bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io


Creative Commons
This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu
creative commons
Attribution-Share Alike 2.5 Canada
You are free:
to Share — to copy, distribute and transmit the work
to Remix — to adapt the work
Free Cultural APPROVED FOR Works
Under the following conditions:
Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.
For any reuse or distribution, you must make clear to others the licence terms of this work.
Any of the above conditions can be waived if you get permission from the copyright holder.
The author's moral rights are retained in this licence.
Disclaimer
Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French
Learn how to distribute your work using this licence

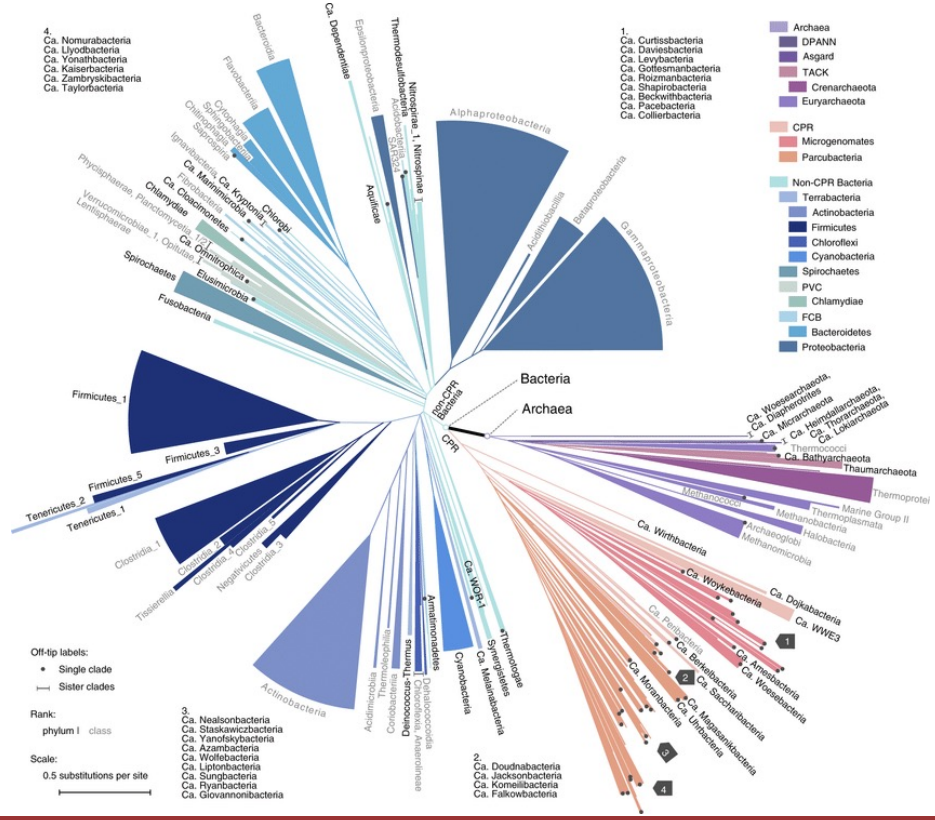
Module 2
Phylogenetics

Fiona Brinkman
Infectious Disease Genomic Epidemiology
April 18-21, 2023

bioinformatics.ca

Department of Molecular Biology and Biochemistry
School of Computing Science
Faculty of Health Sciences

SFU SIMON FRASER UNIVERSITY



Zhu et al 2019. *Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea.* Nature Communications
https://doi.org/10.1038/s41467-019-13443-4

Learning Objectives

By the end of this lecture, you will:

- Understand the fundamentals of character-based evolutionary analysis and phylogenetic analysis
- Learn how to interpret a phylogenetic tree
- Know the basics of how to build a phylogenetic tree
- Appreciate key differences between methods used to build a phylogenetic tree

(slide deck adopted from presentations by Gary Van Domselaar and Will Hsiao)

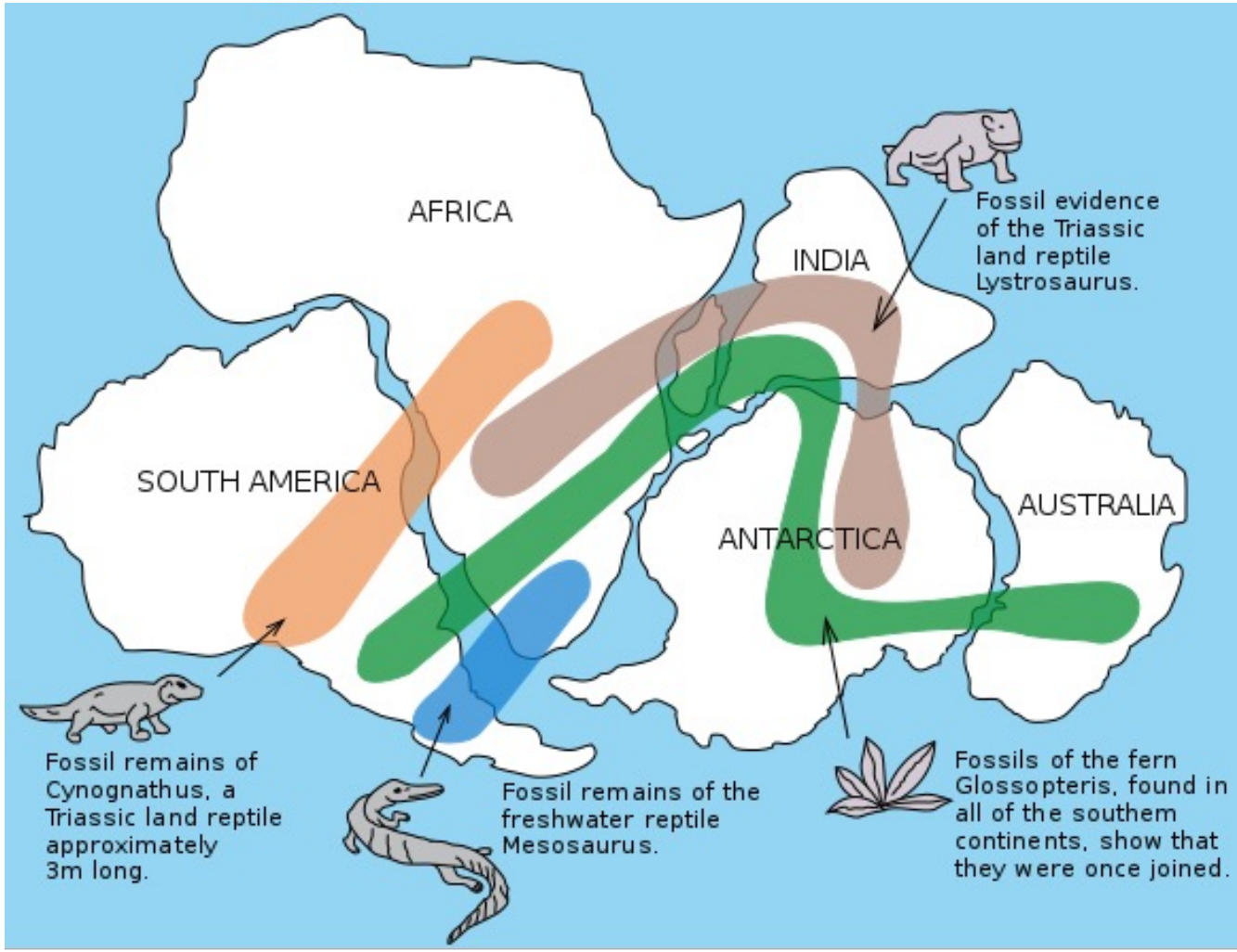
How did evolutionary theory evolve?

Appreciation that the world is not constant, but changing

Plate tectonics appreciated

Discoveries of fossils accumulated

- Remains of unknown but still living species that are elsewhere on the planet?
- Cuvier (circa 1800): the deeper the strata, the less similar fossils were to existing species



ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;

AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

LONDON:

JOHN MURRAY, ALBEMARLE STREET.

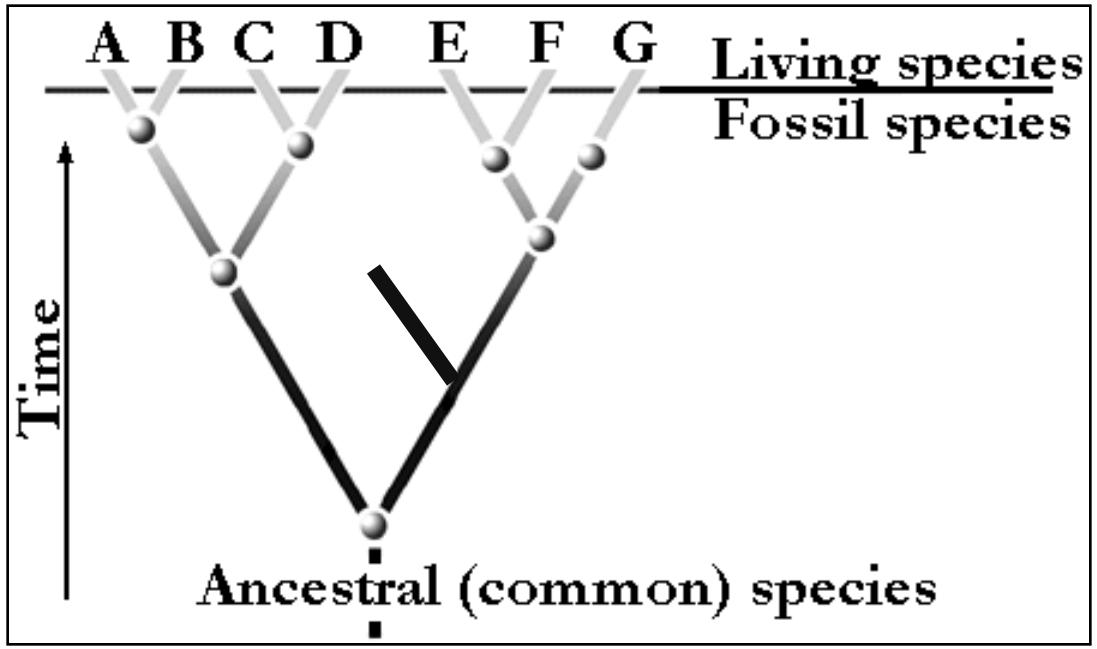
1859.

Part of Darwin's Theory

All organisms are derived from common ancestors by a process of branching.

This explained…

- The fossil record
- Similarities of organisms classified together (shared traits inherited from common ancestor)
- Similar species in the same geographic region



Our common ancestor

- Earth is thought to be approximately 4.6 billion years old

- It was estimated that life on Earth occurred as far back as 4.1 billion years
  - Carbon isotope dating
  - Fossil of microbial mat

- All cellular organisms on Earth share a common ancestor (**LUCA**) dated back to more than 3.8 billion years
  - Shared genetic code
  - Amino acid chirality

Nothing in biology makes sense except in the light of evolution – Theodosius Dobzhansky

Evolution – descent with modifications

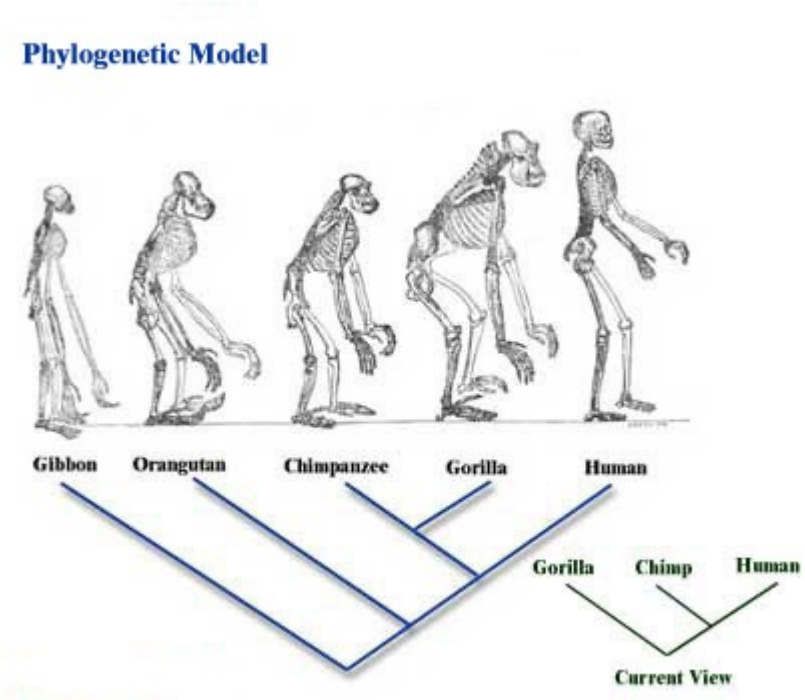
- Change in the **heritable** characteristics of biological populations over successive generations
- **Natural selection** provides one mechanism for evolution (though there is also **Neutral Evolution**)
- Evolutionary processes give rise to biodiversity
  - Estimated over 8 million living **species** (Mora et al PLoS Biol 2011)
  - ~2.3 million species recognized/named and 80% of which are cataloged in databases (https://www.catalogueoflife.org/data/metadata)
- Microbial evolutionary processes blur the boundary of species
  - Asexual reproduction (clonal)
  - Horizontal gene transfer
    - Infection by phages

Some terminology

- **Systematics**, study of the interrelationships of living things
- **Taxonomy**, the science of naming and classifying organisms *(evolutionary theory not necessarily involved)*
- **Phylogenetics**, the field of systematics that focuses on evolutionary relationships between organisms *or genes/proteins* (**phylogeny**).

Phylogenetics

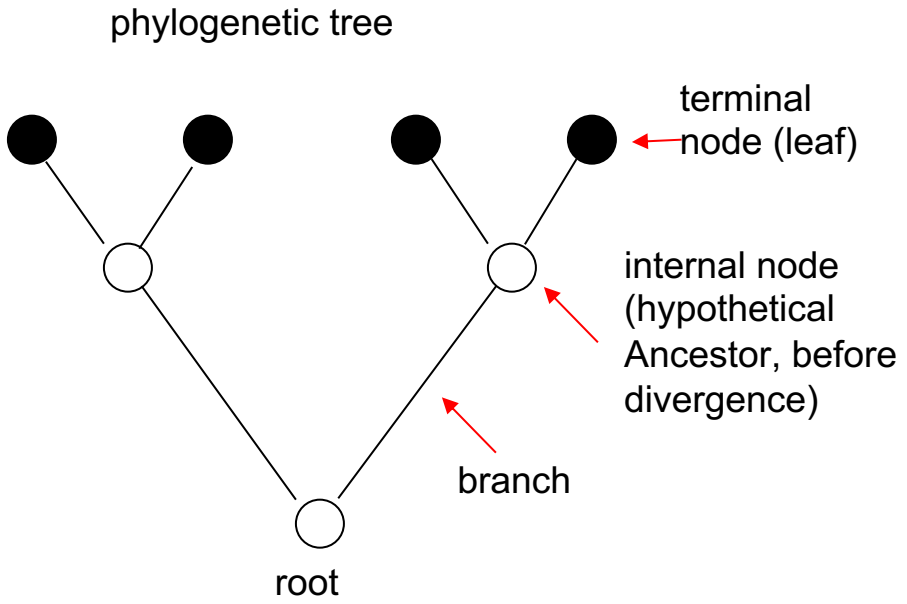
- Usually involves **molecular sequence data** (DNA and proteins), but can also be based on **morphological features** (e.g. bone structure, flower structure).
- The inferred evolutionary relationships are usually depicted as **phylogenetic trees**.
- Methods have several assumptions, such as a **bifurcating** pattern of divergence, and sequences are **homologous** (shared ancestry).



Phylogenetic tree depicting evolutionary relationship among several primate species.

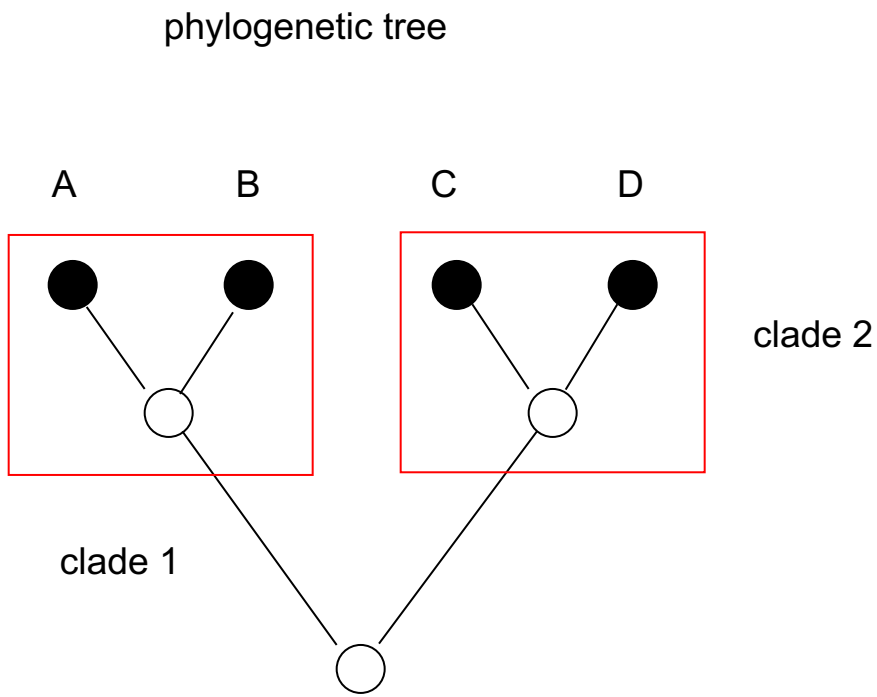
Phylogenetic Tree Terminology

- Consists of **nodes** connected by **branches**.
- The terminal nodes are called **leaves** or **operational taxonomic units** (**OTUs**) and represent sequences or species for which data was obtained. They usually represent living species.
- **Internal nodes** represent hypothetical ancestral species before a divergence event.
- The ancestor of all the sequences or species in a given tree is called the **root**. Not all trees have a root.



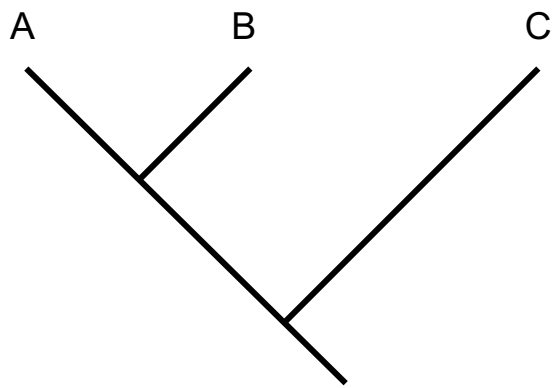
Phylogenetic Tree Terminology

- A **clade** or **monophyletic group** is a group of species or sequences that includes the most recent common ancestor of all of its members and all of the descendants from the most recent common ancestor.
- **Sister taxa** are species or clades arising from the same node.
- Sequences A and B are sister taxa. Sequences C and D are sister taxa.
- Clade 1 and clade 2 are sister taxa.



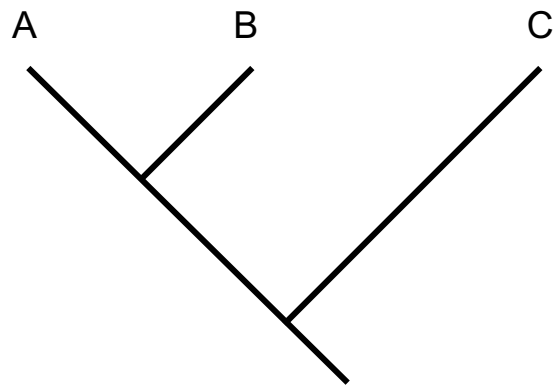
Tree Types: Cladogram

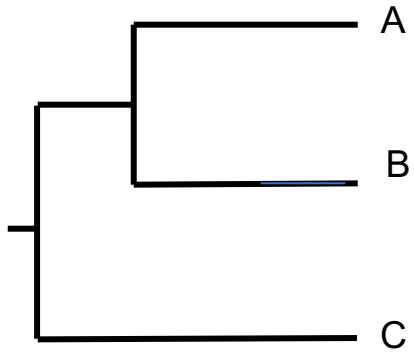
- Only shows the branching order
- **The branch lengths don't have meaning**.
- In this example, a tree for three sequences (A, B, and C) is shown.
- This tree indicates that A and B share a common ancestor more recently than either does with C.



Tree Types: Cladogram

- Only shows the branching order
- **The branch lengths don't have meaning**.
- In this example, a tree for three sequences (A, B, and C) is shown.
- This tree indicates that A and B share a common ancestor more recently than either does with C.

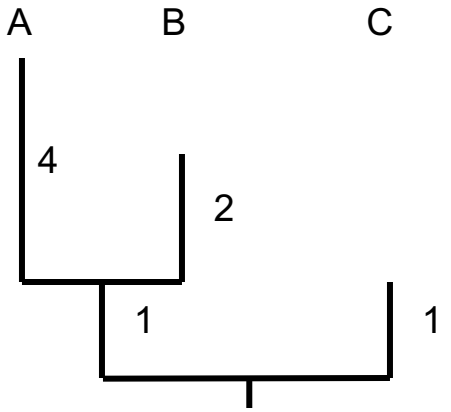




Different shapes can be used to show the same thing

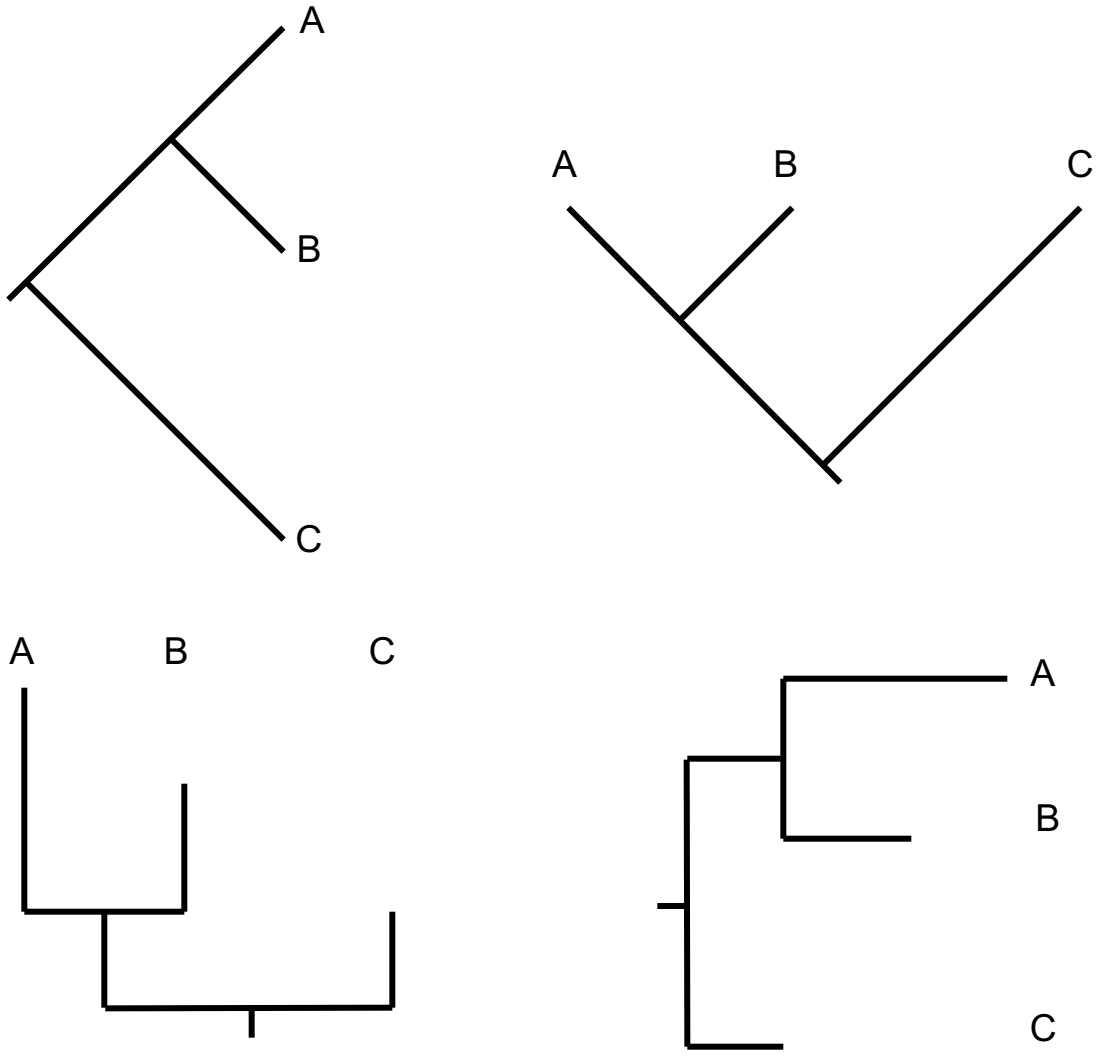
Tree Types: Phylogram

- Shows the branching order **and has scaled branches** to indicate some attribute of similarity, such as the number of sequence changes.
- This tree indicates that A has acquired more substitutions than B in the time since they shared a common ancestor.
- Branch lengths may be indicated using a number, or by adding a scale bar.
- When drawn vertically, the distance between any two nodes is the sum of the vertical branch lengths between them.
- When drawn horizontally, the distance between two nodes is the sum of the horizontal branches



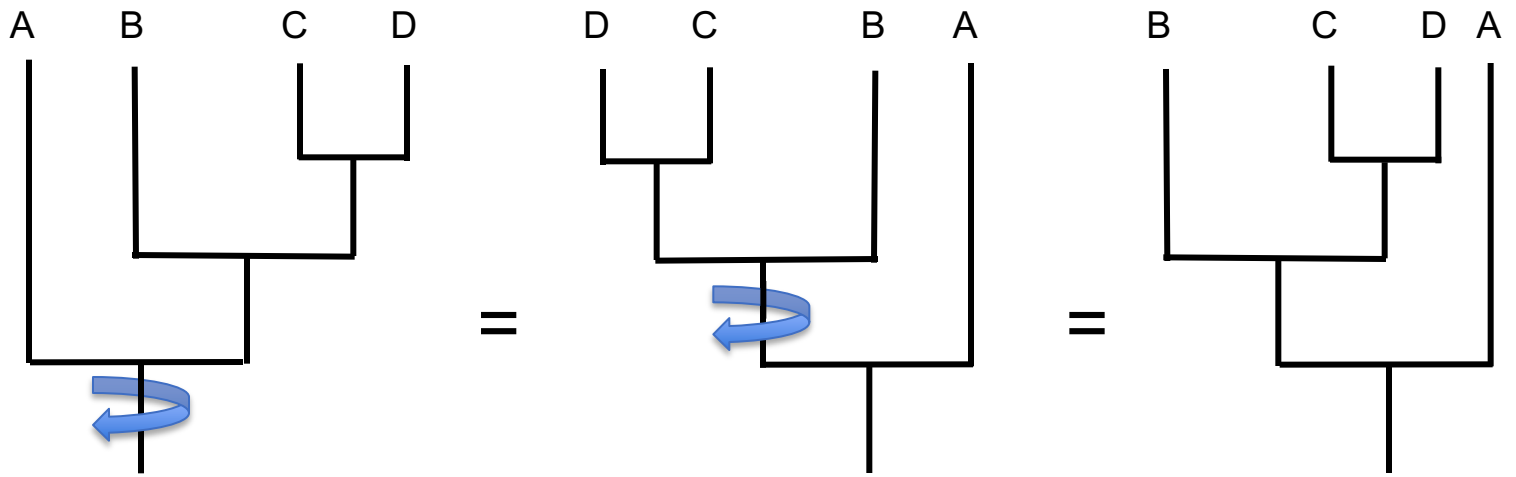
- The distance between A and B is 6
- Horizontal branch lengths not meaningful (just used to separate out taxa).

Tree Orientation



Trees can be drawn vertically or horizontally

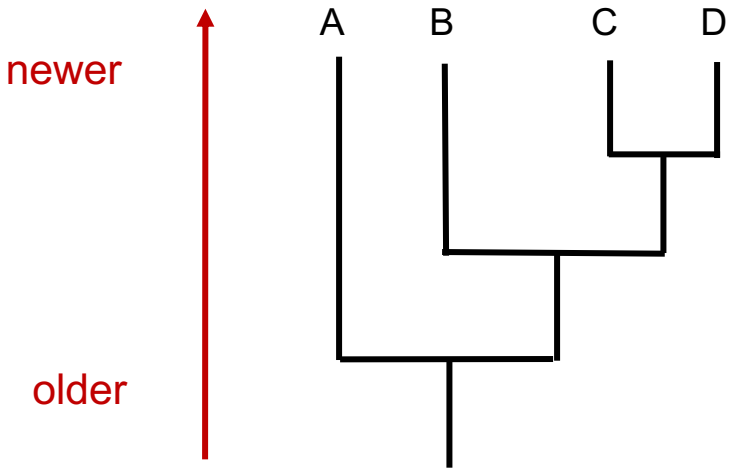
Order of Leaves



- Rotating the branches of a tree doesn't change the relationships depicted by the tree.
- The relatedness is depicted by the tree topology

Rooted Trees

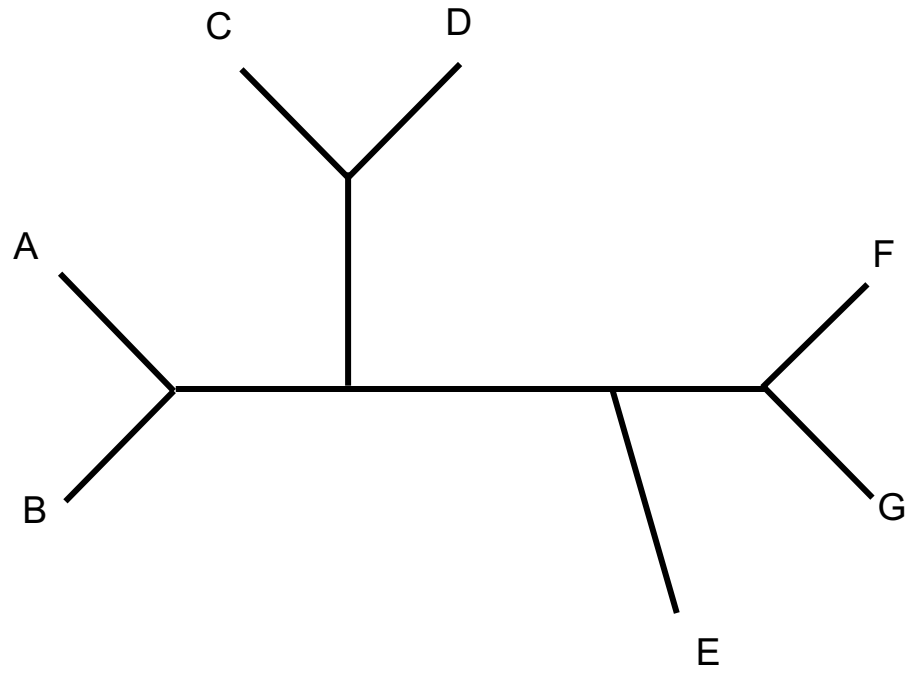
- The root is the ancestor of all the sequences in the tree.
- A tree with a root shows the order of descent (i.e. the direction of evolution).



rooted tree

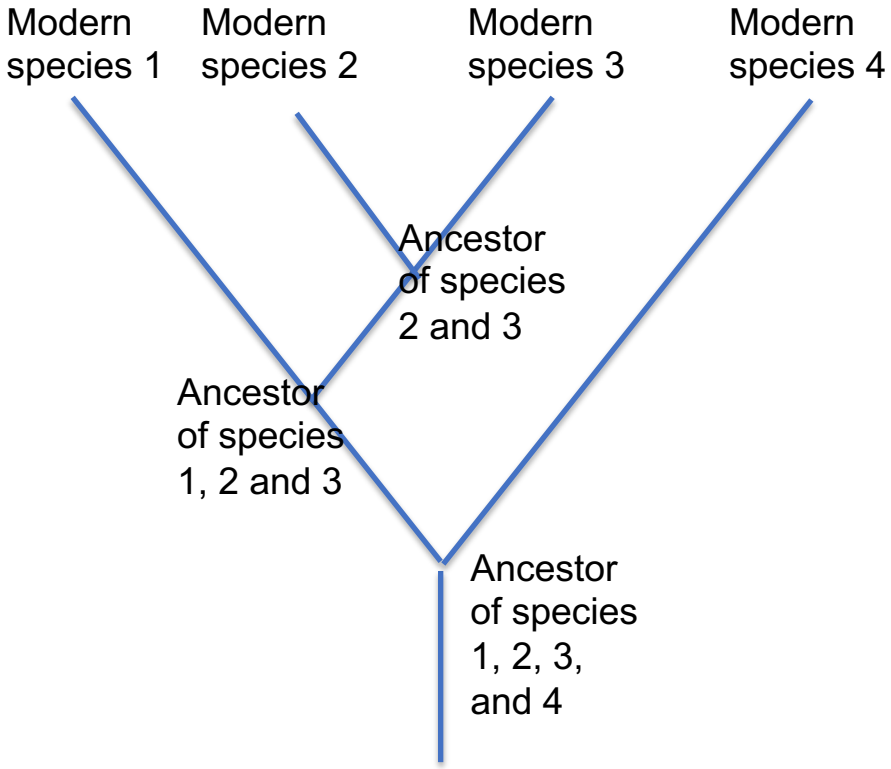
Unrooted Trees

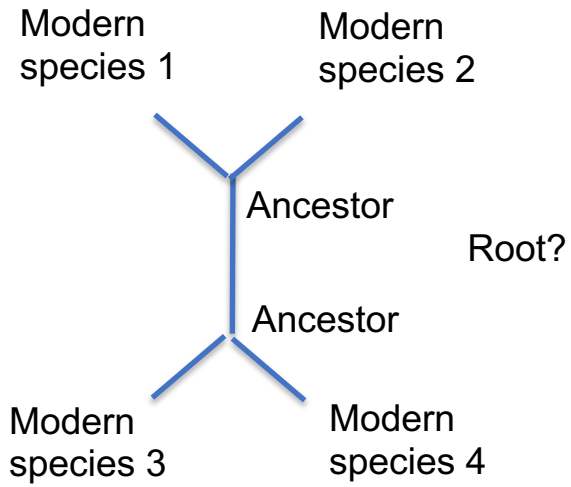
- An unrooted tree lacks a root.
- An unrooted tree doesn't show the direction of evolution.
- Less informative than a rooted tree.
- Often drawn in **radial** format.



unrooted tree
drawn in **radial** format

Rooted vs Unrooted



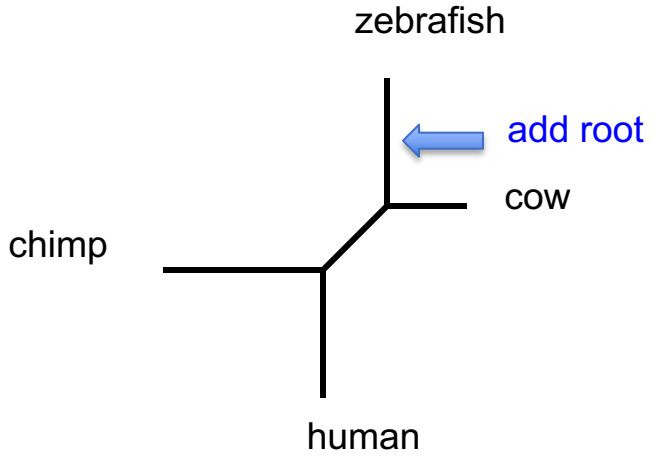


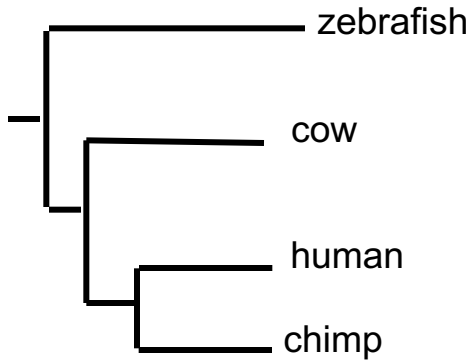
Only a rooted tree shows which ancestors led to which species

Rooting a Tree

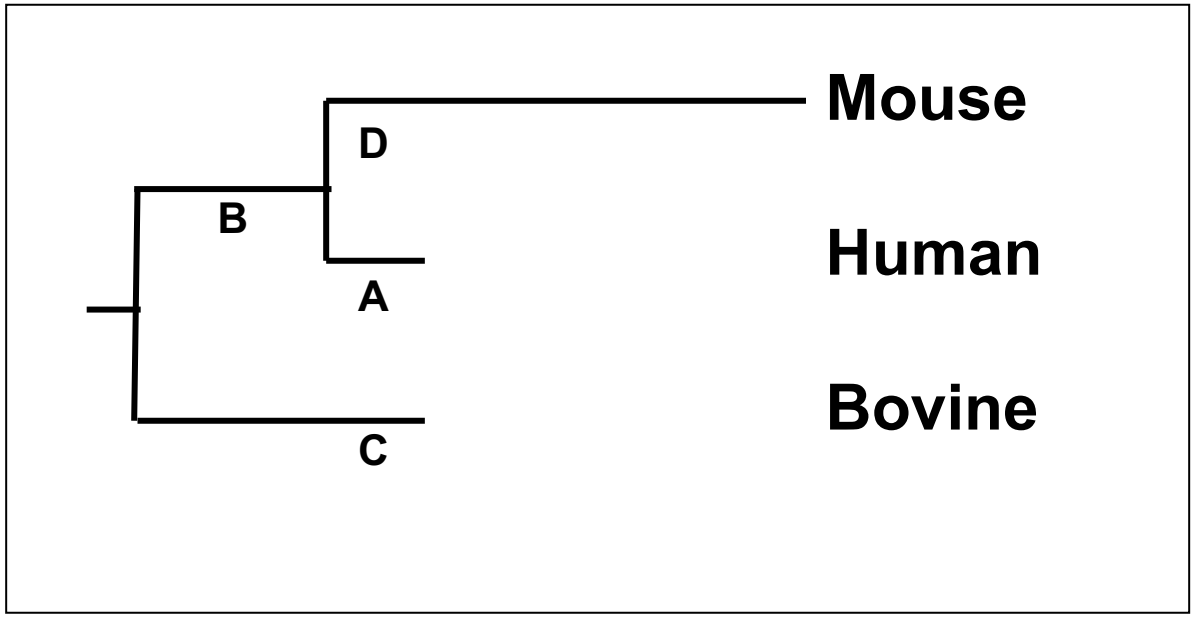
- It is often possible to place the root by including an **outgroup** (a sequence or species that is thought to be ancestral, or more distantly related to the **ingroup** members than they are to each other).

- For example, to root a tree of mammalian sequences, you might use a sequence from a non-mammalian vertebrate as an outgroup.



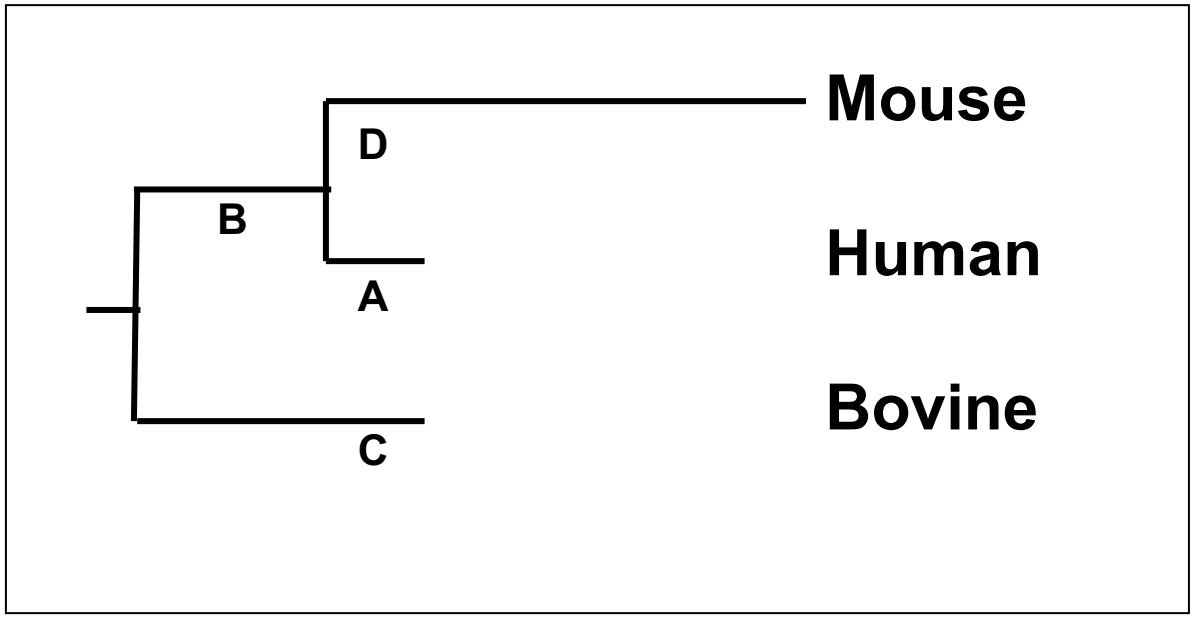


Example: What does this tree tell you?



Example: What does this tree tell you?

- Phylogram (branch lengths meaningful)
- Appears to be rooted
  (but need to check methods to confirm!)
- Mouse lineage has undergone some accelerated evolution
- Human and Mouse share a common ancestor
- …however due to the mouse accelerated evolution, human is actually has more sequence similarity with bovine vs mouse!

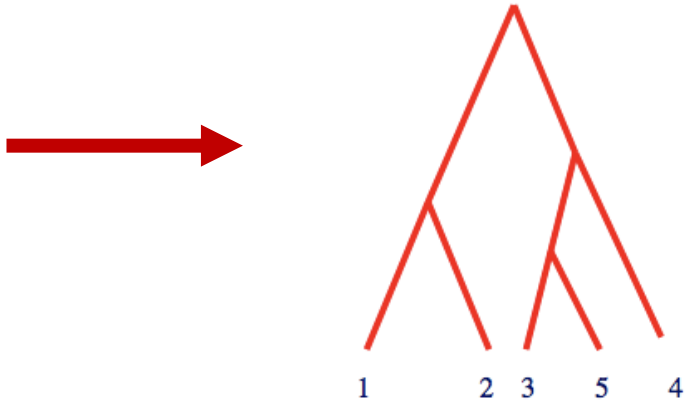


Fun fact: We DO share more sequence similarity with cows vs mice, on average, even though we have more recent shared ancestry with mice!

(Note: The tree is drawn though with exaggerated branch lengths to illustrate the point)

So, how do we build a tree?

TTATTAA…
AATTTAA…
AAAAATA…
AAAAAAT…
TATATAT…



Building a Tree

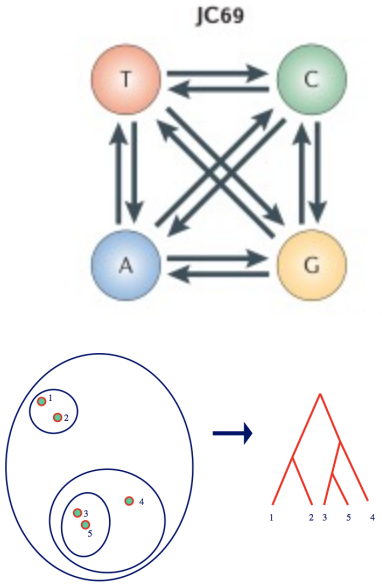
Building a phylogenetic tree from sequence data requires:

1. **A multiple sequence alignment (MSA)**
2. A model of evolution
3. A tree building algorithm

Objective: identify a tree that best describe the data, given the model of evolution used

```
A   TTATTAA
B   AATTTAA
C   AAAAATA
D   AAAAAAT
```

Each aligned column is homologus




```
bioinformatics
bioinformatics
bioinformatios
--oinformatios
---informatios
---information
---information
```

time

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSSSGFIFSS--YAMYWVRQAPG
LSLTCTVSGTSFDD--YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCLISDFYPGA--VTVAWKADS--
AALGCLVKDYFPEP--VTVSWNSG---
VSLTCLVKGFYPSD--IAVEWESNG--
```

Remember: The sole purpose of multiple sequence alignments is to place *homologous (shared ancestry) positions* of *homologous sequences* into the *same column*.

Tree Building

Four steps:

1) Construct a multiple sequence alignment
   - *from good quality sequences ideally labelled with relevant contextual data*
2) Determine an evolution (substitution) model to use
3) Build the tree
4) Evaluate the tree

Contextual data and sequence quality discussed more in future modules

There are 2 main tree building methods

1. Distance-based: transform the sequence data into pairwise distances

2. Character-based: use the aligned sequences directly during tree building

Distance Methods

- Creates a tree based on a distance matrix built using a multiple sequence alignment.
- The simplest measure of distance between two sequences is to count the number of sites at which they differ.
- Common distance methods are unweighted pair-group method with arithmetic mean (UPGMA) and neighbor joining (NJ).

sequences

```
A    TTATTAA
B    AATTTAA
C    AAAAATA
D    AAAAAAT
```

distances

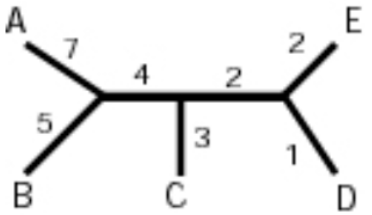
	A	B	C	D
A	0			
B	3	0		
C	5	4	0	
D	5	4	2	0

Distance Methods

	A	B	C	D	E
A	0				
B	12	0			
C	14	12	0		
D	14	12	6	0	
E	15	13	7	3	0

Easy

Not Easy



- Input is an n x n matrix M, where M_{ij} is the distance between sequences i and j.
- The goal is to build a tree where each leaf corresponds to a sequence in M, where the distances measured from the tree between leafs i and j (d_{ij}) correspond to M_{ij}.
- *A tree that exactly fits the matrix often doesn't exist* (the distances must satisfy several mathematical properties if a perfectly matching tree is to exist).
- *We want to find the tree that most closely matches the matrix*.

Finding the tree with the "best-fit" turns out to be an NP-complete problem.

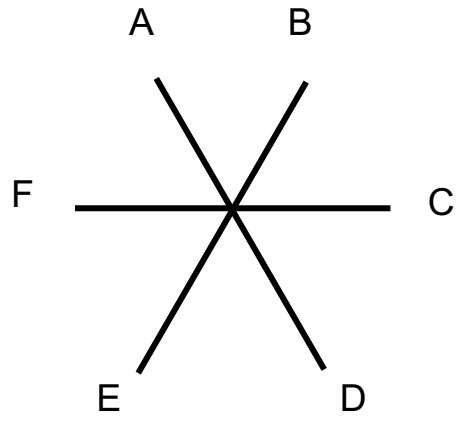
Need heuristic (approximate) approaches.

Neighbor Joining (NJ)

Heuristic method which joins at each step the two closest sub-trees that are not already joined.

Distance M

	A	B	C	D	E	F
A	0					
B	5	0				
C	4	7	0			
D	7	10	7	0		
E	6	9	6	5	0	
F	8	11	8	9	8	0

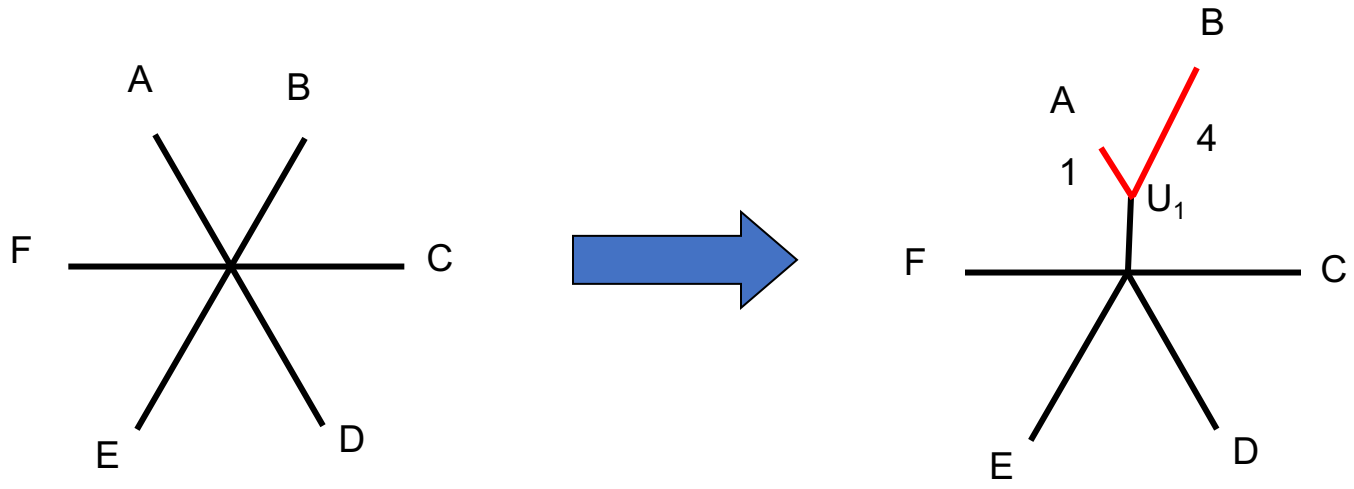


Q Matrix

	A	B	C	D	E
B	-13				
C	-11.5	-11.5			
D	-10	-10	-10.5		
E	-10	-10	-10.5	-13	
F	-10.5	-10.5	-11	-11.5	-11.5

- Start out with a distance matrix and a completely unresolved tree.
- A ‘Q’ matrix is calculated from the original distance matrix to determine the average distance from each node to all the other nodes.

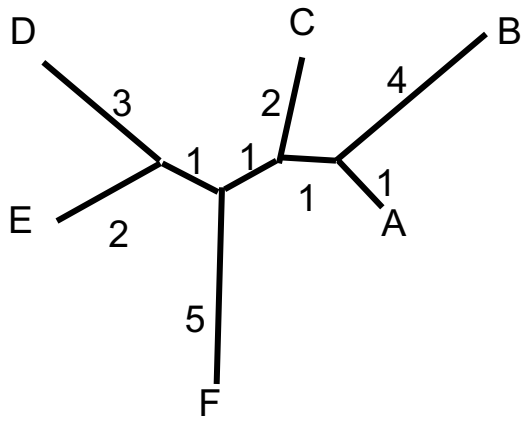
NJ example



- Based on these distances between the new node and the two sequences (A and B), the nodes can be reconnected by a new node (U_1). This is referred to as 'star decomposition'.
- The process is repeated until the tree is fully resolved.

NJ Result

- NJ produces an unrooted tree with branch lengths.
- Tree can be rooted if one of the sequences is known to be an outgroup.
- NJ won't necessarily find the optimal tree.
- However, extensive testing has shown that this method works quite well, and it is relatively fast.



Distance Methods – Watch out

- Distance methods throw some information away
- Many distinct data sets can yield the same distance measures.

```
WNPFKELERAGQRVRDAVISAAPAVATVGQAAAIARG
WNPFKELERAGQRVRDAV-----AVATVGQAAAIARG
WNPFKELERAGQRVRDAVISAA----------------
```

i.e. Gaps are not incorporated into distance-matrices.
So, what portion of the above multiple sequence alignment would be used by a distance-based method?

Distance Methods – Watch out

- Distance methods throw some information away
- Many distinct data sets can yield the same distance measures.

```
WNPFKELERAGQRVRDAVISAAPAVATVGQAAAIARG
WNPFKELERAGQRVRDAV-----AVATVGQAAAIARG
WNPFKELERAGQRVRDAVISAA----------------
```

i.e. Gaps are not incorporated into distance-matrices.
So, what portion of the above multiple sequence alignment would be used by a distance-based method?

- In contrast, character-based methods operate directly on the aligned sequences and are generally regarded as giving more accurate trees.

Character Based Methods

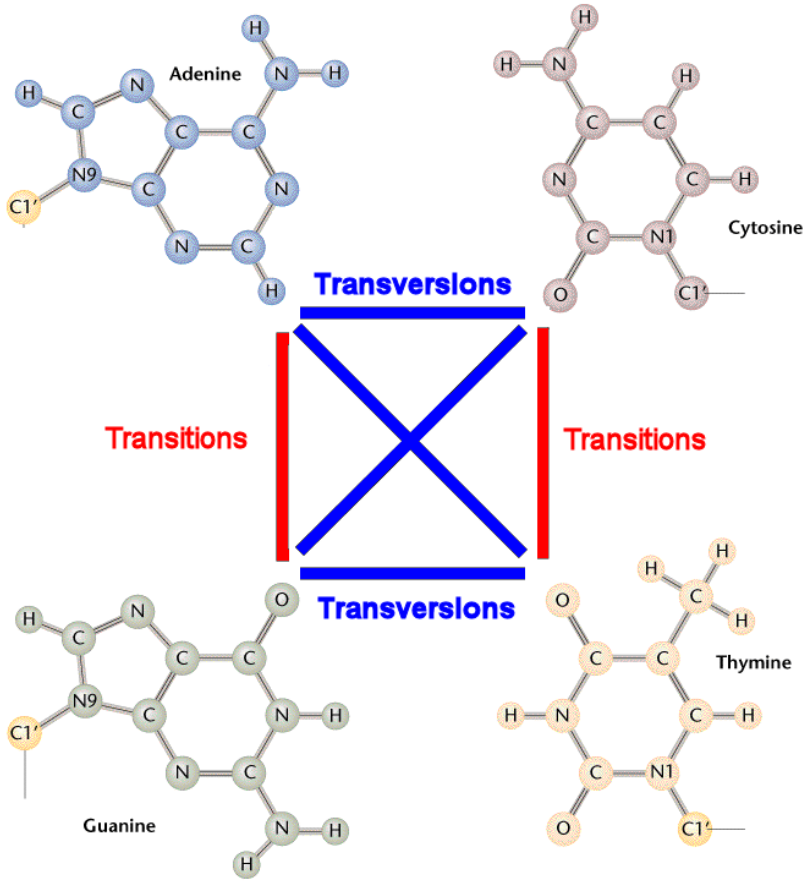
- Character methods (also called discrete methods) operate on sequences rather than on the distances.
- The two major character methods are maximum parsimony (MP) and maximum likelihood (ML).
- MP involves finding the tree that describes the sequences using the fewest evolutionary steps.
- ML involves finding the tree that is most likely to have produced the data, given a probabilistic model of sequence evolution.

Maximum Likelihood Trees

- Maximum likelihood (ML) tries to find the tree that maximizes the probability of observing the data (i.e. the multiple sequence alignment).
- Requires a **model of sequence evolution**, a **tree**, and the **observed data**.
- In other words, given data D and a model M, find the tree T, such that Pr(D|T,M) is maximized.

Evolutionary Model: Transitions and Transversions

- Transitions are the interchange of two **purines** (A←→G) or two **pyrimidines**. (C←→T).
- Transversions are the interchange of a purine with a pyrimidine
- There are twice as many possible transversions, but they are not favoured, transitions are generated at a higher frequency than transversions.
- Transitions are less likely to result in amino acid substitutions.

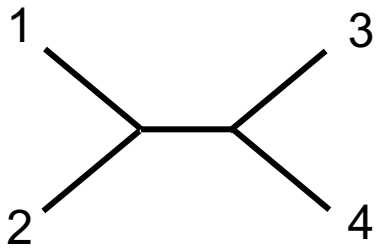


Maximum Likelihood

Consider the following four sequences and the tree ((1,2),(3,4)):

```
1 ATATT
2 ATCGT
3 GCAGT
4 GCCGT
```

((1,2),(3,4))

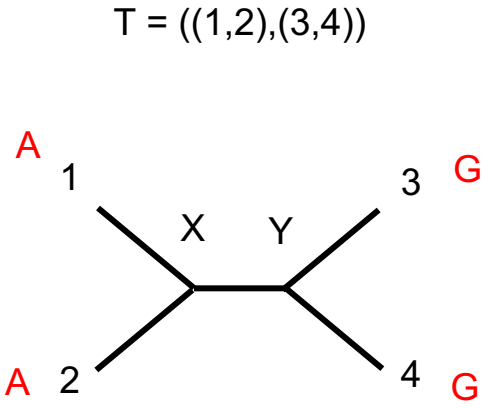


What is the probability of D_1 (i.e position 1), given this tree, and a model of evolution that says the P(transition) = 0.3, P(transversion) = 0.1, and P(no change) = 0.6?

Maximum Likelihood

What is the probability of D_1 (i.e. position 1), given this tree, and some model of evolution?

```
1 ATATT
2 ATCGT
3 GCAGT
4 GCCGT
```



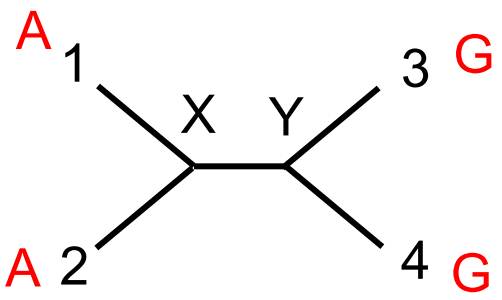
We calculate the probability of the tree for every possible reconstruction of the ancestral sites X and Y. In other words we set X and Y to G, A, T, or C. There are 16 values to calculate (4 x 4).

Maximum Likelihood

What is the probability of D_1 (i.e position 1), given this tree, and some model of evolution?

```
1 ATATT
2 ATCGT
3 GCAGT
4 GCCGT
```

T = ((1,2),(3,4))



With the 16 values calculated we can determine the probability of the column given the tree and the model.

$$Pr(D_1|T,M) = P_{case1} + P_{case2} + ... + P_{case16}$$

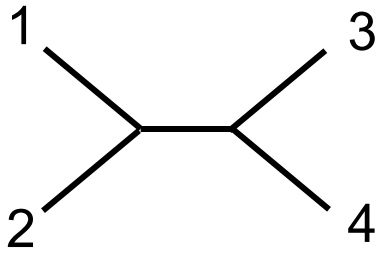
Maximum Likelihood

Other positions, $Pr(D_2|T,M)$ to $Pr(D_5|T,M)$, are calculated in a similar manner. Once we have these we can calculate Pr(D|T,M) as:

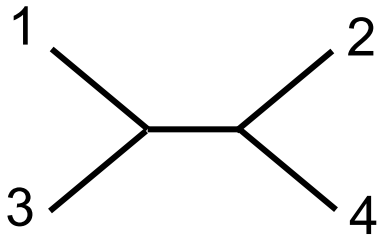
$$Pr(D|T,M) = Pr(D_1|T,M) * Pr(D_2|T,M) * ... * Pr(D_5|T,M)$$

This is the likelihood of one particular tree--need to look at all three to find the one that gives the largest value for Pr(D|T,M).

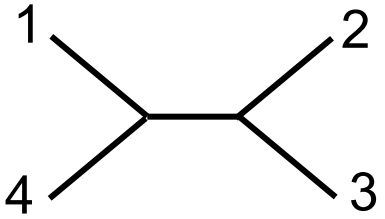
((1,2),(3,4))



((1,3),(2,4))



((1,4),(2,3))



Maximum Likelihood

- Requires searching through many possible trees
- Very computationally intensive.
- The evolutionary model can also include time--this means considering each topology and best branch lengths for each topology.
- The result of ML is dependent on the model of evolution used.

Other Approaches

Bayesian Approaches:

- Applies Bayes theorem to estimate a probability distribution for population parameters of interest.
- Has the ability to incorporate prior information for events (i.e. a prior distribution for outbreak onset time).

Recombination-aware approaches:

- Recombination invalidates most approaches, since the columns of your MSA should be homologous. *Its important to detect possible recombination before performing phylogenetic analysis.*

Ultrafast Sample placement on Existing tRees (UShER):

- Enables rapid SARS-CoV-2 analysis – placing a sample on a very large tree.

Will cover more (Bayesian approaches, time trees) in the phylodynamics course module

So, what is the best tree building approach?

- No single method is the best for all circumstances!
- Depends on
  - the size and complexity of the dataset
  - the speed of your computing resources
  - why you want to perform the analysis

Identifying a best model? Try a couple and see how well they fit the data. jMODELTEST carries out statistical selection of different evolutionary models of nucleotide substitution. http://darwin.uvigo.es/software/jmodeltest.html

Tree Evaluation: Bootstrapping

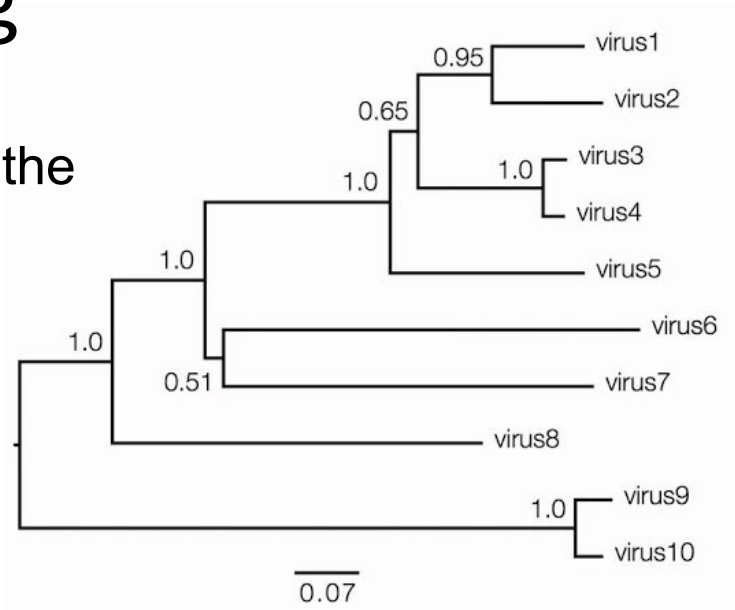
- Bootstrapping tests whether the whole dataset is supporting the tree, or if the tree is just a slight winner among nearly equal alternatives.

- Involves generating many new data sets (usually 100 to 1000) that are *slightly perturbed* from the original data set (such as random sampling of some columns and replacement, so some columns are represented more and others are removed). These are called bootstrap replicates.

- For each replicate a tree is built using the same method as the original tree.

- The original tree is then labeled at the node of a cluster with numbers indicating how often each cluster occurs in the trees made from the replicates (or a consensus tree is built and labeled).

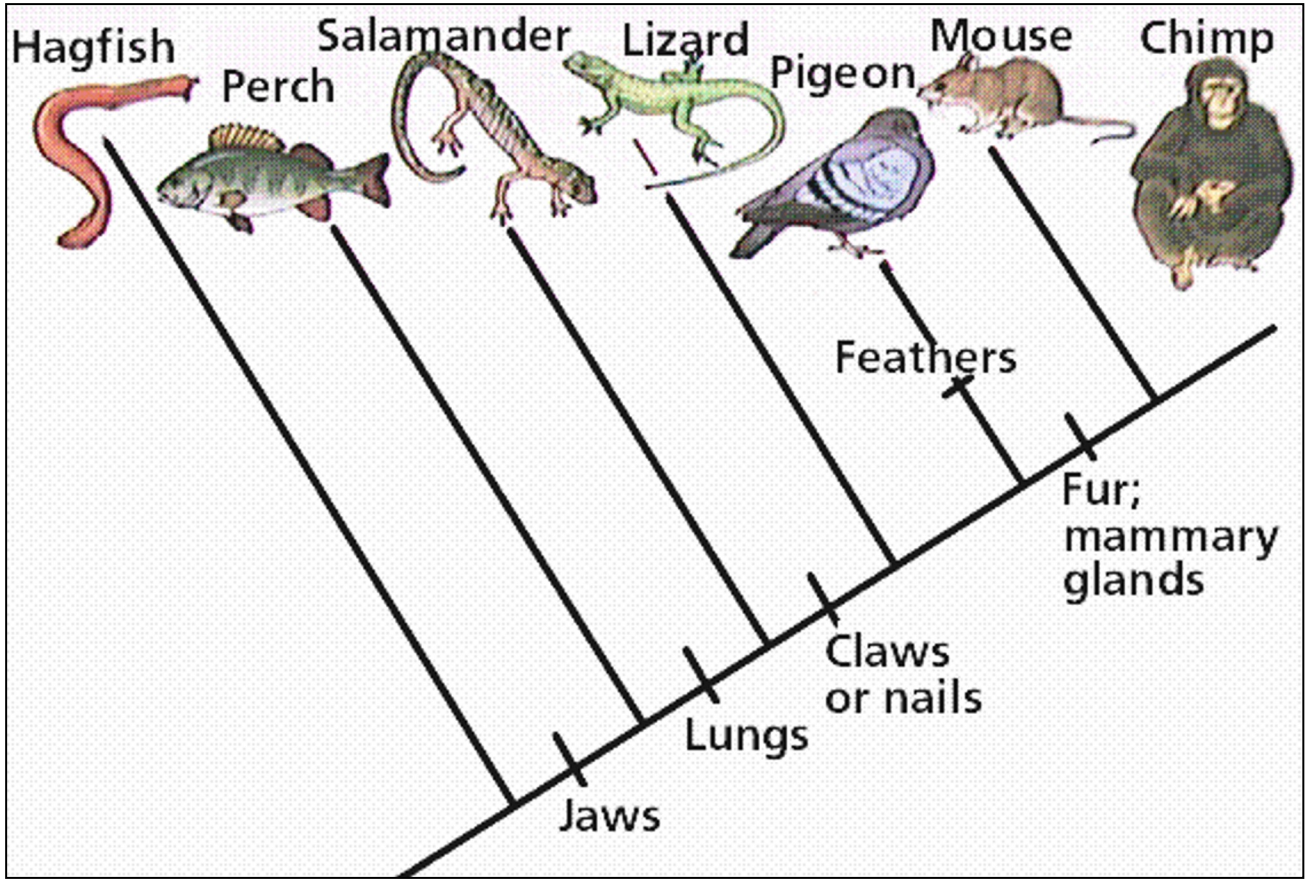
- A bootstrap value of > 70% is considered good support for a given cluster.



High bootstrap values provide support, *as long the multiple sequence alignment is robust.*

Also useful: Identify unique characters vs homoplasy

- Unique characters occurred once and are unreversed
- Examples: Fur, feathers, certain nucleotide changes
- Very useful as support for inferring relationships



Genomic Epidemiology Example:

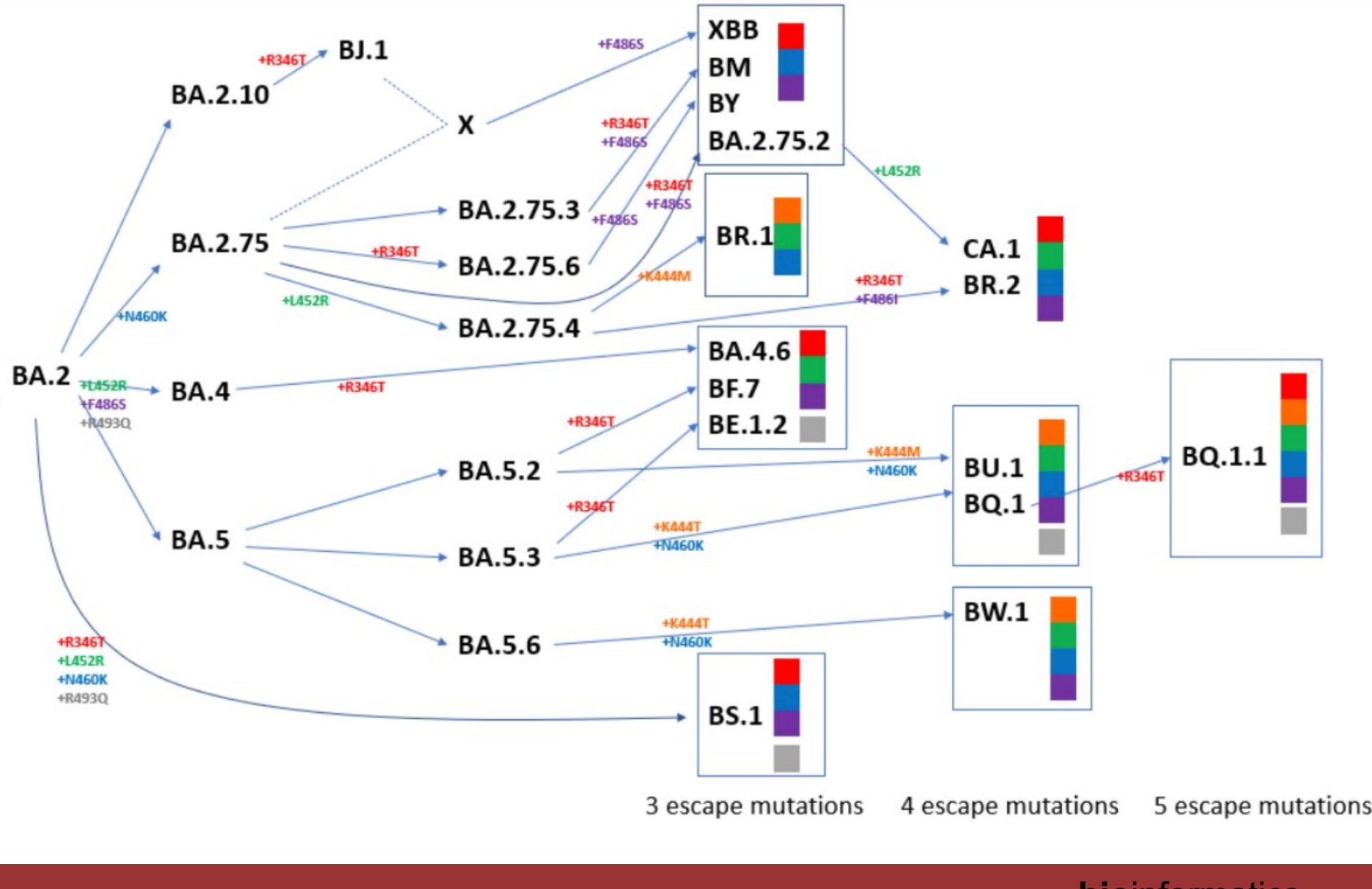
You find an indel (insertion/deletion character) only in sequences associated with an outbreak in a certain geographic location. Phylogeny suggests a new sequence of interest elsewhere may have originated from that location. The new sequence contains the unique indel, further supporting the phylogeny.

Also useful: Identify convergent evolution

Example:

"The Great Convergence" for SARS-CoV-2 viral variants.

Immune evasive mutations (coloured) have been occurring independently in different viral variants, again and again. Evidence of functional benefit



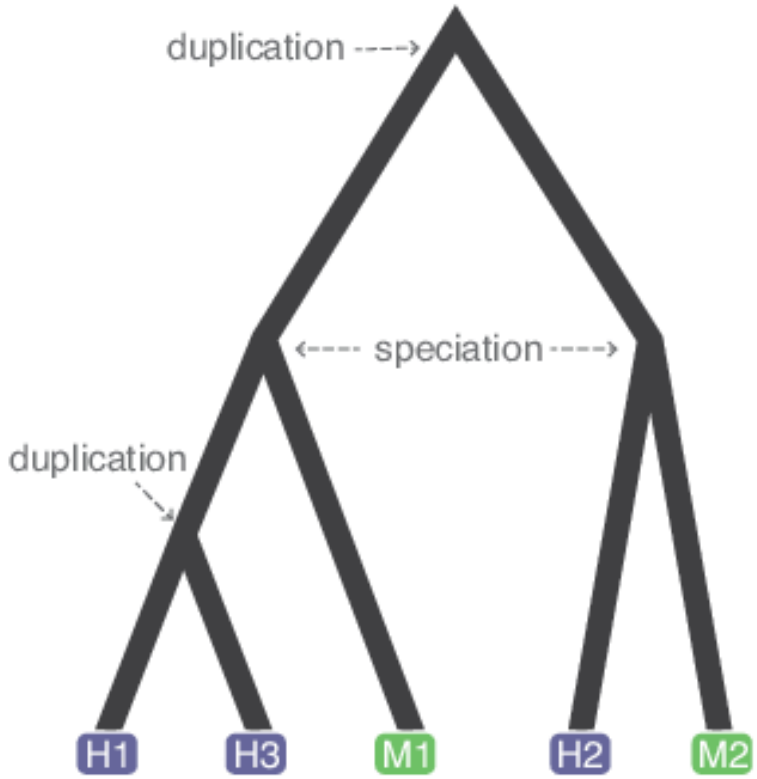
Also useful: Identify orthologs, paralogs, xenologs

Homologous or not?: *Often* determined by arbitrary threshold level of similarity determined by alignment

- **orthologs** - Homologs produced by speciation.
  They *tend* to have similar function.
  Common way to Infer: The gene tree matches the species tree. Rough way to Infer: Reciprocal best BLAST analysis

- **paralogs** - Homologs produced by gene duplication.
  They *tend* to have differing functions.
  Common way to Infer: Multiple copies in same species

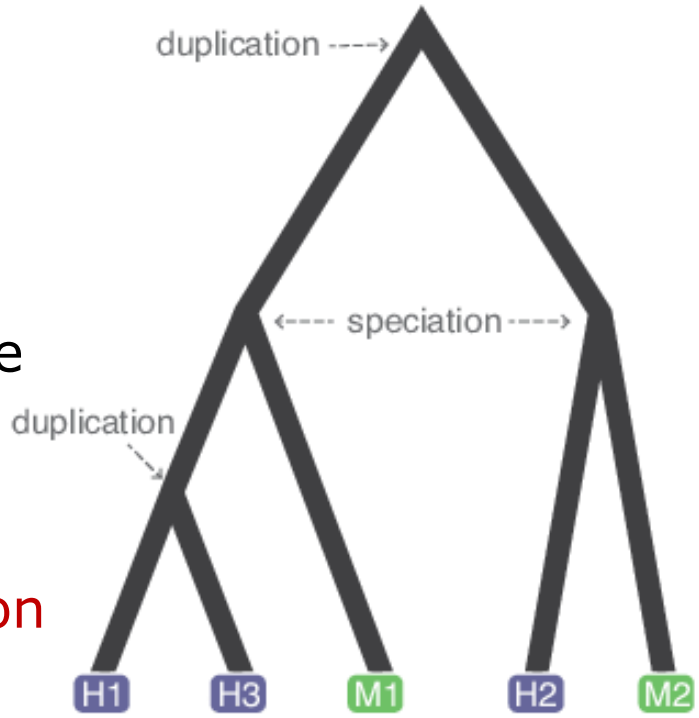
- **xenologs** -- Homologs resulting from horizontal gene transfer between two organisms.
  They *tend* to impart novel adaptive functionality to the recipient organism.
  Common way to Infer: A gene doesn't match the species tree where the rest of the gene tree does match.
  Or you ID a novel genomic insertion (phage, genomic island)

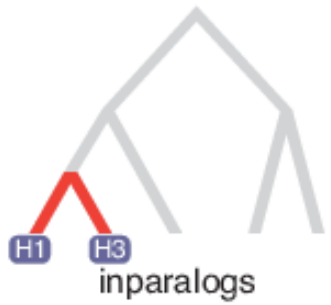


Also useful: Identify orthologs, paralogs, xenologs

We can refer to paralogs in the context of a species relationship:
In-paralogs:
Duplication after species divergence
Out-paralogs:
Duplication before species divergence

Useful for inferring function:
In this tree (right) M2 and H2 are likely functionally similar. The function of H1 and H3 vs M1 is less clear.





Out-paralog example: A large paralogous family of bacterial transporter genes that evolved before divergence of many bacterial species.
Note: Selective gene loss in different species can complicate analysis.
In-paralog example: A recent duplication of some *Neisseria* opacity genes in a lineage, as part of its evolving immune evasion.

Closing comments

- Phylogenetic inference allows us to *estimate/infer* the evolutionary relationships between genes and/or species
- Data quality is paramount
- We covered the basic terminology, methods, interpretation
- More complicated and efficient methods are needed to deal with genomic size data
- We will look at some of the techniques and approaches in the rest of the workshop.

Note that phylogeny is best interpreted when you have well organized contextual data (geography, patient information)… More about that soon!

But first: AWS Setup

Workshop Sponsors: