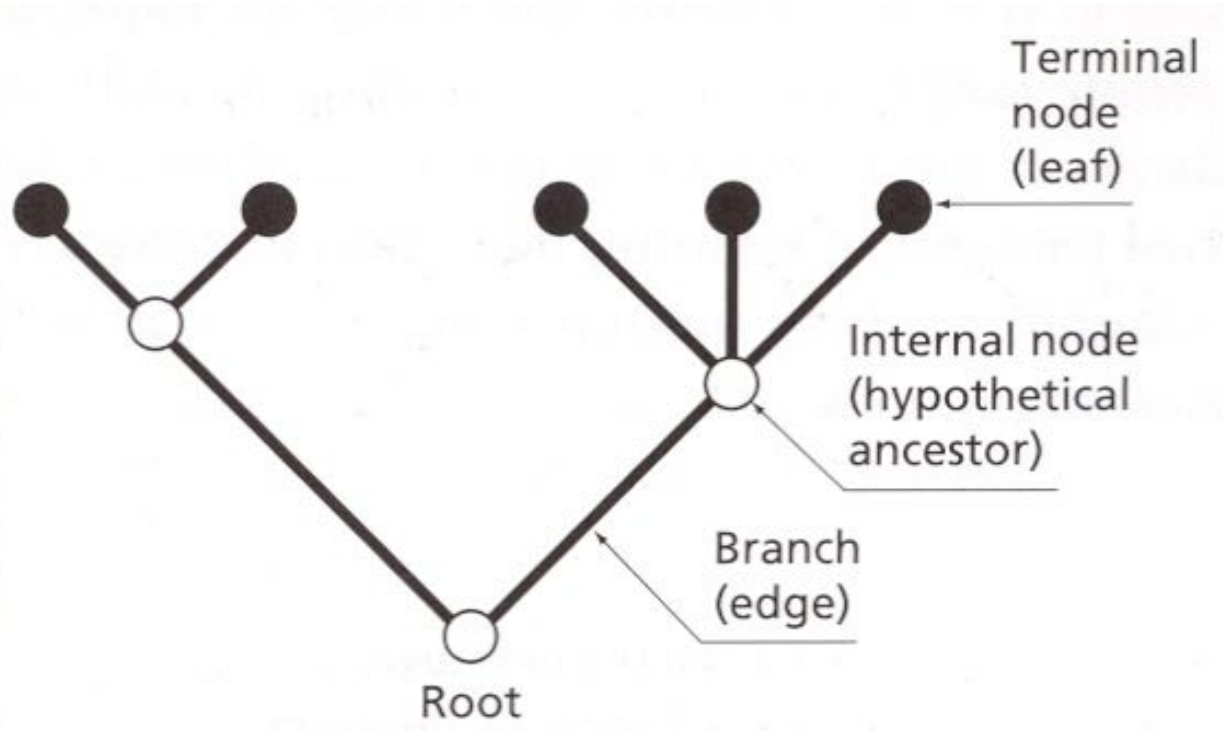
# Gaining epidemiological insights from phylogenomic trees

Tree based on 1,036 complete influenza A virus genomes (H1N1)

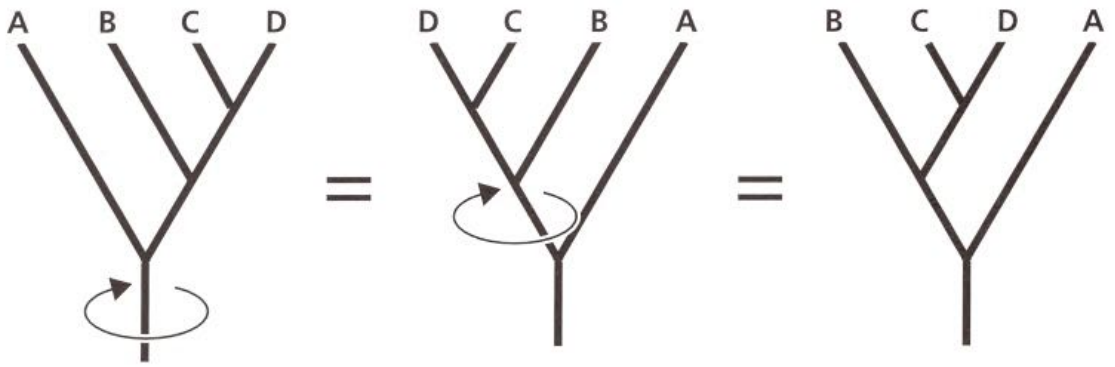



*Holmes E C et al. 2011 J. Virol.*

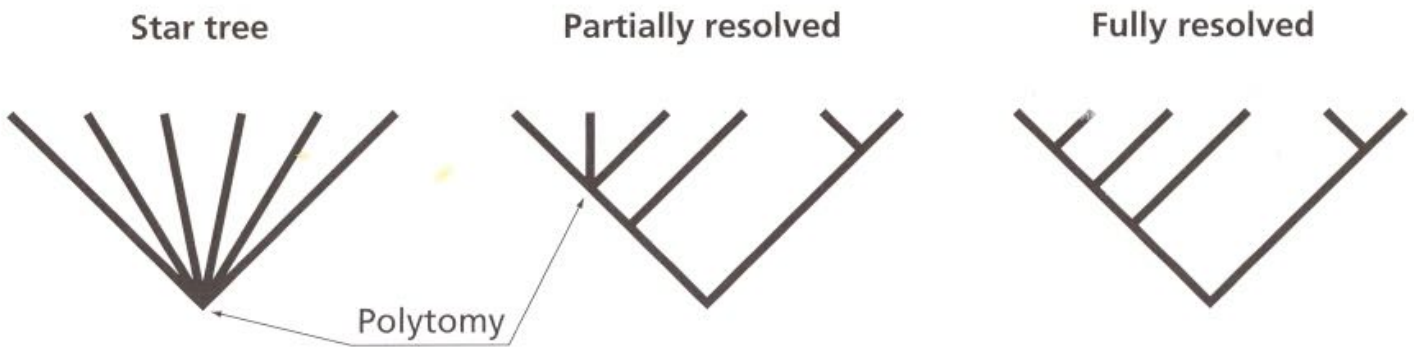
# The parts of a tree



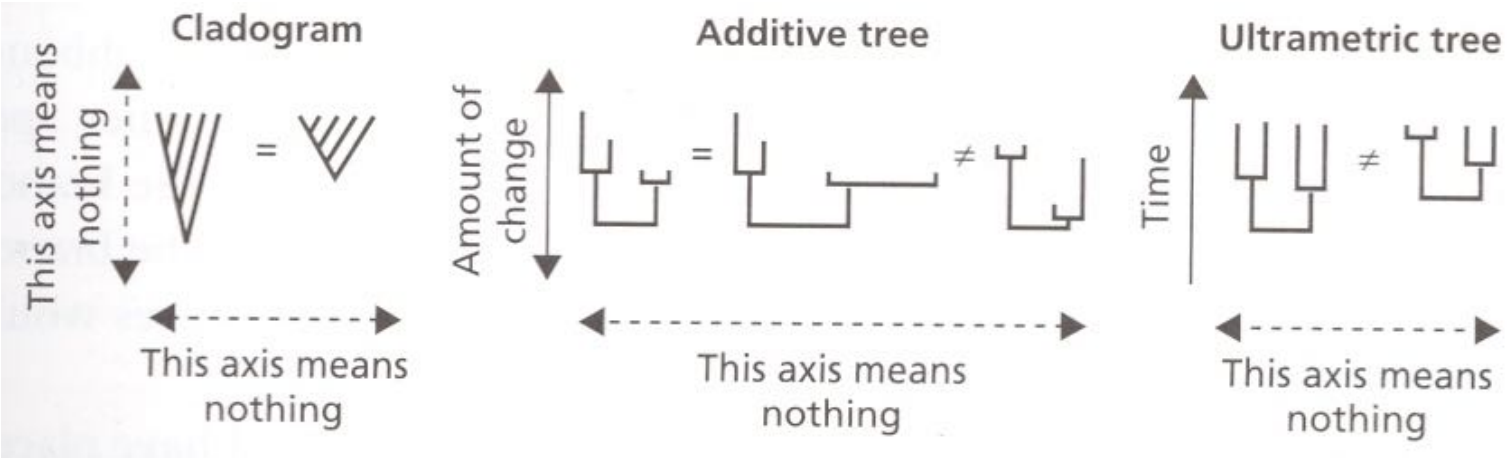
# Trees are like mobiles



# Tree not always strictly bifurcating



# Different ways to depict a tree

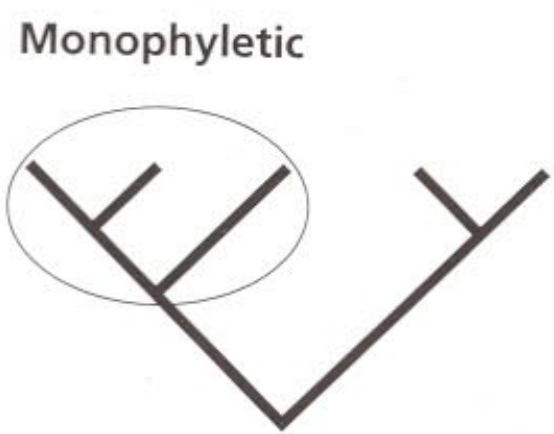



Topology only

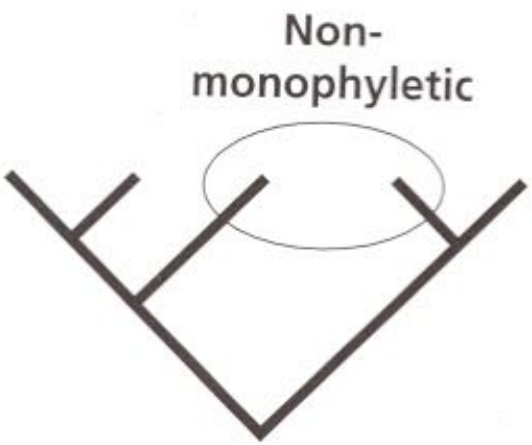
Shape of the tree

Topology +
Branch lenghts

# Monophyly vs Non-Monophyly



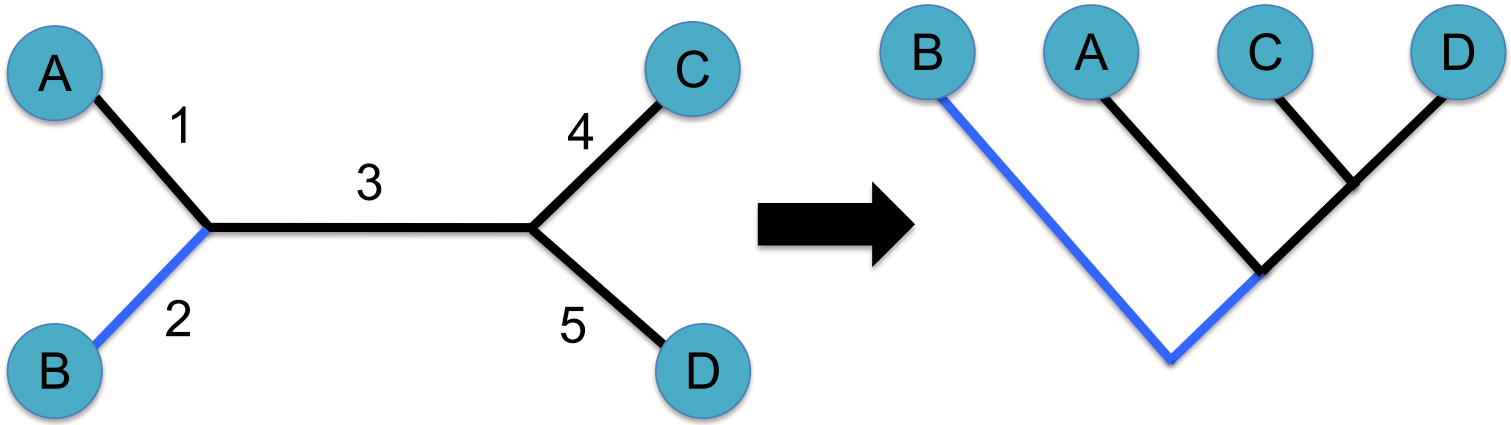

All descendents derived from one ancestor
AND all descendants included




Does not include all descendants

# Rooted vs unrooted trees

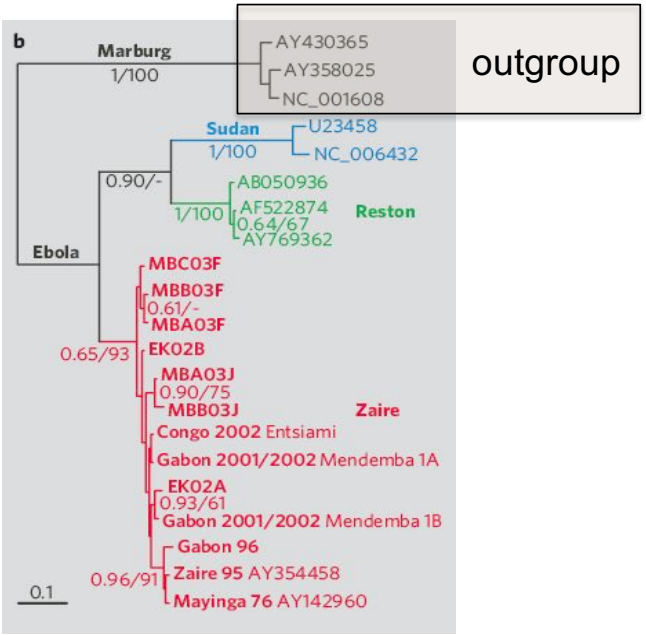
Multiple options for placing the root



# Two ways of rooting a tree

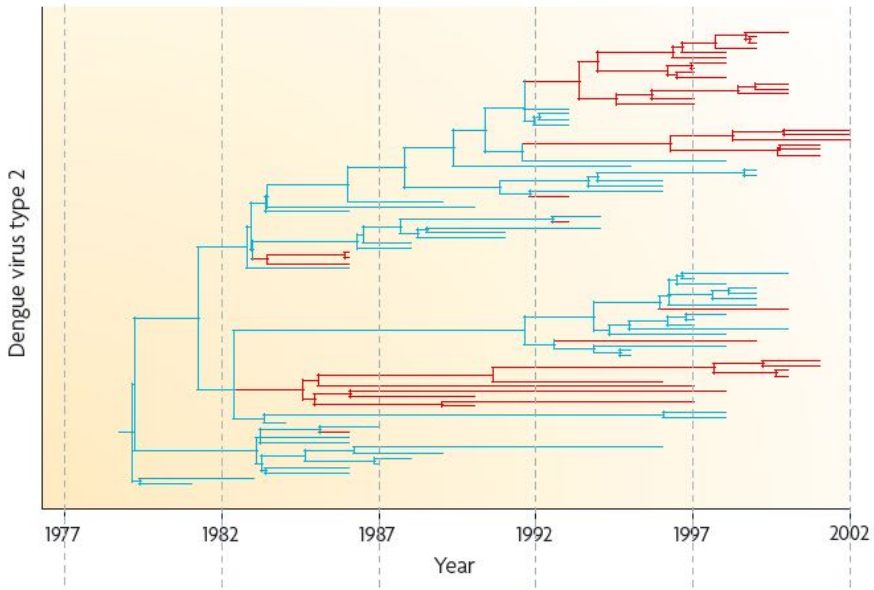
1. Using an **outgroup**:

One or more taxa that fall outside the group of interest




2. Using a **molecular clock**:

Orientates the tree along a time axis



# Number of possible trees rises quickly!

| Taxa | Unrooted trees | Rooted trees |
|---|---|---|
| 3 | 1 | 3 |
| 4 | 3 | 15 |
| 5 | 15 | 105 |
| 6 | 105 | 945 |
| 7 | 954 | 10,395 |
| 8 | 10,395 | 135,135 |
| 9 | 135,135 | 2,027,025 |
| 10 | 2,027,025 | 34,459,425 |
| 20 | 2.22E+20 | 8.20E+21 |
| 30 | 8.69E+36 | 4.95E+38 |

# The basic steps of phylogenetic analysis

1) Collect homologous sequences
2) Conduct multiple alignment
3) Fit an appropriate substitution model
4) Estimate tree(s) under that model
5) Test the reliability of the estimated tree(s)
6) Interpret and apply the phylogenetic tree
7) Potentially repeat steps 4-6 using different tree building methods and/or additional data

# The basic steps of phylogenetic analysis

1) Collect homologous sequences
2) Conduct multiple alignment
3) Fit an appropriate substitution model
4) Estimate tree(s) under that model
5) Test the reliability of the estimated tree(s)
6) Interpret and apply the phylogenetic tree
7) Potentially repeat steps 4-6 using different tree building methods and/or additional data

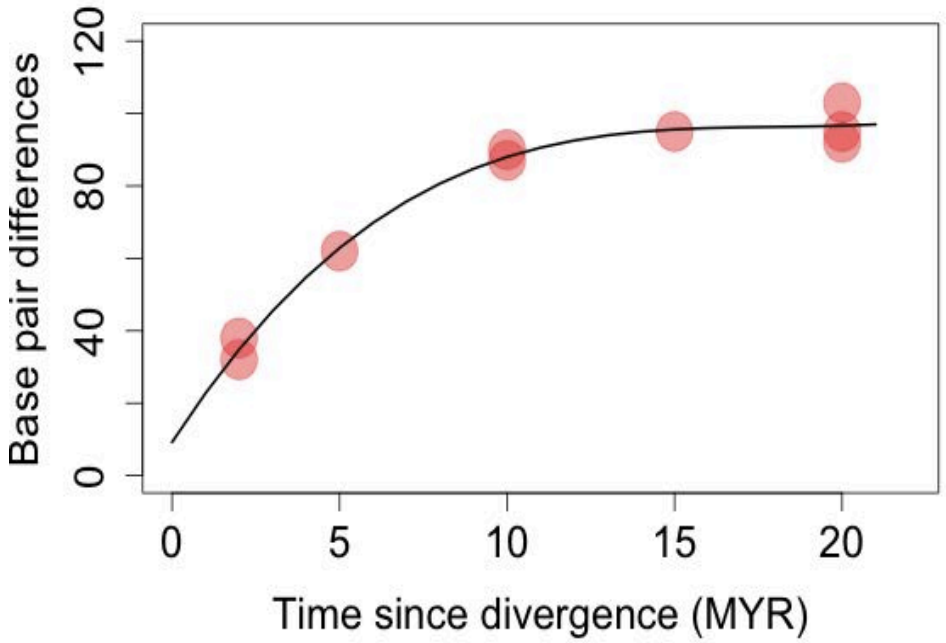
# Models of substitution

How to measure distance between two sequences?

Easiest measure would be number (or proportion) of different sites

=> Problem of multiple ‘hits’ at the same site

mtDNA data from bovine mammals



# Jukes - Cantor model

All nucleotides undergo changes at the same rate

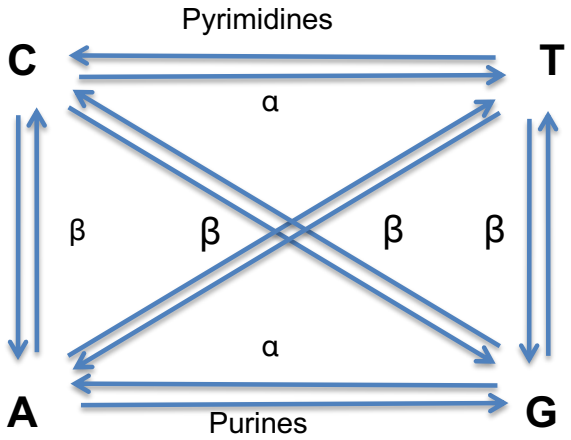
Nucleotide frequencies are the same

$$q_A = q_C = q_G = q_T = ¼$$

| | A | T | C | G |
|---|---|---|---|---|
| A | - | α | α | α |
| T | α | - | α | α |
| C | α | α | - | α |
| G | α | α | α | - |

# Kimura 2-parameter model

Transitions (α) (purine to purine or pyrimidine to pyrimidine subsitutions) are more common than transversions (β)



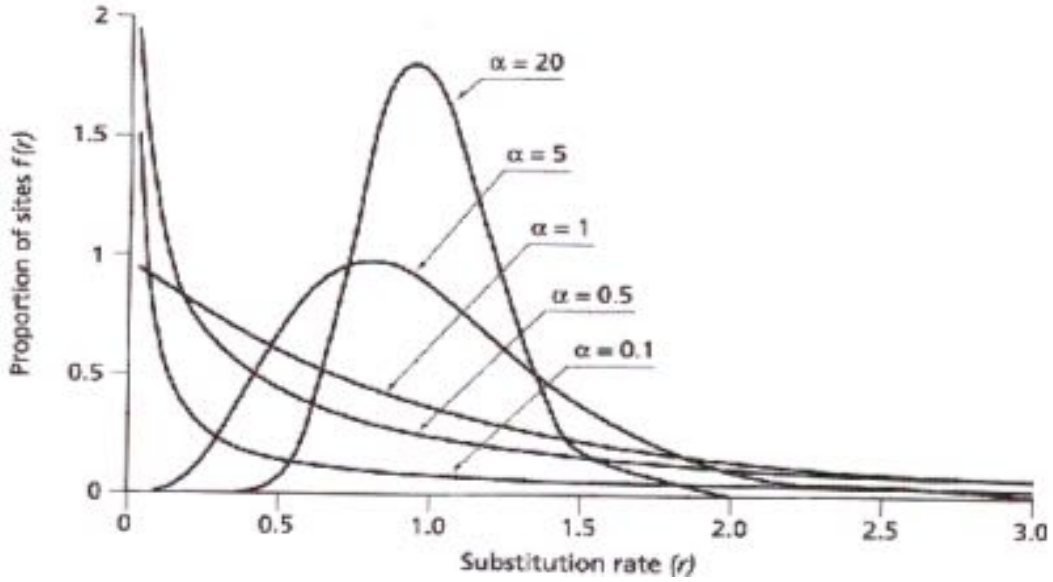

| | A | T | C | G |
|---|---|---|---|---|
| **A** | - | β | β | α |
| **T** | β | - | α | β |
| **C** | β | β | - | β |
| **G** | α | β | β | - |

# Substitution models

- **Kimura**: different probabilities for transitions and transversions
- **HKY**: different probabilities for transitions and transversions, also takes into account that frequencies of nucleotide bases might differ
- **GTR** (general time reversible model): the most general substitution model because each type of substitution is allowed to have its own rate
- **Codon models**: assign different substitution rates and parameters to the three codon position (only for coding sequences)

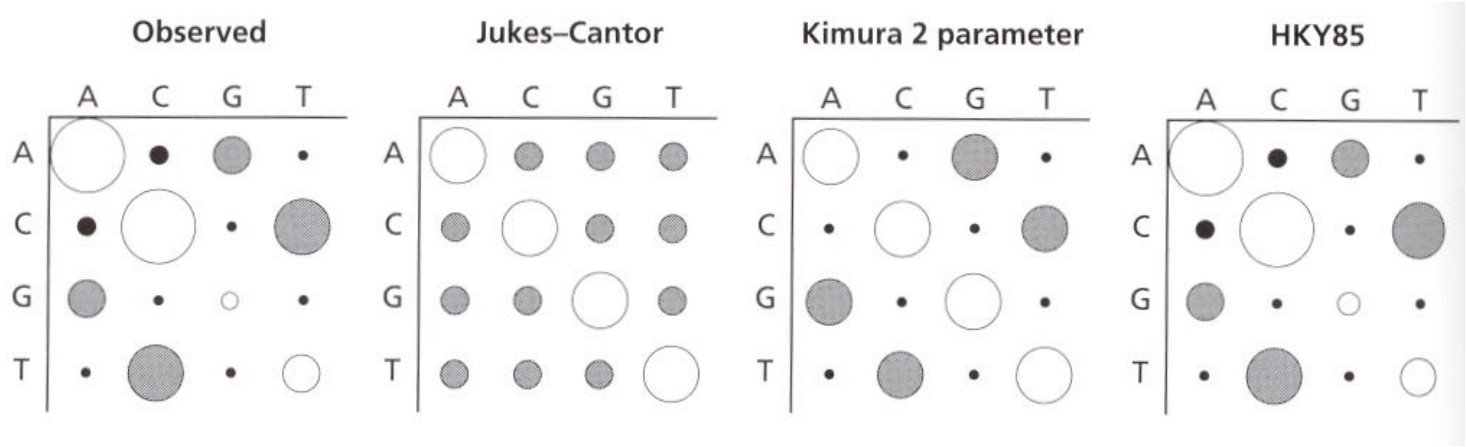
# Variation among sites

Some sites undergo changes more frequently than others - can be expressed using a gamma distribution




slow fast

# Finding a substitution model



# Choosing the right model

jModeltest

Available from: http://darwin.uvigo.es/software/jmodeltest.html

Fits up to 88 candidate models fit to your sequence data

Table 1. Substitution models available in jModelTest. Any of these models can include invariable sites (+I), rate variation among sites (+G), or both (+I+G).

| Model | Reference | Free parameters | Base frequencies | Substitution rates | Substitution code |
|---|---|---|---|---|---|
| JC | (Jukes and Cantor 1969) | 0 | equal | AC=AG=AT=CG=CT=GT | 000000 |
| F81 | (Felsenstein 1981) | 3 | unequal | AC=AG=AT=CG=CT=GT | 000000 |
| K80 | (Kimura 1980) | 1 | equal | AC=AT=CG=GT; AG=CT | 010010 |
| HKY | (Hasegawa, Kishino, and Yano 1985) | 4 | unequal | AC=AT=CG=GT; AG=CT | 010010 |
| TNef | (Tamura and Nei 1993) | 2 | equal | AC=AT=CG=GT; AG; CT | 010020 |
| TN | (Tamura and Nei 1993) | 5 | unequal | AC=AT=CG=GT; AG; CT | 010020 |

# Estimating phylogenies

## General approaches for building trees

- Distance based methods
- Maximum parsimony
- Maximum likelihood
- Bayesian methods

# Estimating phylogenies

## Involves two processes:

- Finding the topology
- Estimation of the branch lengths

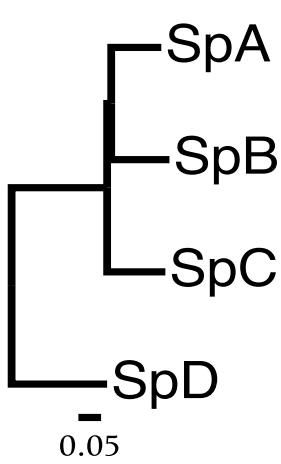
## Optimality criterion

- How well do the data fit a particular tree topology?
- Is used to compare and rank different trees
- Allows to search for the best tree (under given criterion)

# Distance-based methods

```
SpA  ATGCAGGTA
SpB  ATGCTGCTA
SpC  ATGCAGCTC
SpD  TAGCAGGAC
```

| | SpA | SpB | SpC | SpD |
|---|---|---|---|---|
| SpA | - | | | |
| SpB | 2/9 = 0.22 | - | | |
| SpC | 0.22 | 0.22 | - | |
| SpD | 0.44 | 0.66 | 4/9 = 0.44 | - |



# Distance-based methods

Basic procedure

- Calculate pairwise distances among all sequences (according to some substitution model)
- Use distances to build tree (according to some rule e.g. “neighbour joining” method)

Important features

- Very quick way to generate tree, even for large data sets
- No attempt to evaluate alternative trees
- Information about character state change is lost

# Maximum parsimony

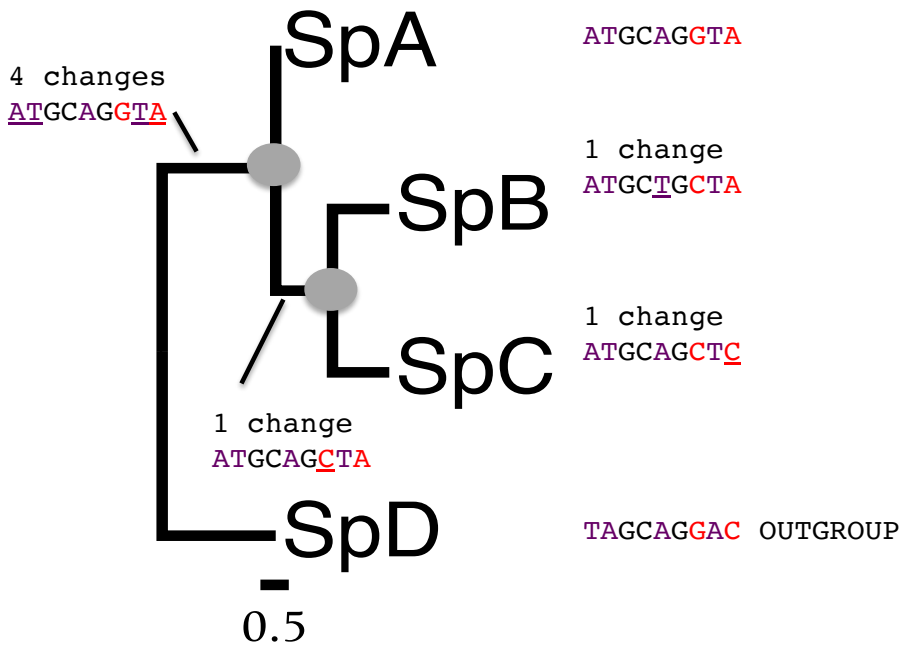
```
ATGCAGGTA
ATGCTGCTA
ATGCAGCTC
TAGCAGGAC
```

3 characters are constant

4 variable characters are parsimony-uninformative

2 parsimony-informative characters

# Maximum parsimony




A minimum of seven changes required to explain the data

3 characters are constant

4 variable characters are parsimony-uninformative

2 parsimony-informative characters

# Maximum parsimony

Basic procedure

- Optimality criterion: parsimony score
- The minimum number of steps (events) necessary to explain the data

Important features

- Score easy to compute => fast method
- All substitutions considered equally likely (weighting schemes possible)
- Implicit assumption that rate of change is low (no multiple hits)
- => Potential problem of “long-branch attraction”

# Maximum likelihood

Basic procedure

- Optimality criterion: likelihood score
- Maximize the probability of the sequences, given a tree and its branch lengths and an evolutionary model and its parameters

Important features

- Allows full use of evolutionary models
- Relies heavily on model chosen => can be misleading if there is much variation in the substitution process among lineages
- Computationally much more demanding

# Bayesian phylogenetics

Basic procedure

- Objective: determine the posterior distribution of trees given the sequence data
- Based on this distribution, ‘best’ tree can be identified

Important features

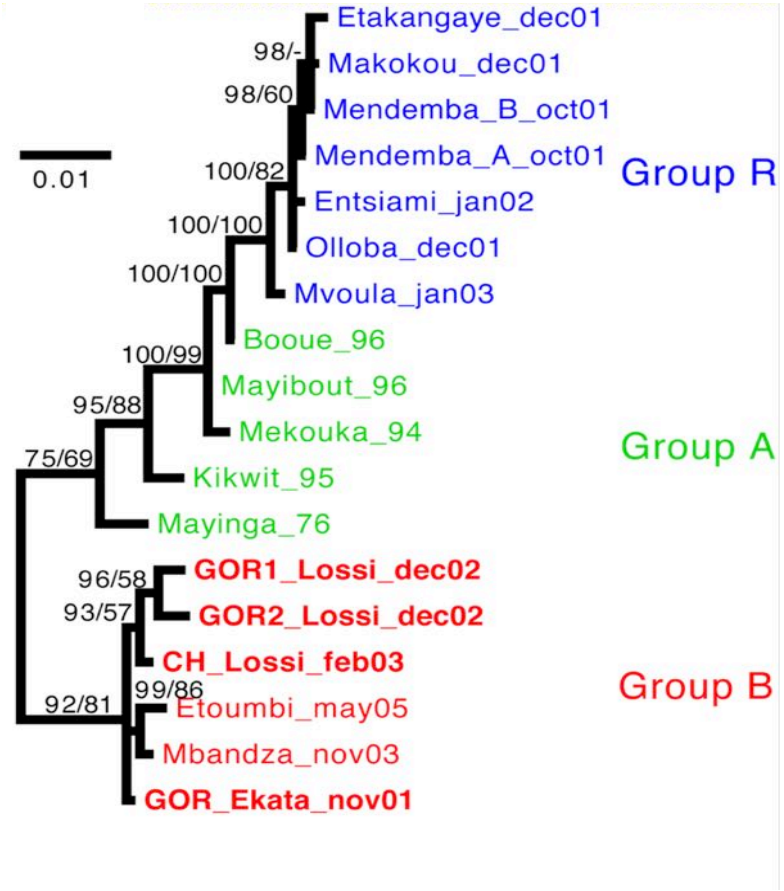
- Allows full use of evolutionary models
- Need to include priors
- Posterior probabilities are approximated through Markov Chain Monte Carlo (MCMC) methods that sample from the posterior
- Clade probabilities provide measure of uncertainty

# How well supported is a grouping?

**Non-parametric bootstrap**

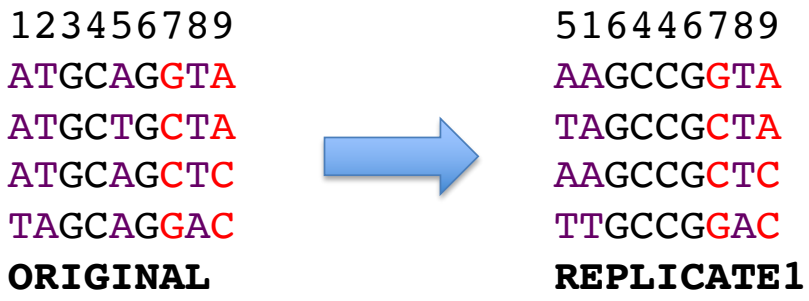
Sample from the original data to create ‘new’ data sets

Count how often a particular clade appears in the resampled data



# Bootstrapping

“new” datasets of same size are generated from original data by sampling columns with replacement




Trees build from these new data sets

The frequency with which a node appears across replicate trees is taken as a measure of confidence for that node

# How well supported is a grouping?

**Posterior probabilities**

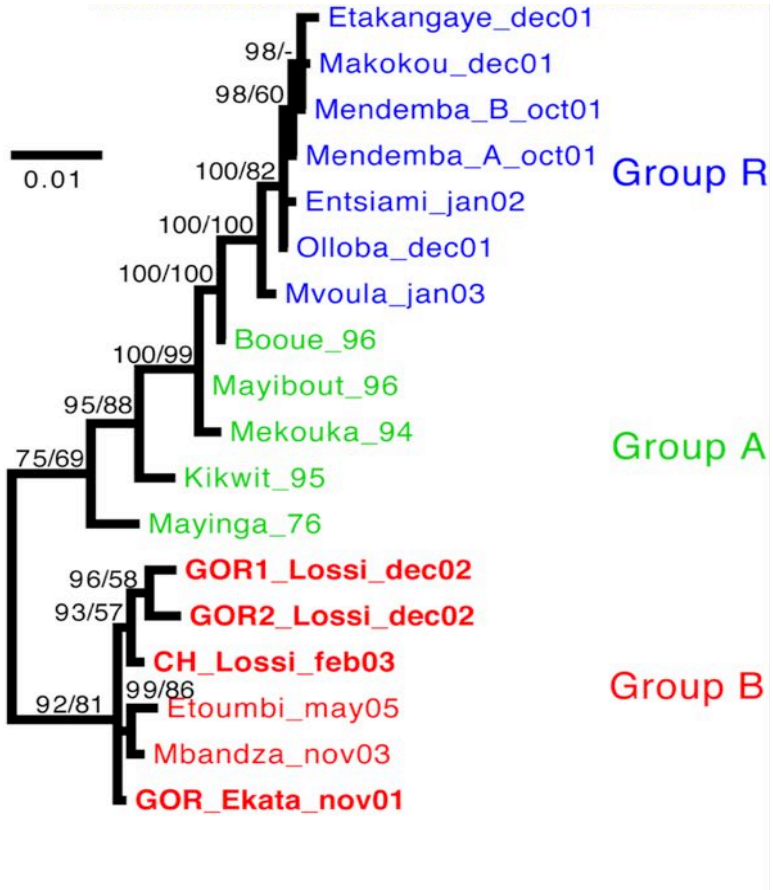
Count the frequency of a clade within the posterior distribution of trees

Less conservative: tend to be much higher than bootstrap values

<u>Strong support:</u>

Bootstrap >0.7

Posterior probabilities >0.95



# Further resources