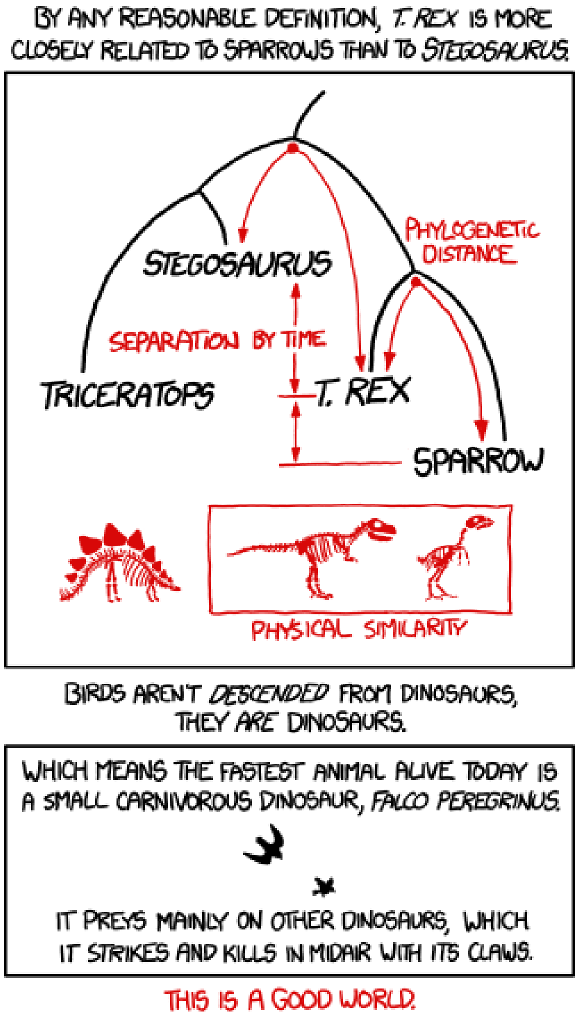
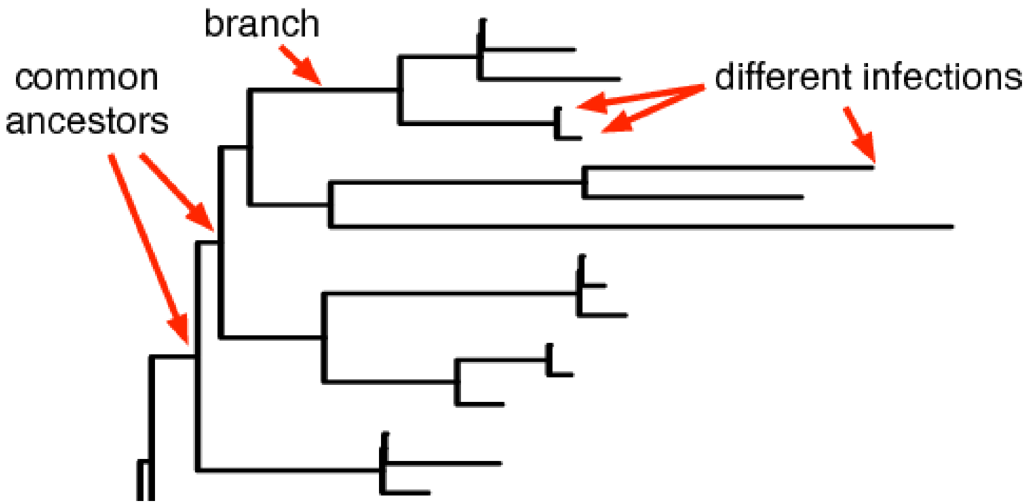
MIMM4750G
Building trees


BY ANY REASONABLE DEFINITION, T. REX IS MORE CLOSELY RELATED TO SPARROWS THAN TO STEGOSAURUS.
STEGOSAURUS
PHYLOGENETIC DISTANCE
SEPARATION BY TIME
TRICERATOPS
T. REX
SPARROW
PHYSICAL SIMILARITY
BIRDS AREN'T DESCENDED FROM DINOSAURS, THEY ARE DINOSAURS.
WHICH MEANS THE FASTEST ANIMAL ALIVE TODAY IS A SMALL CARNIVOROUS DINOSAUR, FALCO PEREGRINUS.
IT PREYS MAINLY ON OTHER DINOSAURS, WHICH IT STRIKES AND KILLS IN MIDAIR WITH ITS CLAWS.
THIS IS A GOOD WORLD.

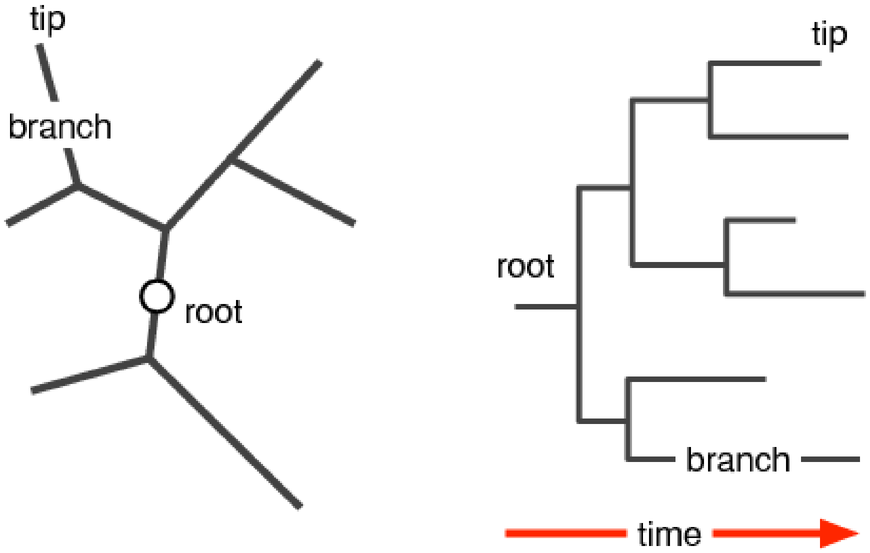
What is a phylogeny?

- A tree-based hypothesis about how populations are related by common ancestors.
- Each population (species/infection) is represented by a tip of the tree.
- Connected by branches to common ancestors (nodes).



What is a root?

- Phylogenies can be rooted or unrooted.
- The root is a hypothesis about what point on the tree represents the earliest time.



- There are different ways to display trees: (*left*) usually used for unrooted trees, (*right*) usually rooted *but not always*.

Phylogenies and infectious disease

- Trees represent how pathogen diversity is structured into subtypes.
- How different pathogen species are related to each other.
- At a population level, a tree tells us something about how a pathogen spread through host populations.

Evolution of pathogenic proteobacteria

- Proteobacteria is a phylum containing many human pathogens.
- Grouped into classes, *e.g.*, α, β, γ, etc.
- This phylogeny reconstructs the emergence of pathogenic species across classes (red branches).
- Study proposes that ancestral Proteobacteria were free-living bacteria.

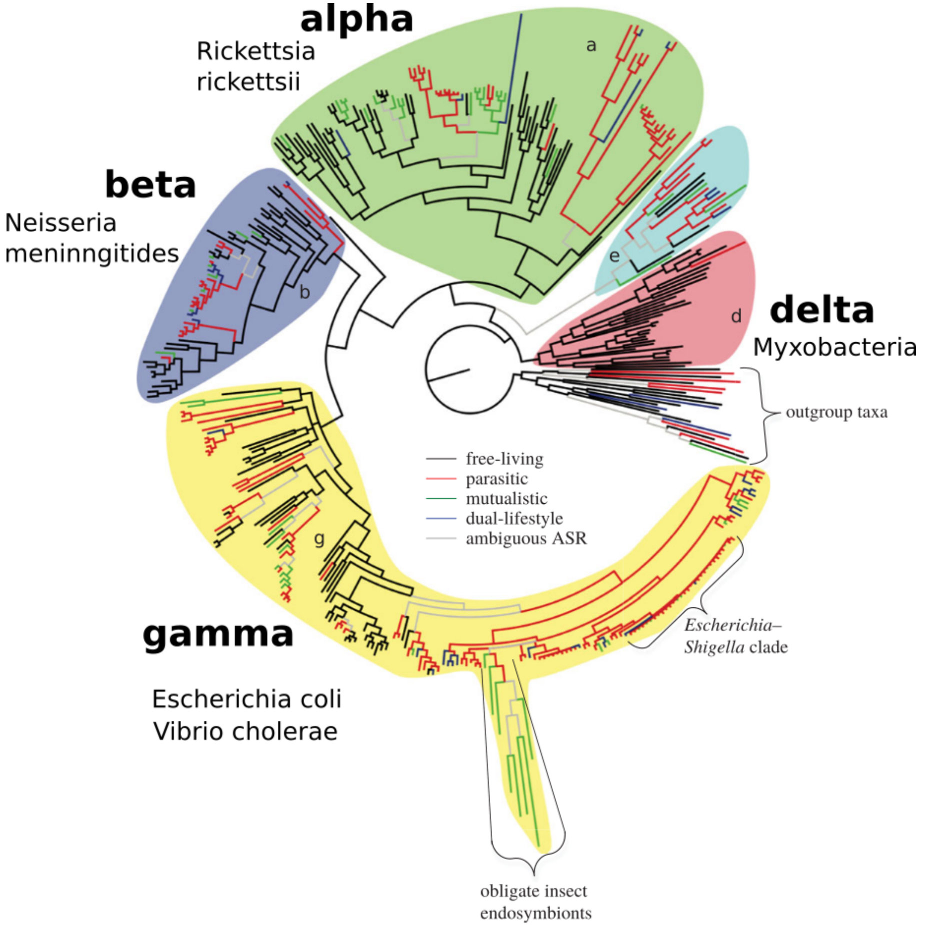
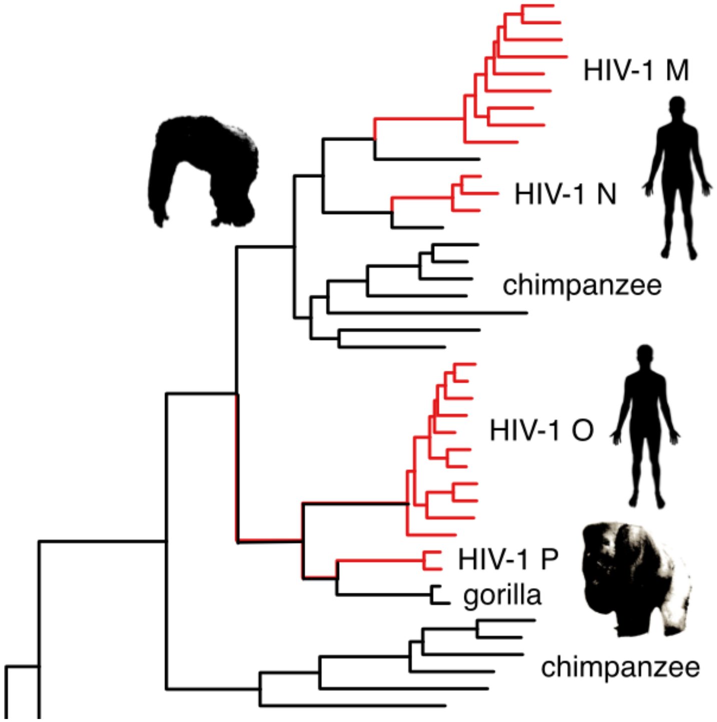


Figure from Sachs *et al.* (2014) *Proc Roy Soc Lond B*, 10.1098/rspb.2013.2146

***INCA Q1** - According to this tree, at least how many times has SIV moved into the human species?*

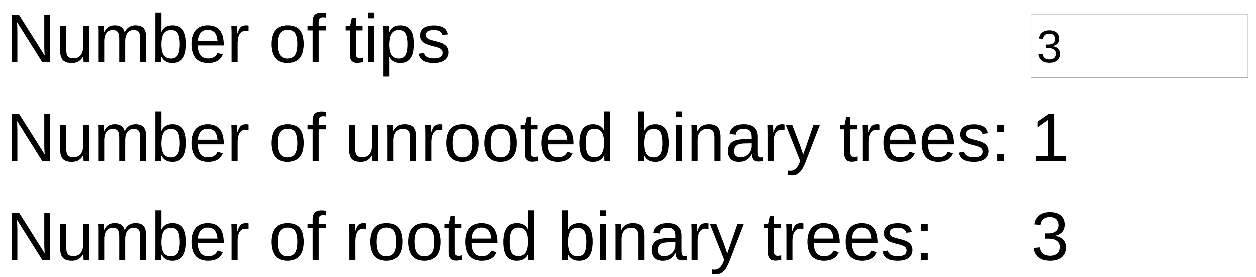


Modified from Joy *et al.* (2015) Origin and Evolution of Human Immunodeficiency Viruses. Global Virology I. Springer, New York.

Ebola virus outbreak in West Africa

How many trees?

- There are an enormous number of possible trees relating even a small number of species!



Number of tips	3
Number of unrooted binary trees:	1
Number of rooted binary trees:	3

INCA Q2 *Why are there always more possible rooted trees than unrooted trees?*

Distance-based methods

- Building a tree can be viewed as a clustering problem!
- We already know how to calculate genetic distances between pairs of sequences.
- Agglomerative (hierarchical) clustering means we group the most similar pair of sequences and progress from there.

UPGMA

- Unweighted pair group method with arithmetic mean.
- Every sequence starts out as a cluster of one ($n_X = 1$).
- Algorithm:

1. Join clusters X, Y with minimum distance:

$$d(X,Y) = \sum_{x \in X} \sum_{y \in Y} d(x,y)/(n_X n_Y)$$

2. Replace X and Y with cluster $X \cup Y$, where:

$$d(X \cup Y, Z) = \frac{n_X d(X,Z) + n_Y d(Y,Z)}{n_X + n_Y}$$

3. Go to step 1 until only one cluster remains (the root).

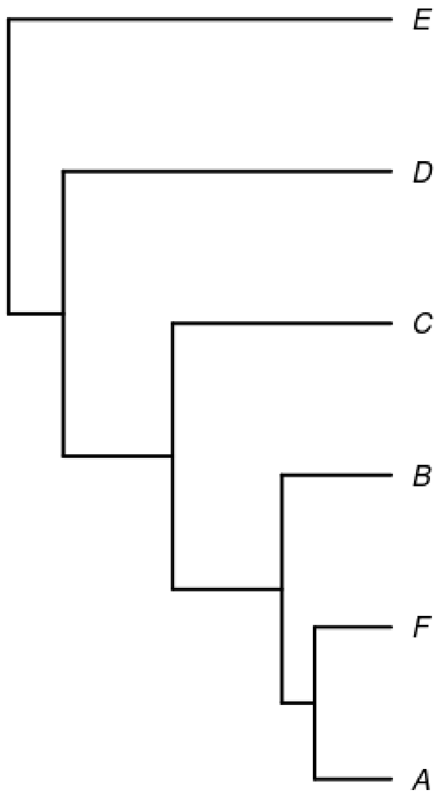
Ultrametric trees

- Because of how UPGMA computes the distances of ancestral nodes, it generates trees where every tip is the same distance from the root.

- This is what you would get if:

  1. we sample each tip at the same moment in time.

  2. the rate of evolution is constant.



Neighbor-joining trees

- Another distance-based clustering method for making trees
- Start with a "star" phylogeny: every tip directly descended from the root
- Add ancestral nodes that minimize the total branch length of the tree

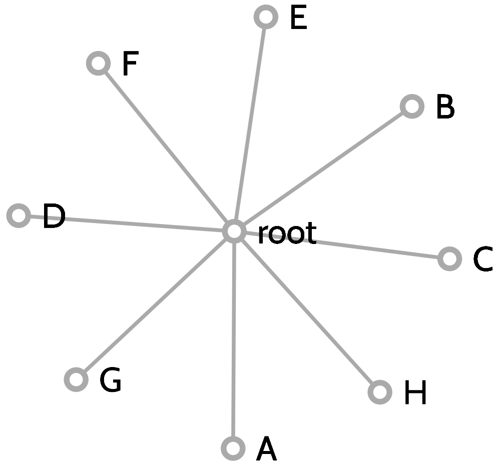
NJ algorithm

1. Calculate distance matrix d_{ij}
2. Calculate vector $u_i = \sum_{j=1}^{n} d_{ij}/(n-2)$
3. Find which i and j that minimize $d_{ij} - u_i - u_j$
4. Place new node ij ancestral to i and j.
5. Calculate new distances from ij to i, j and previous ancestor.

Neighbor-joining

Reset Step!

A	B	C	D	E	F	G	H
0	7	8	11	13	16	13	17
7	0	5	8	10	13	10	14
8	5	0	5	7	10	7	11
11	8	5	0	8	11	8	12
13	10	7	8	0	5	6	10
16	13	10	11	5	0	9	13
13	10	7	8	6	9	0	8
17	14	11	12	10	13	8	0



Pros and cons of NJ

- Distance-based methods are faster than likelihood-based methods (next week)
- Unlike UPGMA, NJ is robust to changing rates of evolution
- Effective for massive data sets.
- Demonstrably accurate for reconstructed trees from simulated data.

Software for NJ

- MEGA
- RapidNJ
- R package ape