Phylogenetics - thinking with trees

Philip Ashton

Anders Gonçalves da Silva, PhD
MDU PHL --- Bioinformatics Team

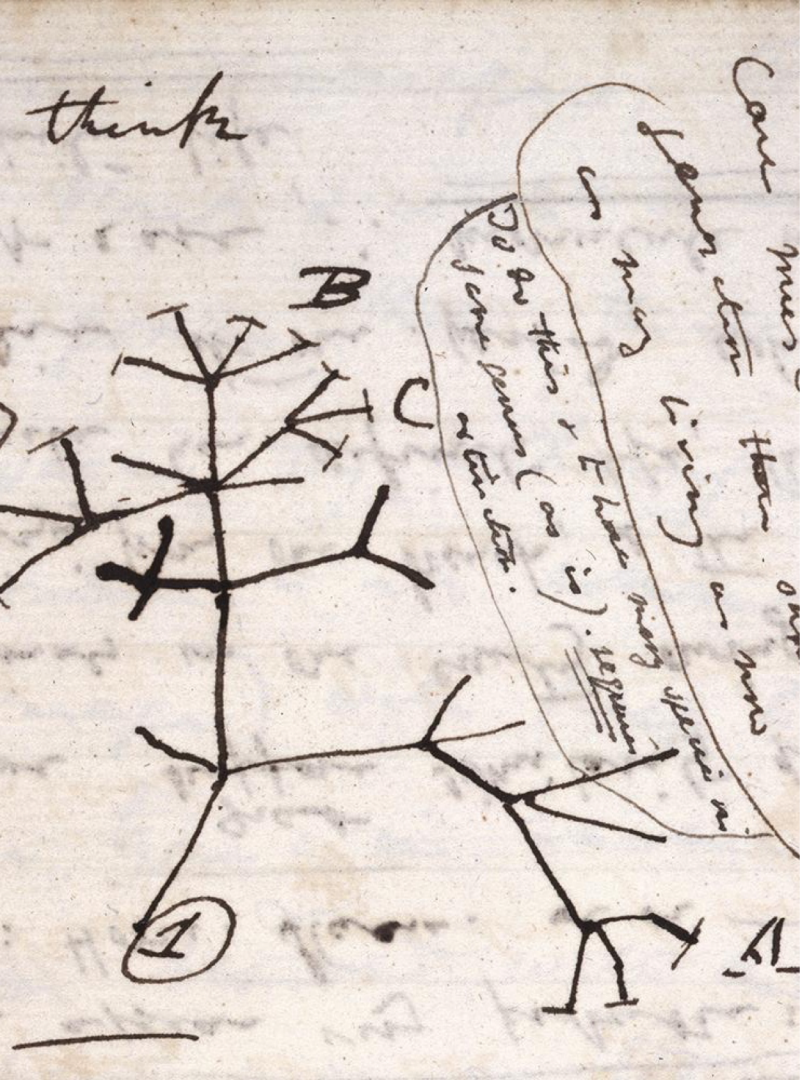
Doherty Institute

PUBLIC HEALTH MICROBIOLOGY

POSTERA CRESCAM LAUDE
THE UNIVERSITY OF MELBOURNE

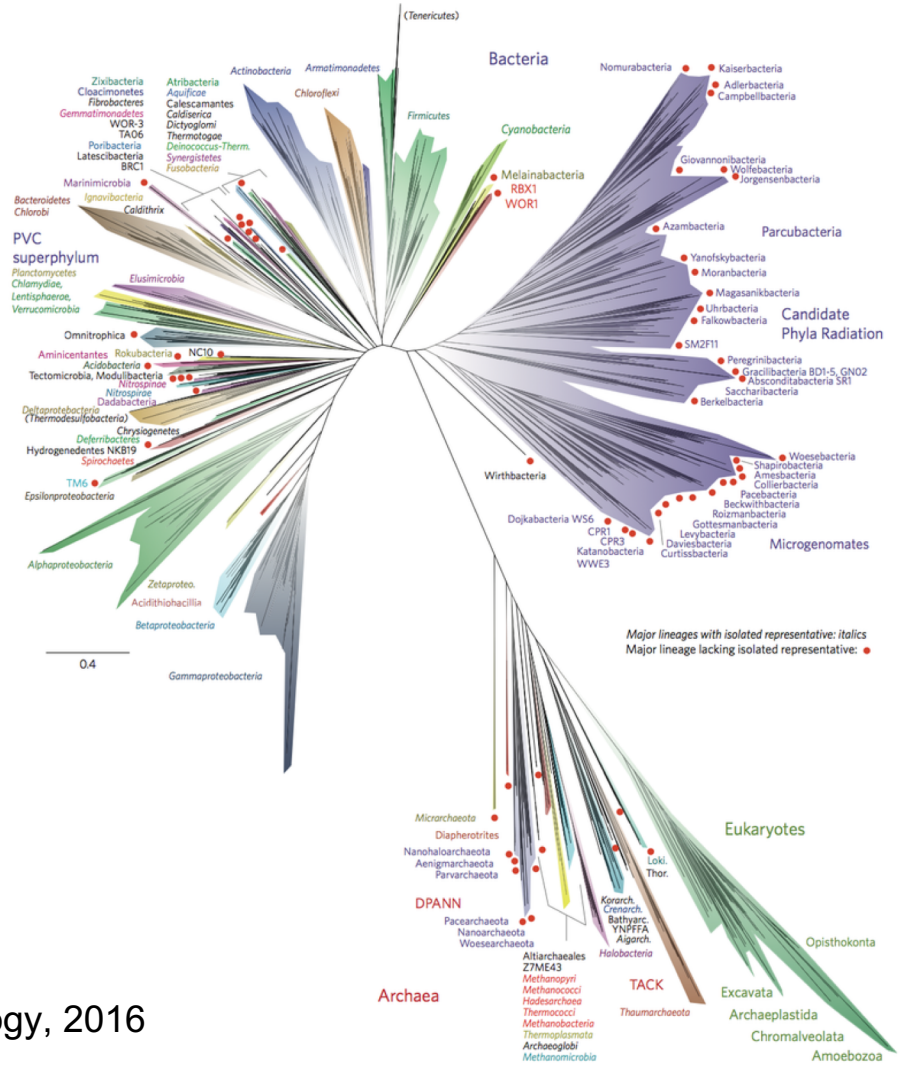
Outline

1. What are trees?
2. How do we make a tree?
3. Topologies
4. Pairwise SNP distances
5. Bootstraps
6. The End
7. Timed phylogenies
8. Branch lengths
9. Models of sequence evolution and DNA substitutions
10. Recombination
11. Why would we use phylogenies in epidemiology?

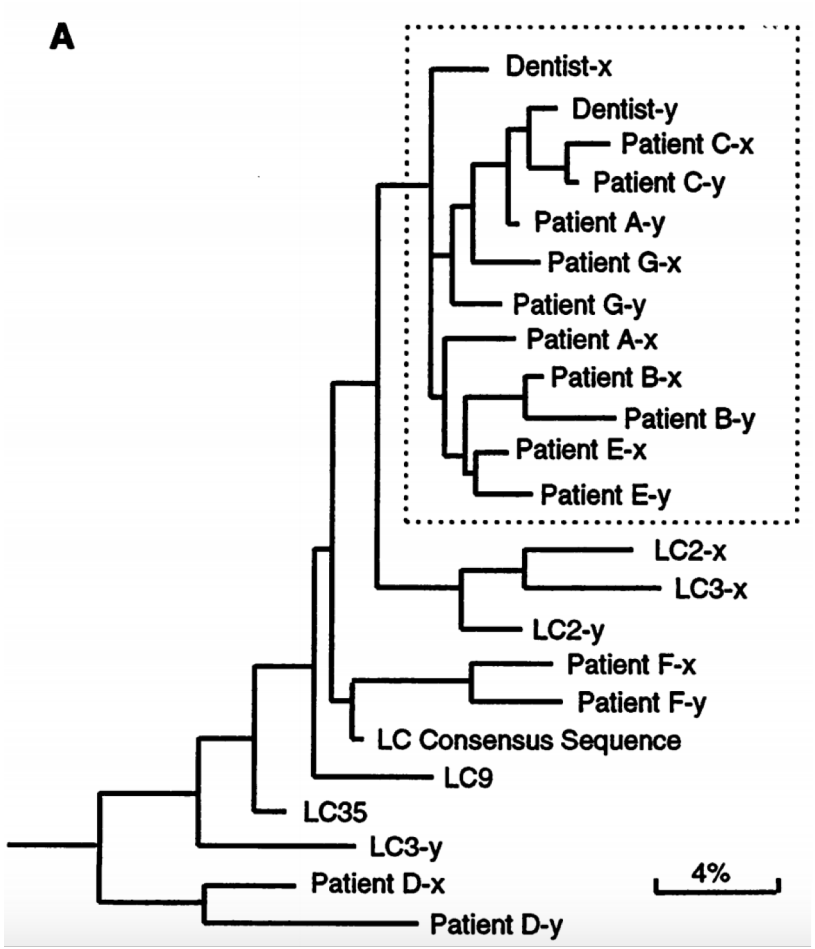


What are phylogenetic trees?

- A graphical summary of the ancestral relationships between organisms
- The reproductive links from individuals, to populations, to species, to all biological diversity



Hug et al., Nature Microbiology, 2016



Why would we use phylogenies in epidemiology?

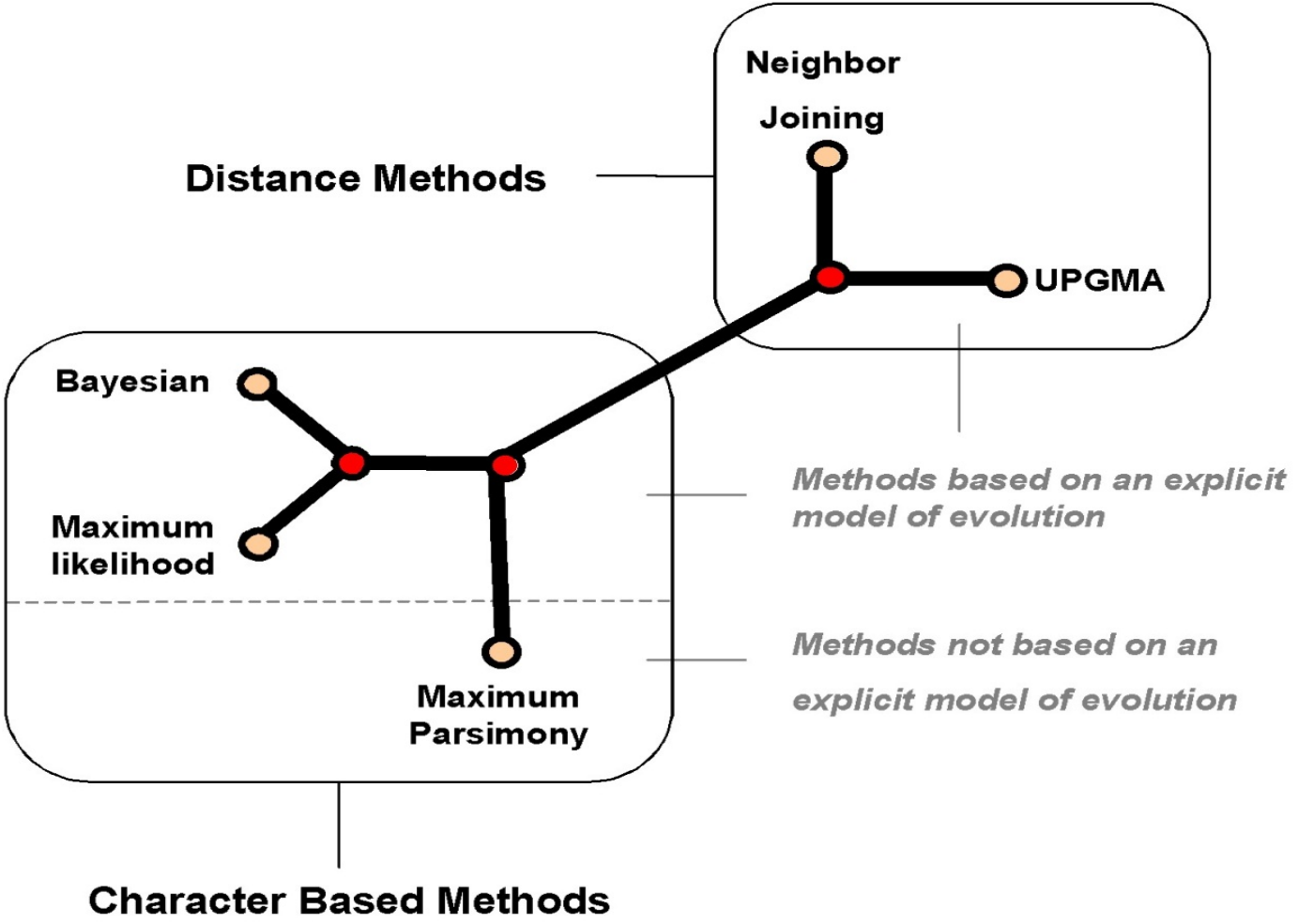
Phylogenies help reduce epi uncertainty.

Science. 1992 May 22;256(5060):1165-71. Paperpile

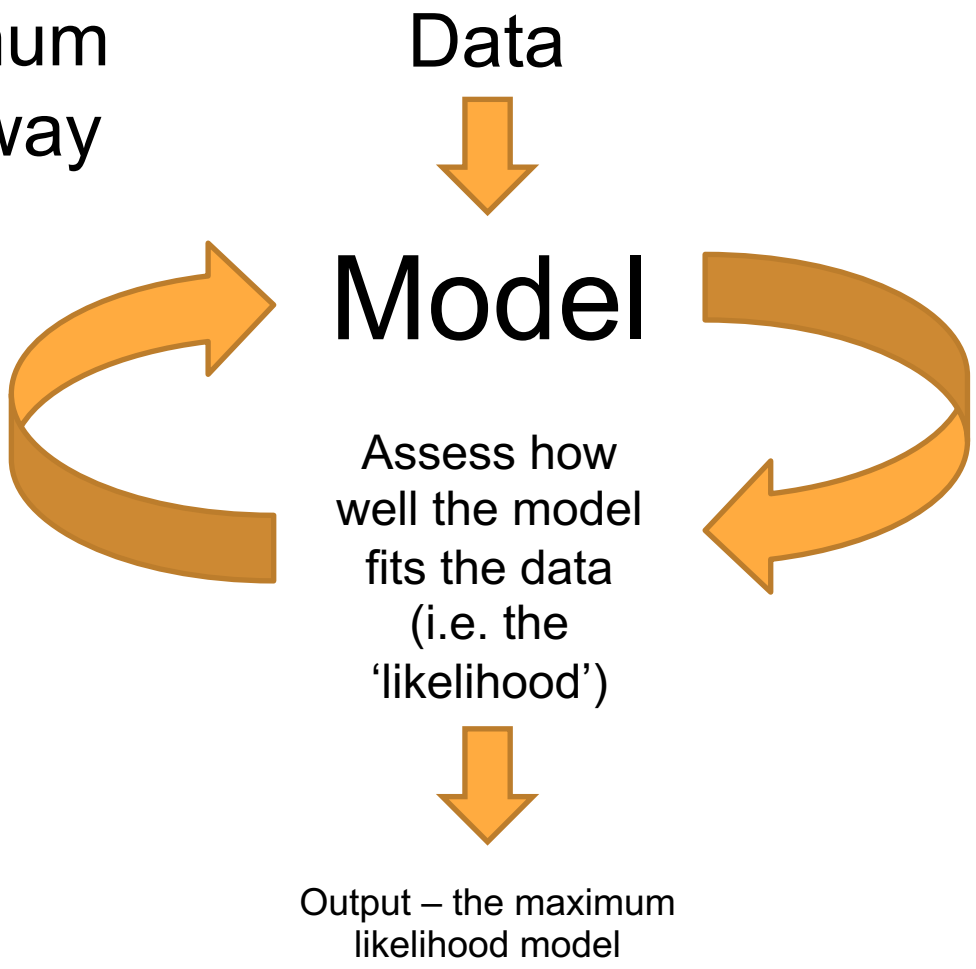
Molecular epidemiology of HIV transmission in a dental practice.

Ou CY[1], Ciesielski CA, Myers G, Bandea CI, Luo CC, Korber BT, Mullins JI, Schochetman G, Berkelman RL, Economou AN, et al.

How do we make a tree?


Neighbor
Joining
Distance Methods
UPGMA
Bayesian
Maximum
likelihood
Methods based on an explicit
model of evolution
Methods not based on an
explicit model of evolution
Maximum
Parsimony
Character Based Methods

The maximum likelihood way



Data

- Usually sequence data
- Can be physical features e.g. does it have a tail, etc.

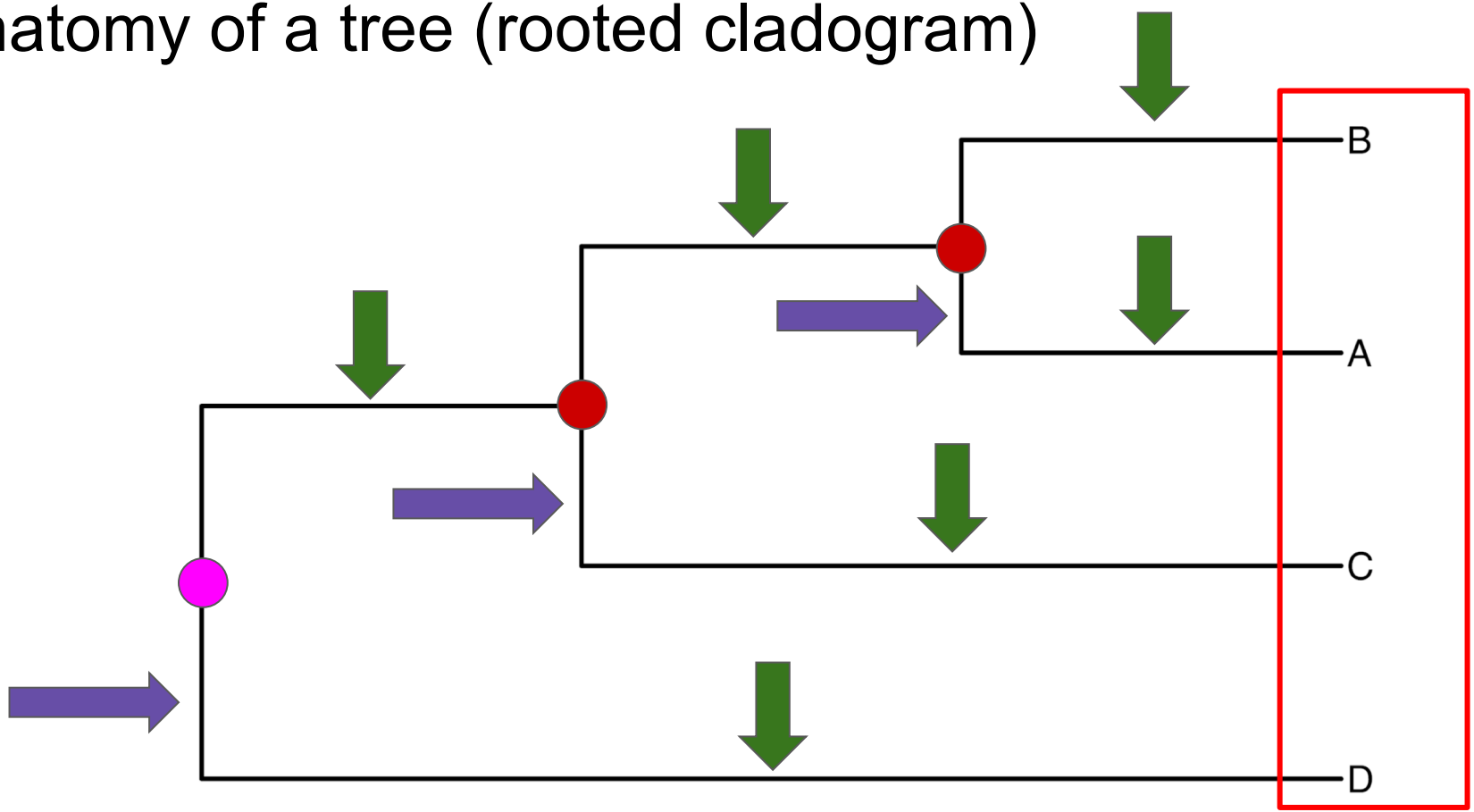
Model

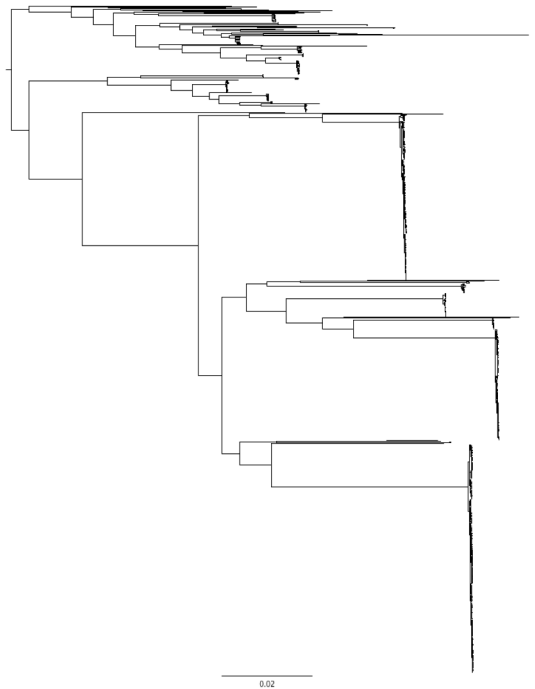
- Assumes tree like evolution
- Contains a few other key features we will go through

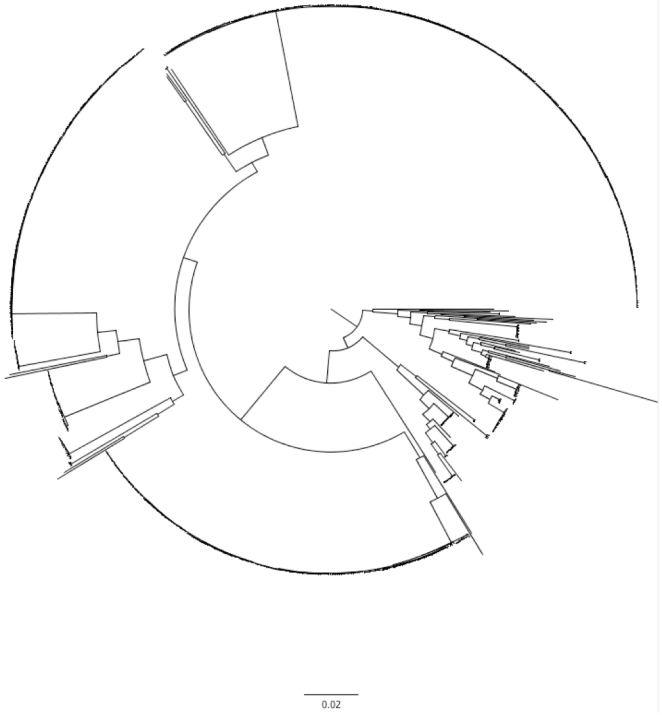
Number of tips	Number of Trees
3	3
4	15
5	105
10	34,459,425
20	8.200795e+21
30	4.951798e+38

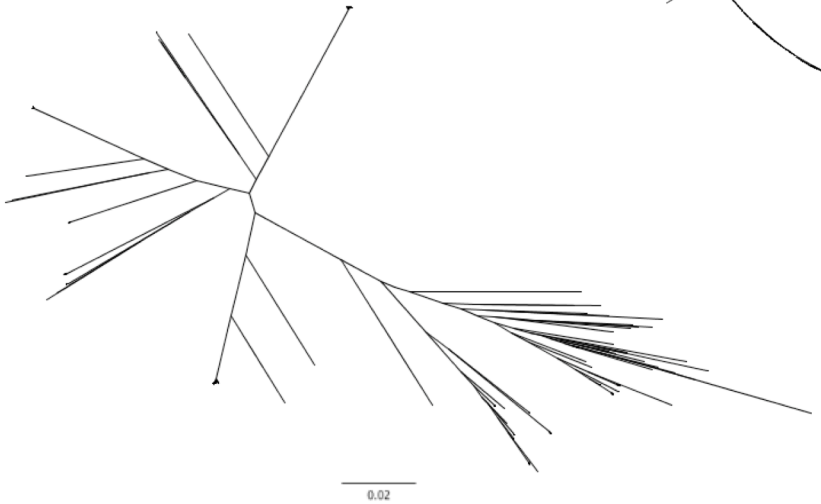
Topologies

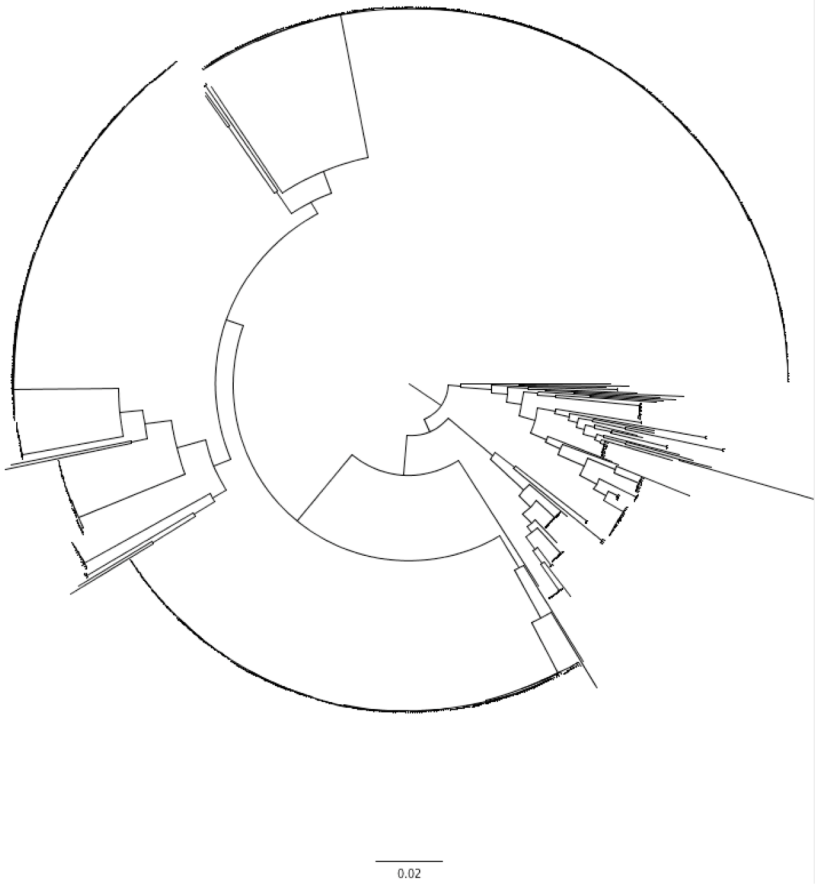
Anatomy of a tree (rooted cladogram)

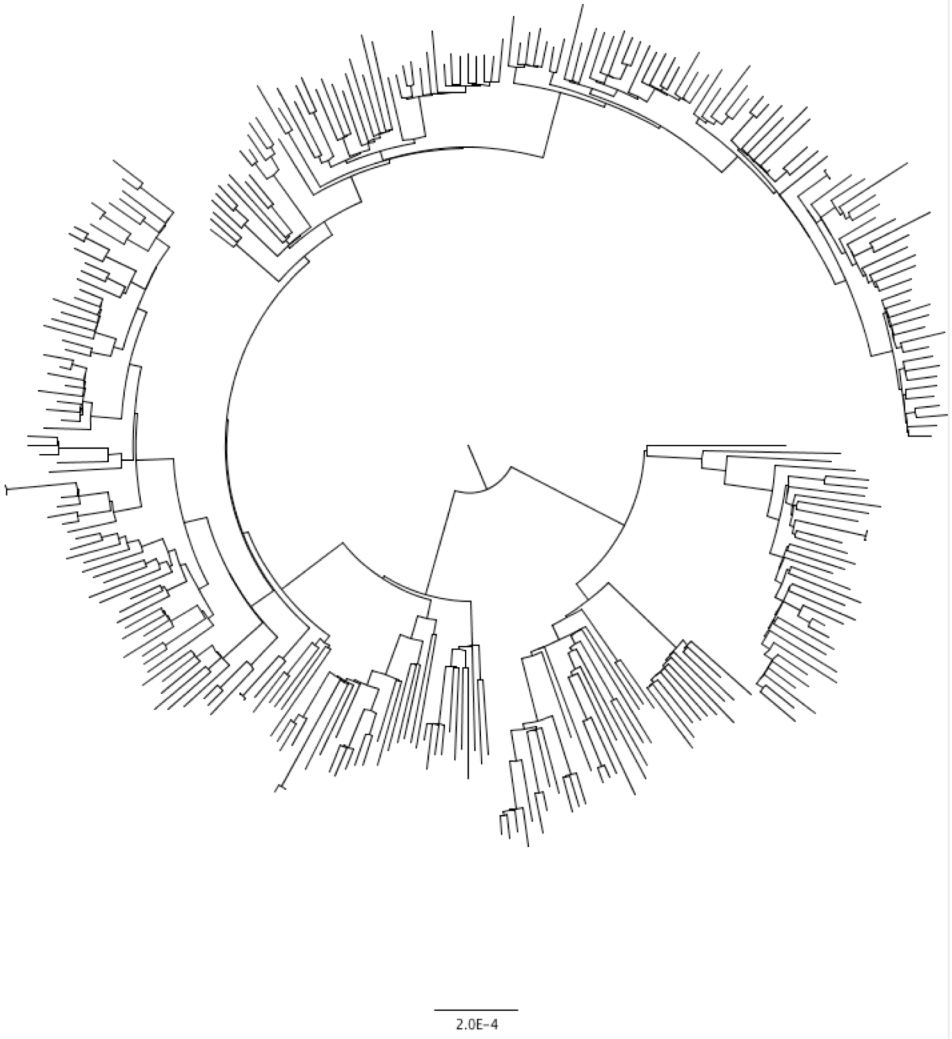



0.02


0.02

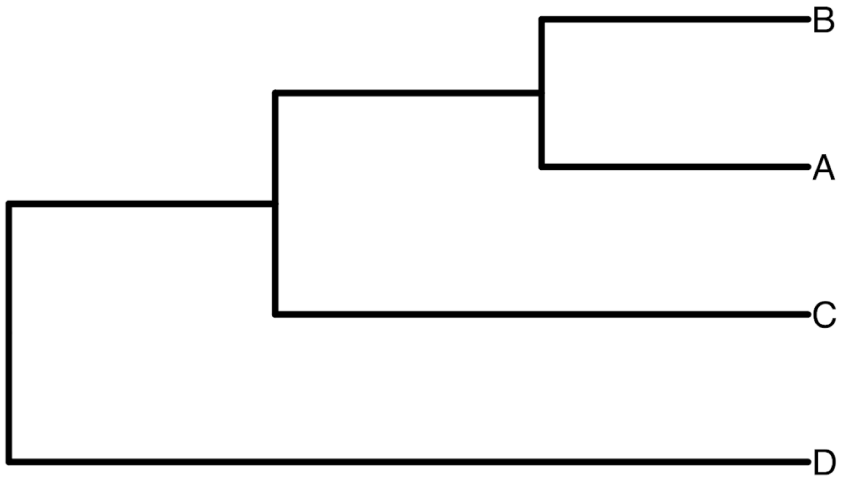

0.02

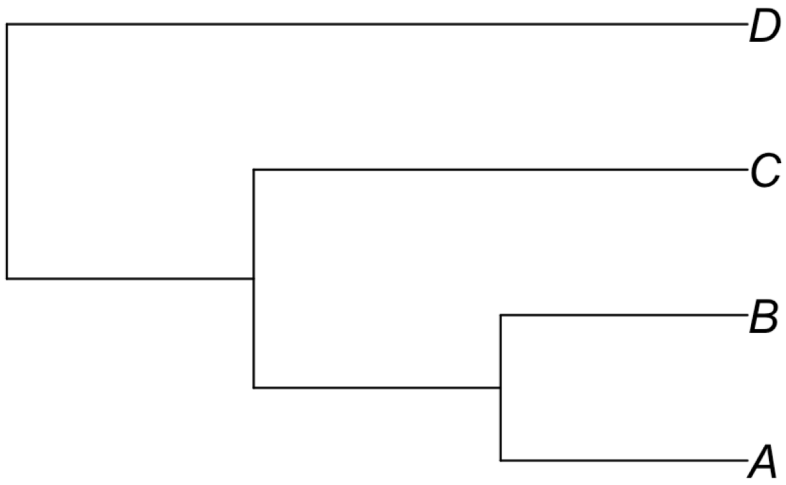

0.02

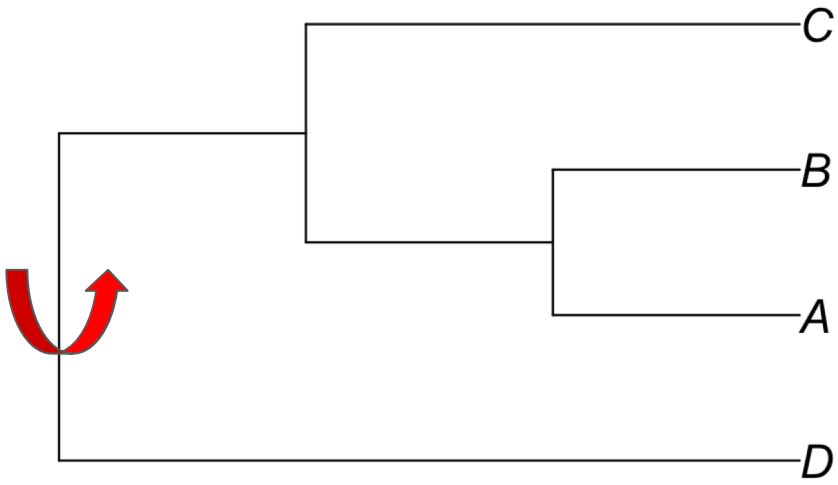

2.0E-4

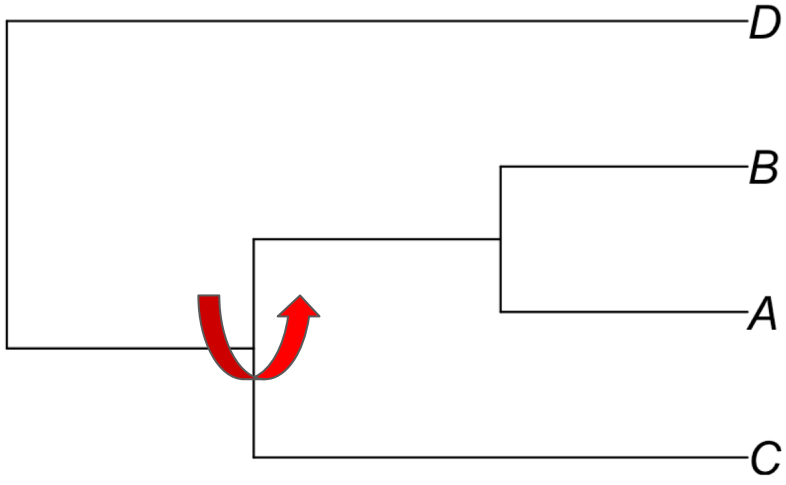
Additional representations: Newick Format

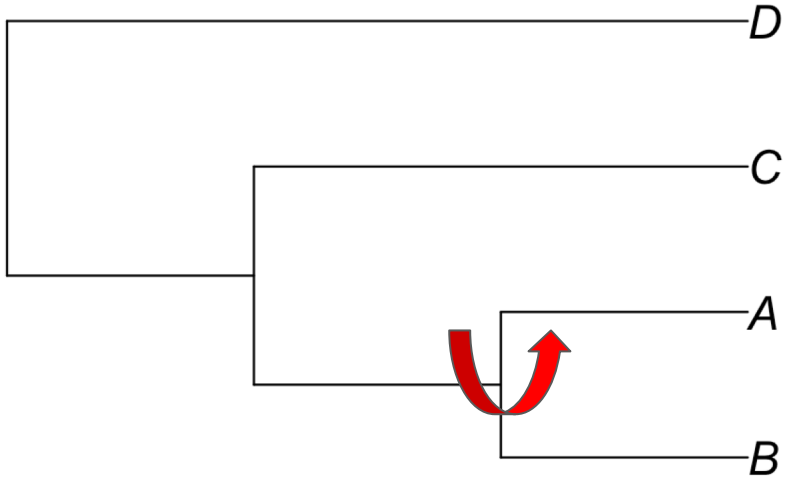
(((B, A), C), D)



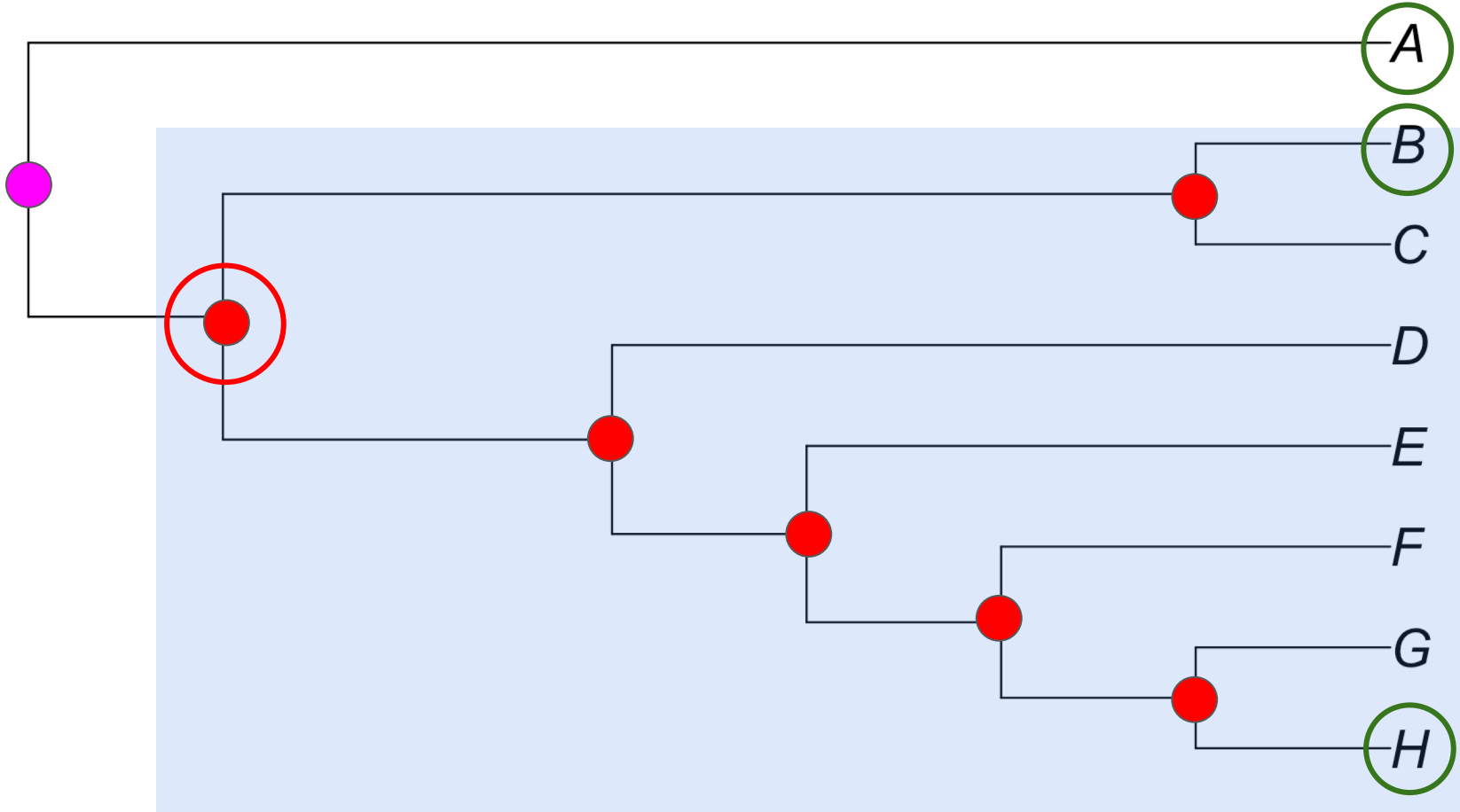

D
C
B
A


C
B
A
D


D
B
A
C

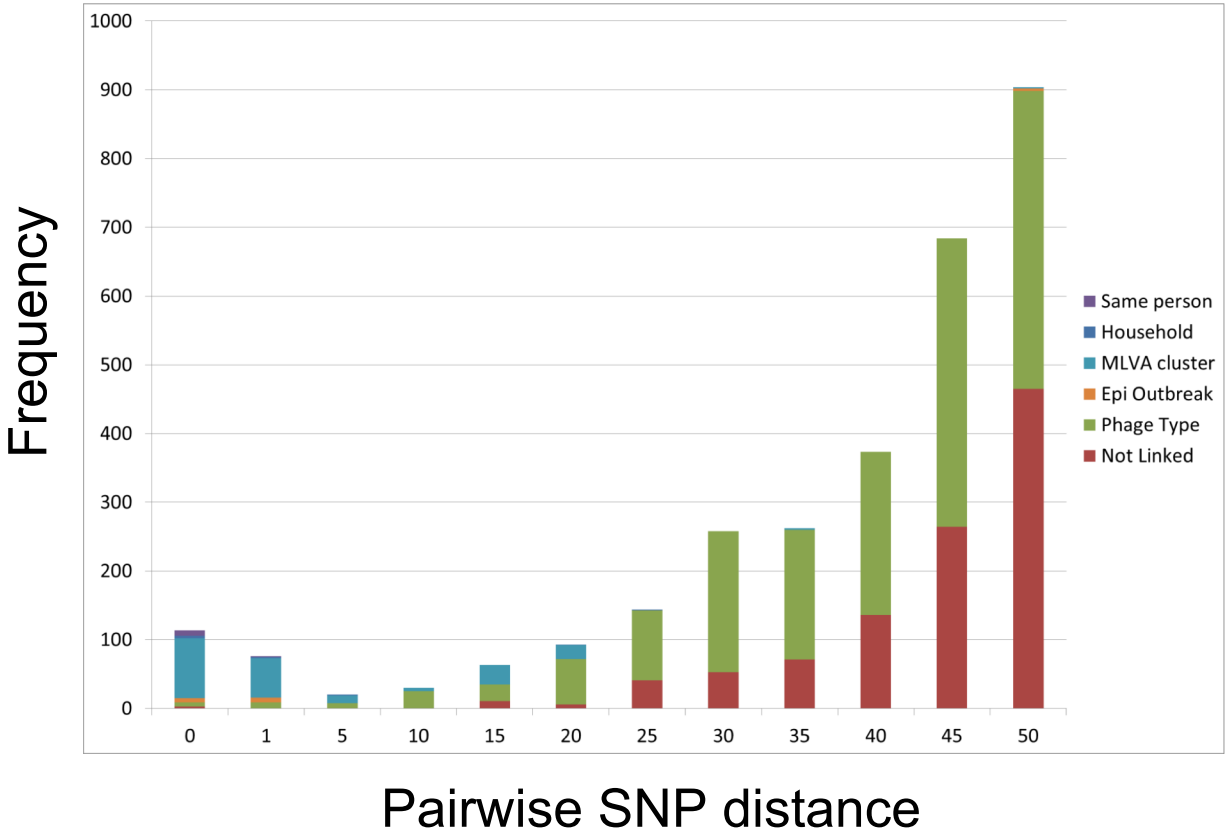

D
C
A
B

Is B closest to A or H?



Monophyletic clade: a group of genomes that consists of all the descendents of a common ancestor

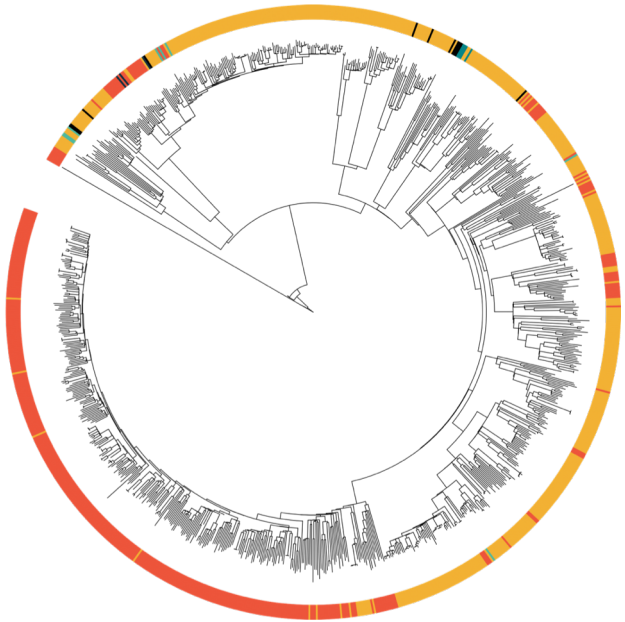
Pairwise SNP distance interpretation

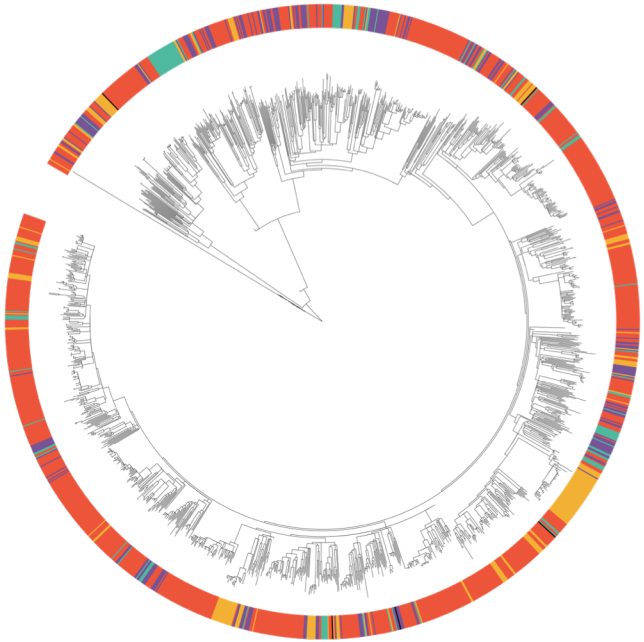


Shiga Toxin producing *E. coli* O157

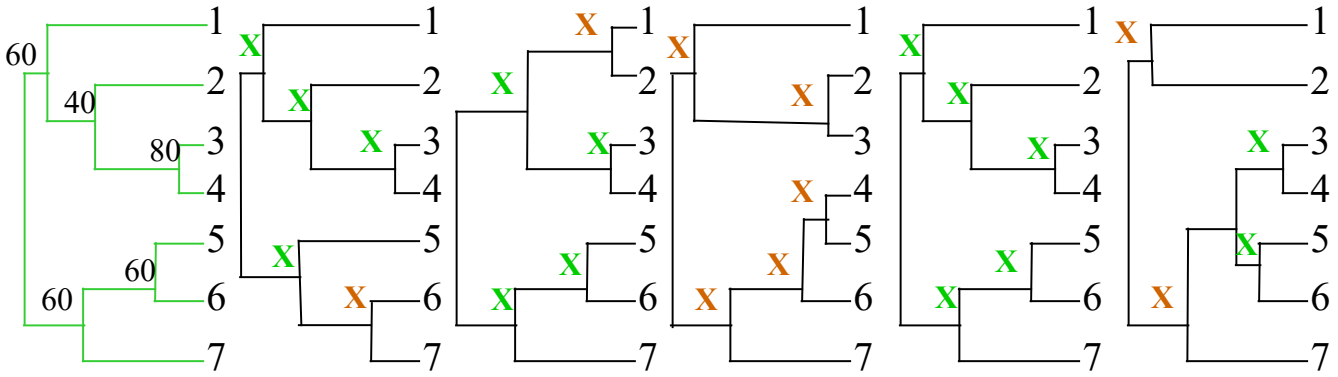
How to identify phylogenetic signal?

“tendency for related species (isolates) to resemble each other more than they resemble species drawn at random from the tree” - Blomberg and Garland, 2002





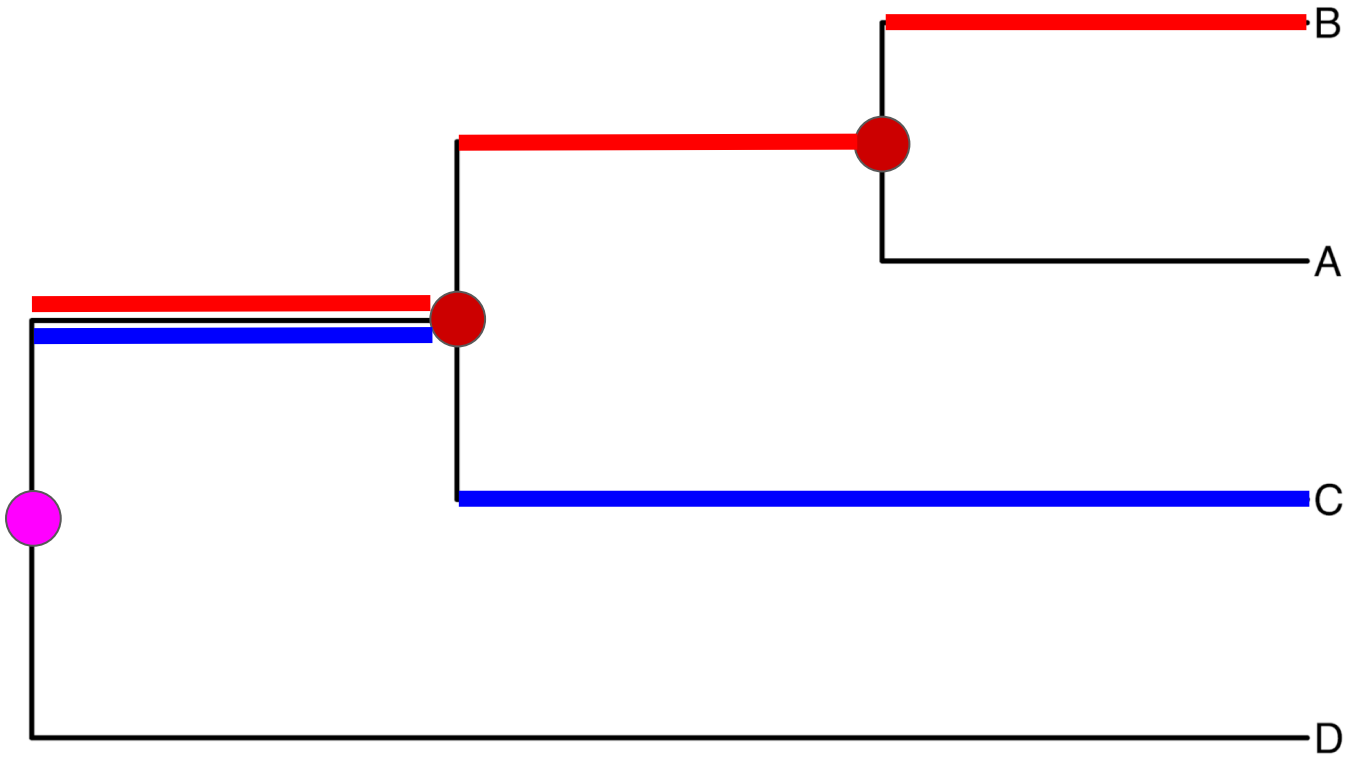
Bootstraps



A well predicted branch would recur >70-80 / 100 simulations

The End

Anatomy of a tree



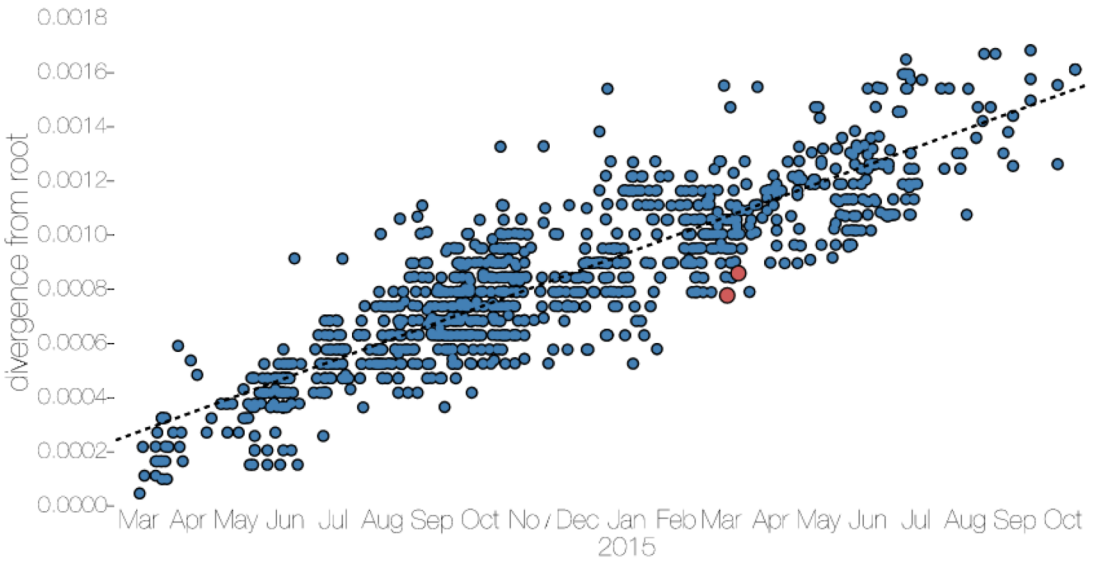
Common mistakes and how to avoid them

1. Common mistakes
    a. Looking along the tips --- order is arbitrary (remember rotations)
    b. Counting nodes
    c. Perceived notions of relationship
2. Avoid them by
    a. Sign-post method
    b. Grouping method

Timed phylogenies

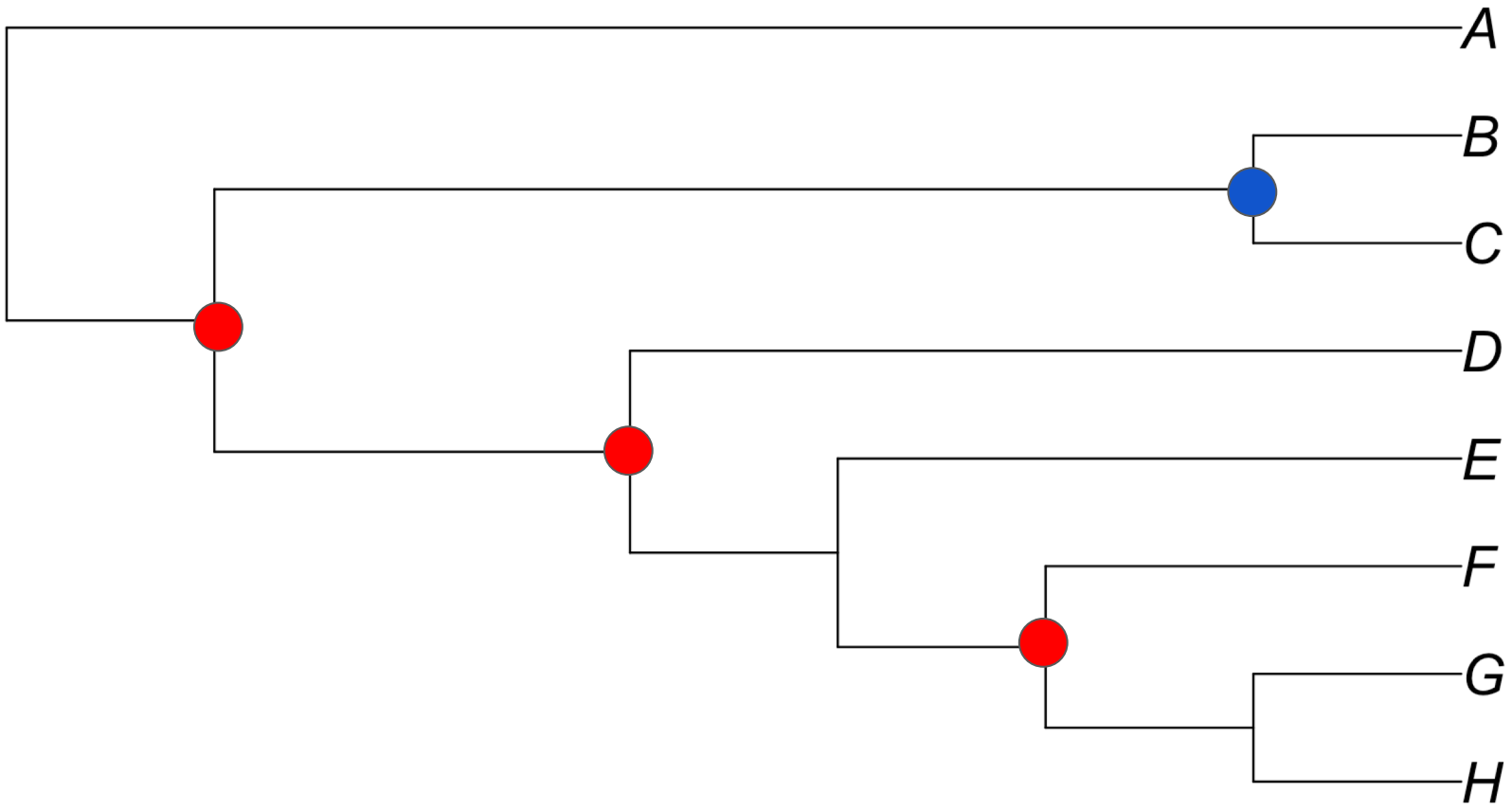
BEAST is a commonly used tool for putting dates on phylogenies. However, it assumes that there is a molecular clock. Tempest, from Andy Rambaut can show you whether there is clock like evolution.

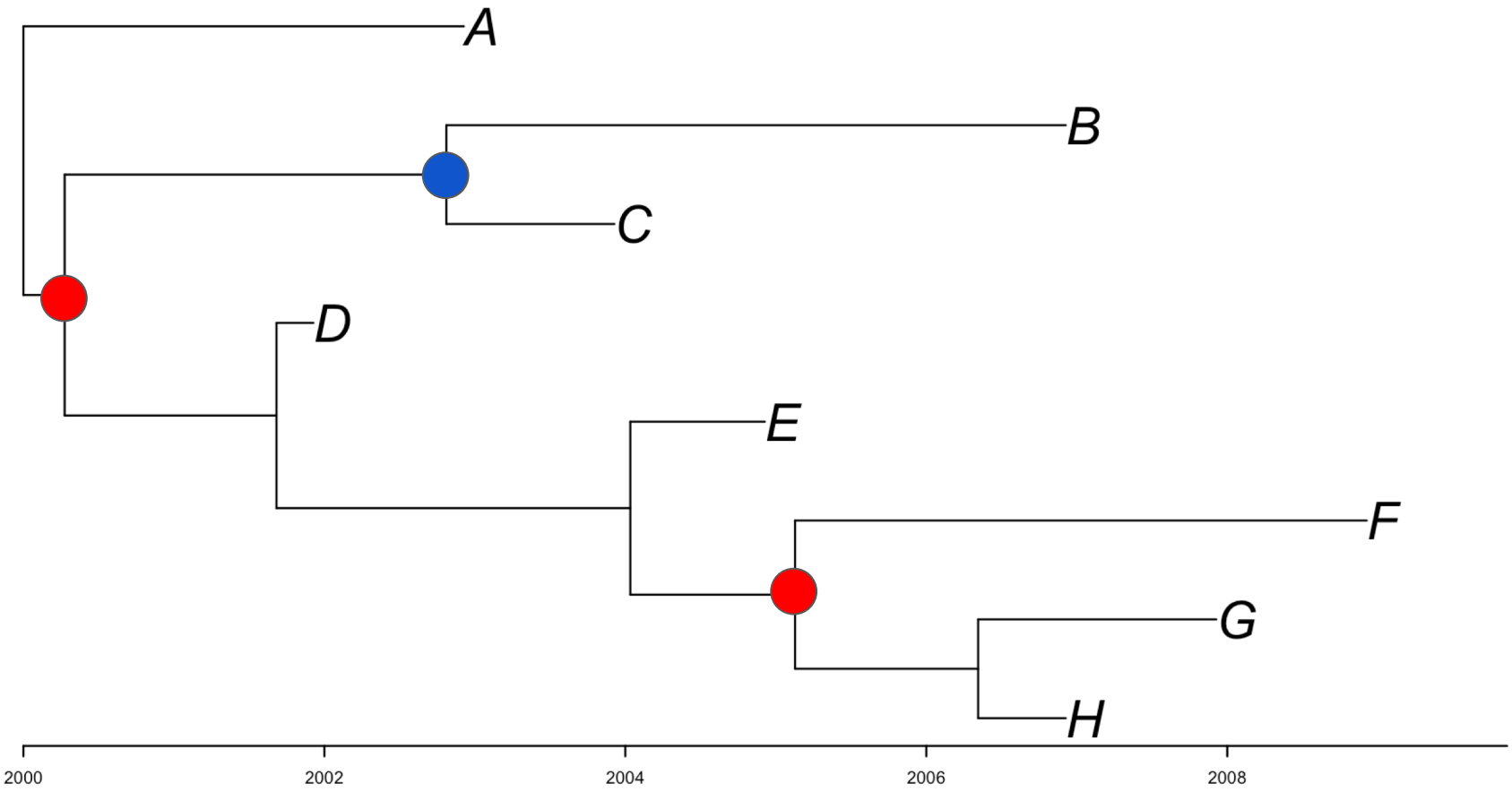
Tip date randomization is another method to identify clock-like evolution.



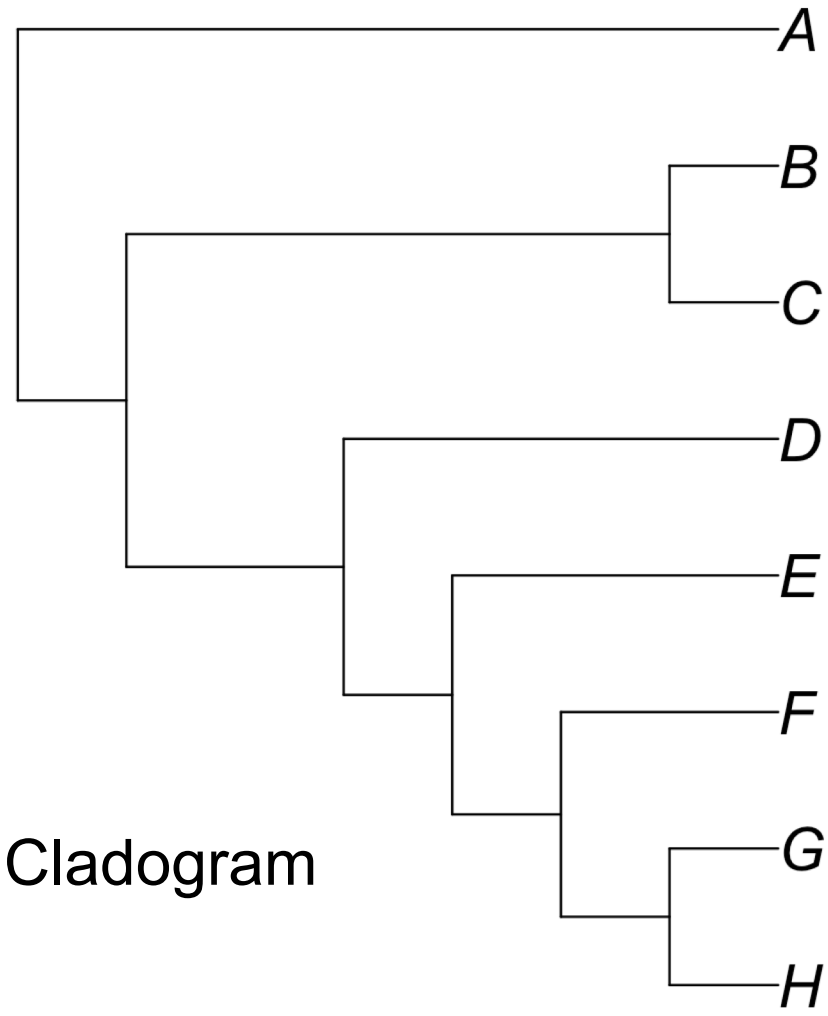
EBOV evolution - Eddie Holmes

Which split happened first?

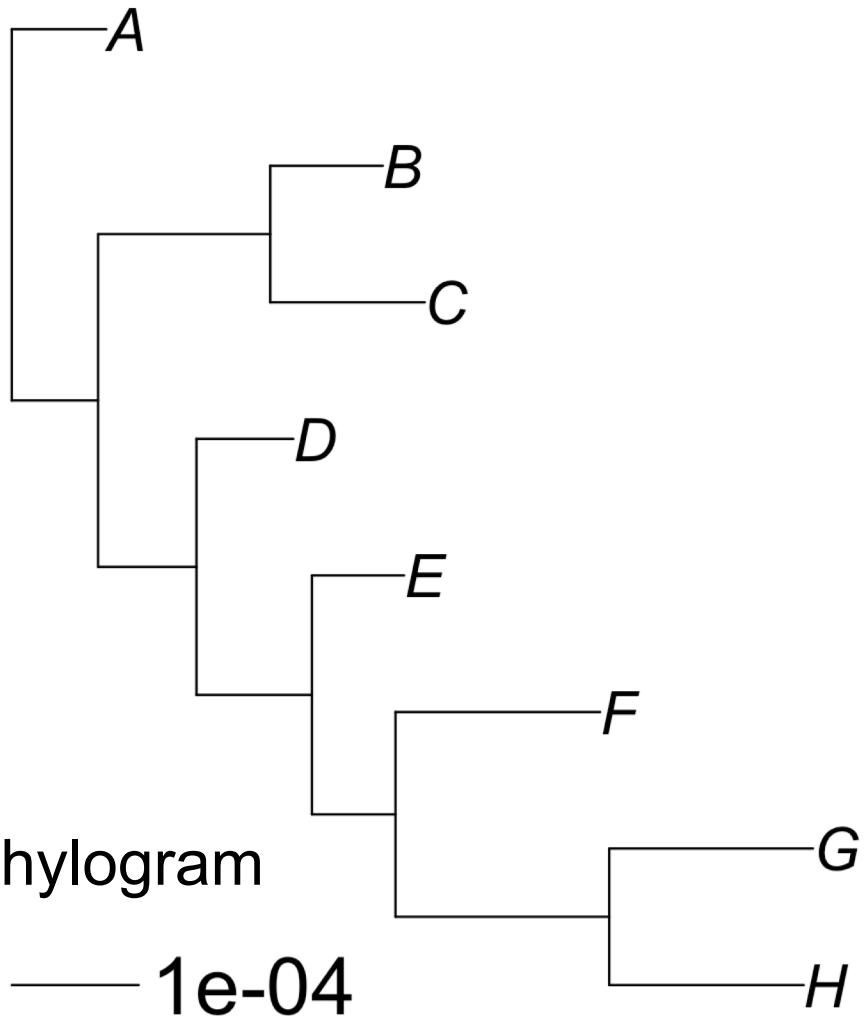



A
B
C
D
E
F
G
H
2000
2002
2004
2006
2008

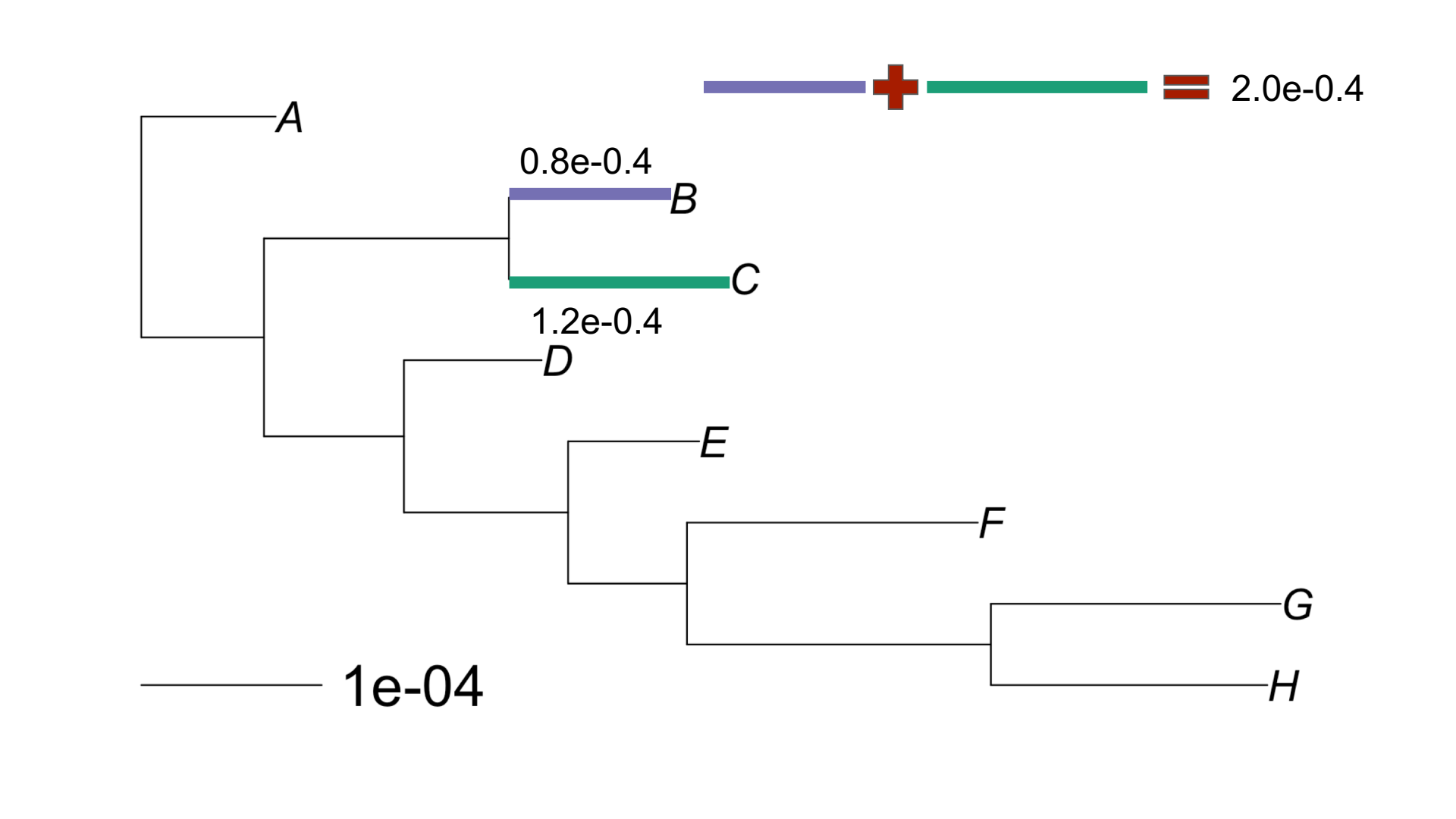
Beyond topologies --- what are the branch lengths all about?



Cladogram



Phylogram


2.0e-0.4
A
0.8e-0.4
B
C
1.2e-0.4
D
E
F
G
H
1e-04

Branch lengths

1. Depend on the evolutionary model
2. Gives the measure of the amount of evolution that has happened at a branch
3. In the case of maximum likelihood trees, they represent the expected (or average) number of substitutions along the branch
4. We sum branch lengths to measure the distance between pairs of nodes

Naive branch length to SNP calculation

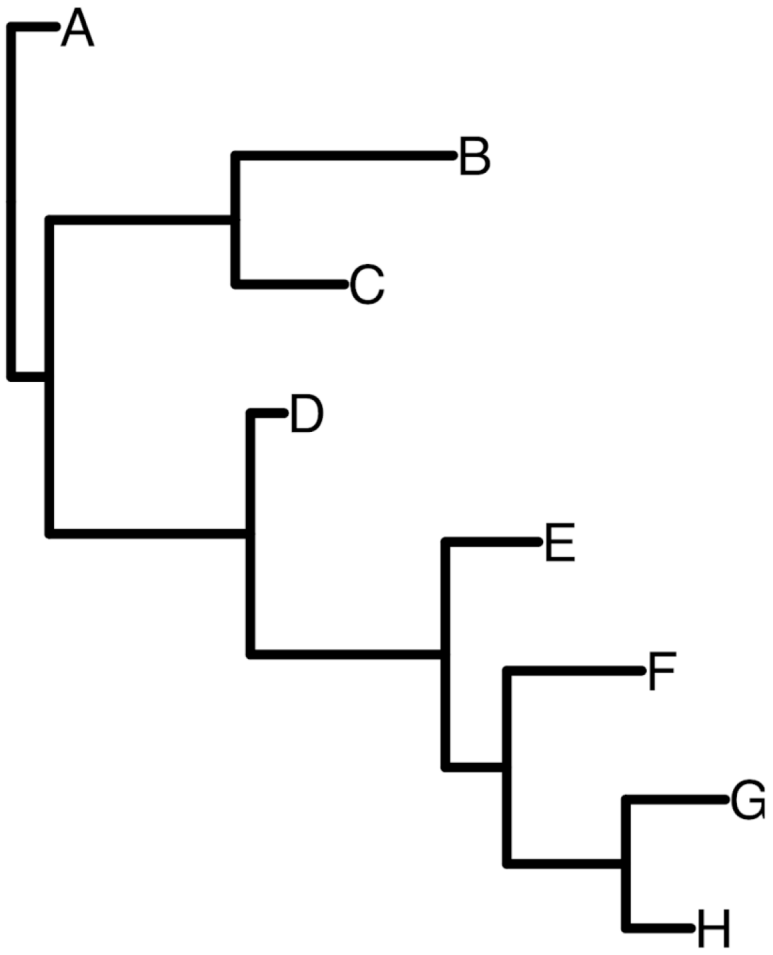
Need:

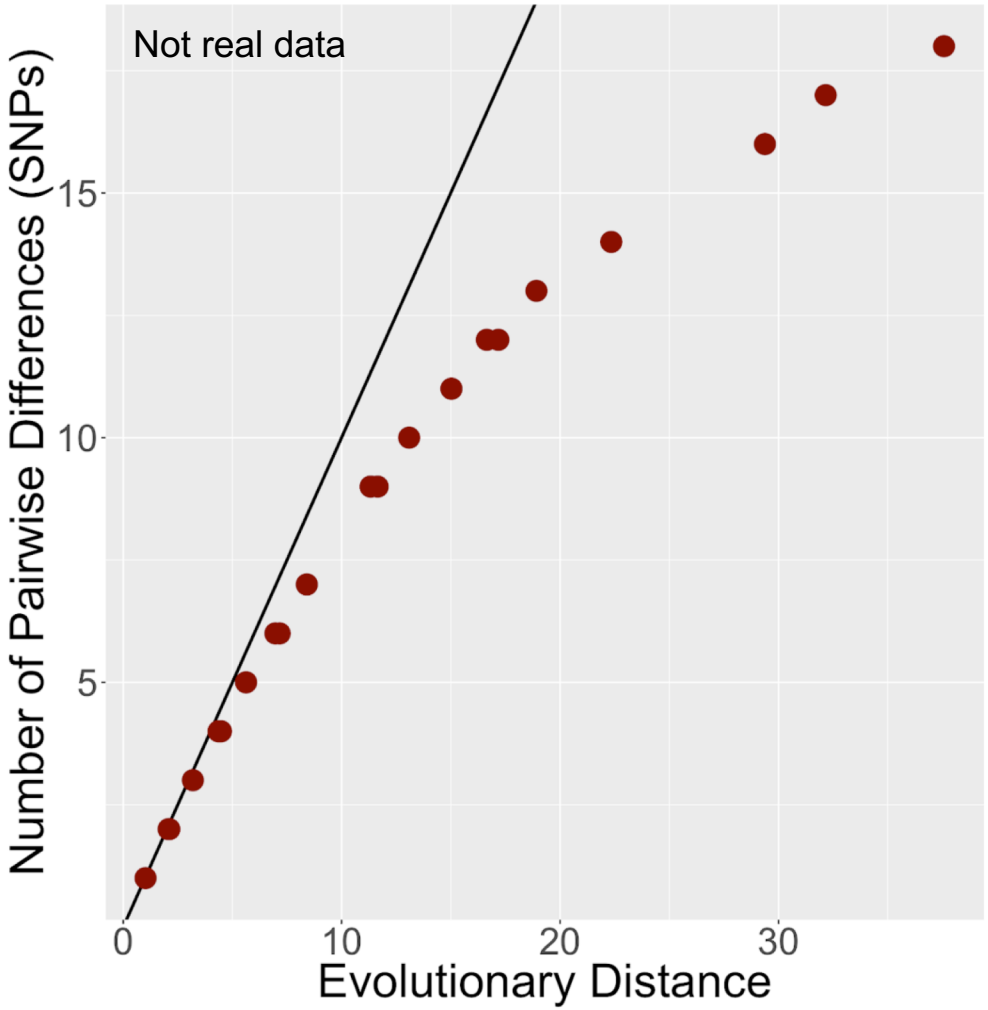
- Scale bar
- Number of SNPs in the alignment
- Total length of tree

Number of SNPs = (Distance on tree / total length of tree)
* Number of SNPs in the alignment

Pairwise distances

1. Count of differences (i.e., pairwise SNPs)
2. Estimate of the number of substitution events
    a. Parsimony
    b. Maximum likelihood/Bayesian model


A
B
C
D
E
F
G
H


Not real data
Number of Pairwise Differences (SNPs)
15
10
5
0
10
20
30
Evolutionary Distance

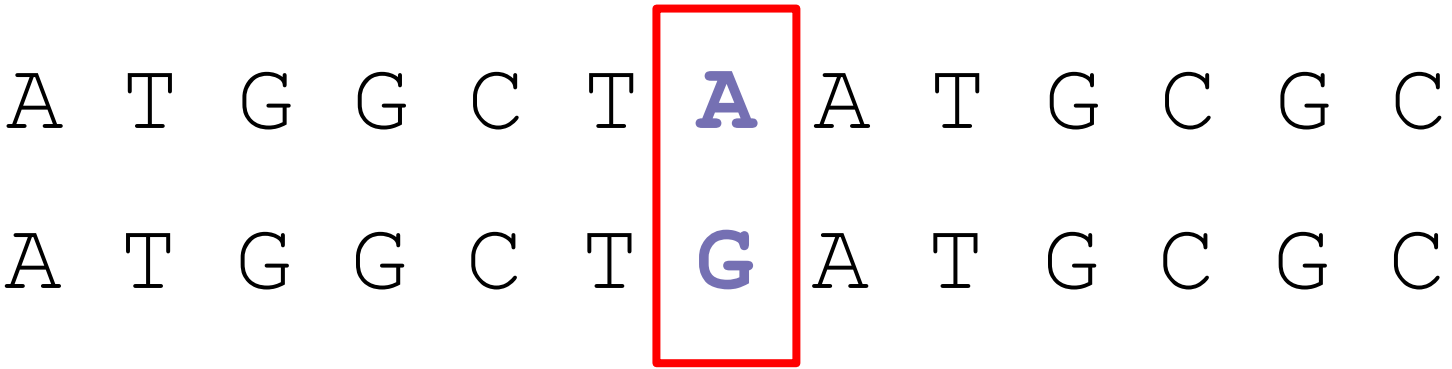
What is happening?

1. Sequences are **FINITE**
2. Leads to **MUTATION SATURATION**
3. This means **MULTIPLE HITS** at a site
4. This means that **OBSERVED PAIRWISE DIFFERENCES** are likely an **UNDERESTIMATE** of the total divergence

Models of evolution

- Parsimony
  - **Goal:** minimise the number of steps
- DNA or Amino acid substitution models
  - **Goal:** Correct the observed differences pairwise differences --- or estimate the actual number of substitutions that have actually happened
  - *Underpins most modern tree inference software*

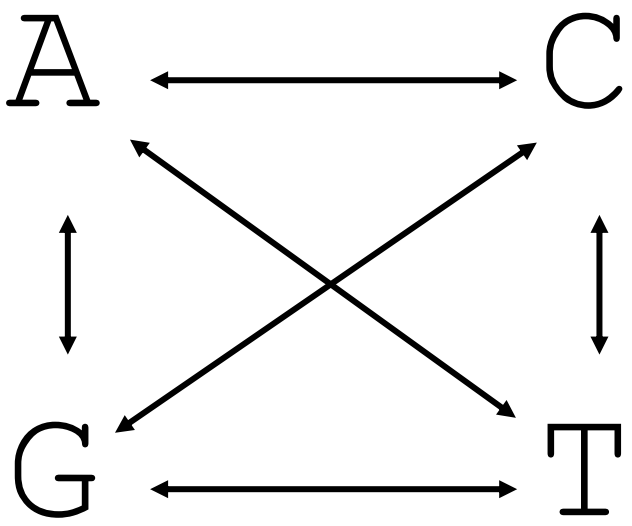
What is a SUBSTITUTION?



Loads of mathematics!!
(https://en.wikipedia.org/wiki/Substitution_model)

But, let us get a gut feeling for the process!

A T G G C T **A** A T G C G C



Prob(SUBSTITUTION)

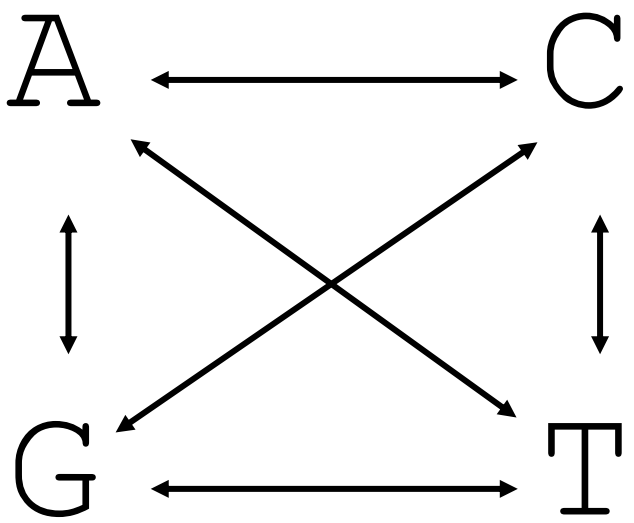
Common parameters in substitution models

1. Base frequencies
2. Transition/Transversion ratio
3. Gamma distributed among site rate heterogeneity
4. Proportion of invariant sites

Jukes-Cantor 1969 (JC69)

Assumptions

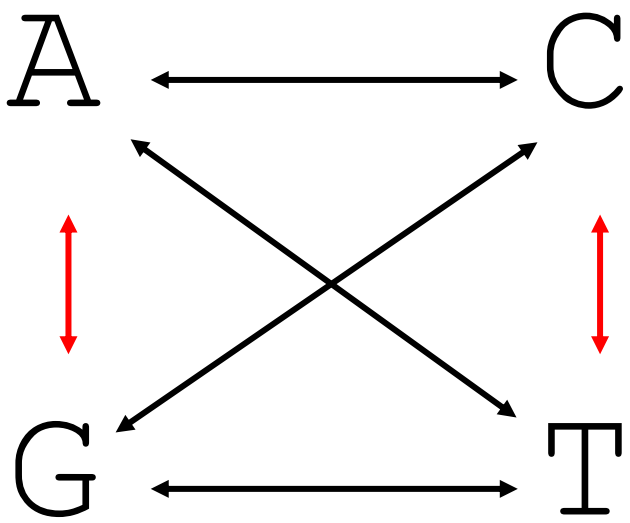
- All BASE FREQUENCIES are the same:
  - A = T = C = G = 0.25
- A SINGLE SUBSTITUTION RATE



Kimura 1980 or Kimura-2-Parameter (K80)

Assumptions

- All BASE FREQUENCIES are the same:
  - A = T = C = G = 0.25
- Allows for unequal transition/transversion ratio
- Transition:
  - A <> G or C <> T
- Transversions:
  - The rest

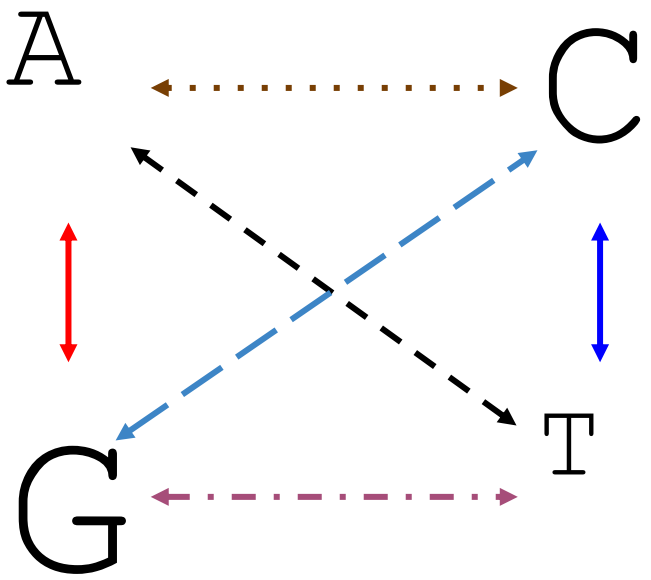


A T G G C T **A** A T G C G C

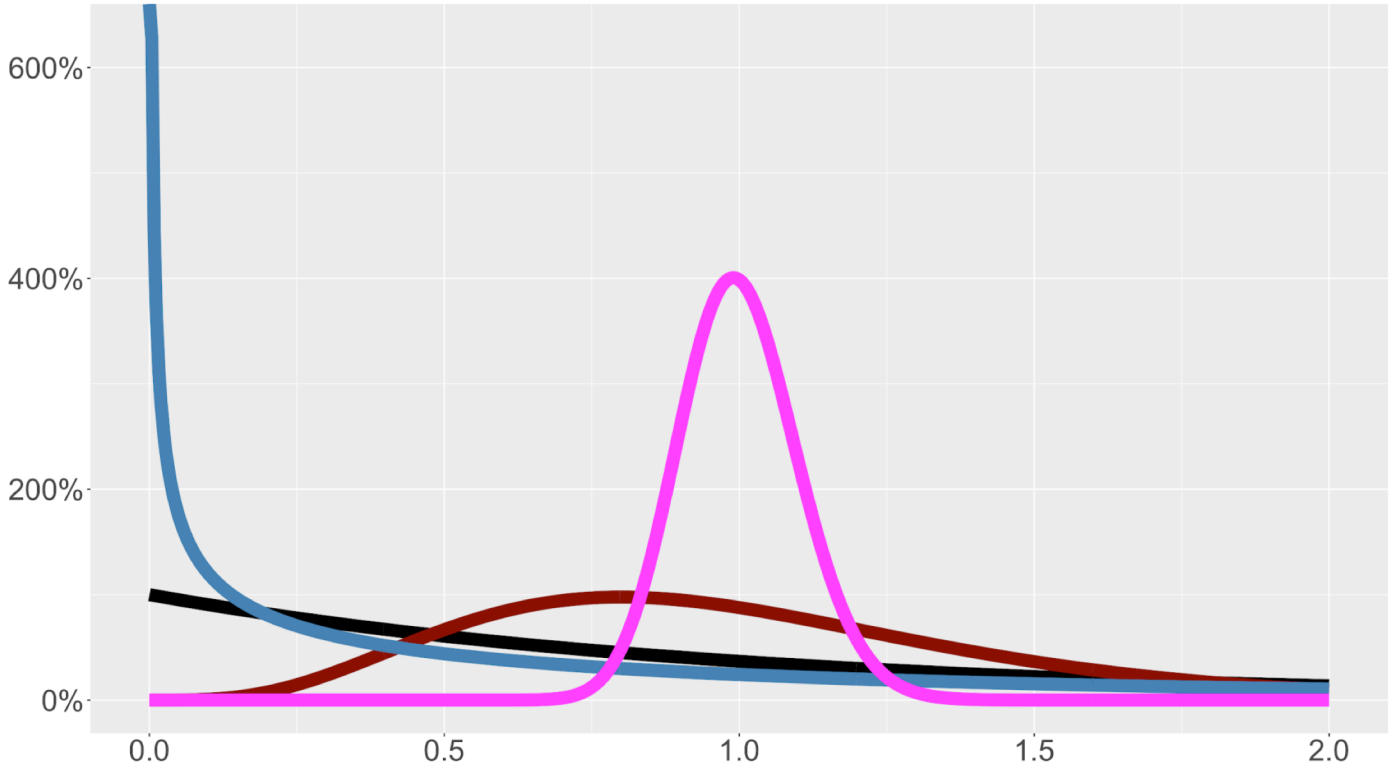
Generalised Time Reversible (GTR)

Assumptions

- BASE FREQUENCIES are different:
  - A ≠ T ≠ C ≠ G
- Allows for individual substitution rates



Among site rate heterogeneity (Γ)



Proportion Invariant Sites (I)

GTR + Γ + I

Pairwise SNP distance calculation

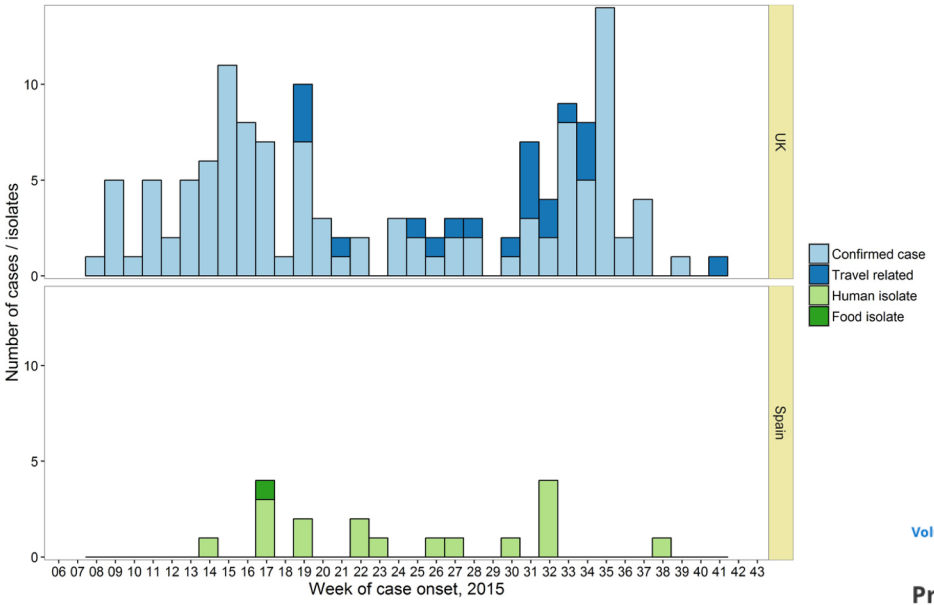
	Strain1	Strain2	Strain3	Strain4	Strain5	Strain6	Strain7	Strain8	Strain9
Strain1	0	161	136	134	118	192	104	110	195
Strain2	161	0	107	105	89	99	79	81	102
Strain3	136	107	0	68	52	138	54	44	141
Strain4	134	105	68	0	34	136	52	26	139
Strain5	118	89	52	34	0	120	36	8	123
Strain6	192	99	138	136	120	0	110	112	9
Strain7	104	79	54	52	36	110	0	28	113
Strain8	110	81	44	26	8	112	28	0	115
Strain9	195	102	141	139	123	9	113	115	0

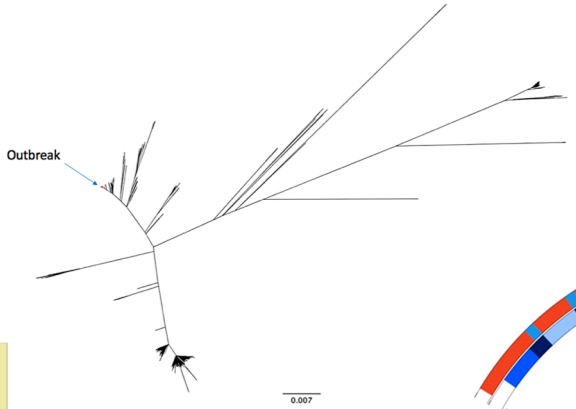
Recombination

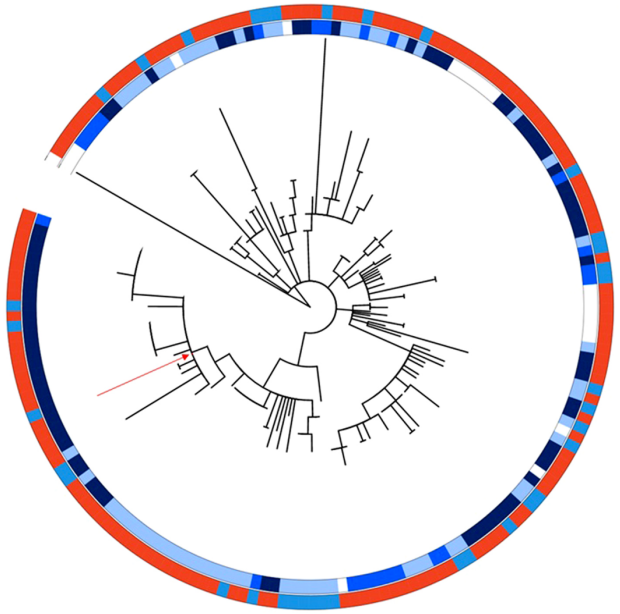
How clonal are bacterial populations?

- Clonal – acquired DNA vertically from parent without recombination (sexual or otherwise)
- Most phylogenetic trees assume that DNA has been transmitted vertically (no recombination)
- Some bacteria undergo a lot of recombination, others are very clonal. Smith, JM et al., 1993, PNAS, v90, p4384
- We can use Gubbins or ClonalFrameML to identify regions of recombination.
- Can use a phylogenetic network approach (SplitsTree) to analyse your data
- You need to know for your bacterium if it is recombinogenic or not, and what the best ways to deal with that level of recombination are
- Mobile elements like phage and plasmids should usually be excluded, or treated separately in a phylogenetic analysis

Why would we use phylogenies in epidemiology?







Volume 145, Issue 2 pp. 289-298

Cited by 27

Access

Prospective use of whole genome sequencing (WGS) detected a multi-country outbreak of *Salmonella* Enteritidis

T. INNS [a1] [a2] [a3], P. M. ASHTON [a4], S. HERRERA-LEON [a5], J. LIGHTHILL [a6], S. FOULKES [a1], T. JOMBART [a7], Y. REHMAN [a1], A. FOX [a8], T. DALLMAN [a3] [a4], E. DE PINNA [a4], L. BROWNING [a9], J. E. COIA [a10], O. EDEGHERE [a1] and R. VIVANCOS [a1] [a2] [a3]