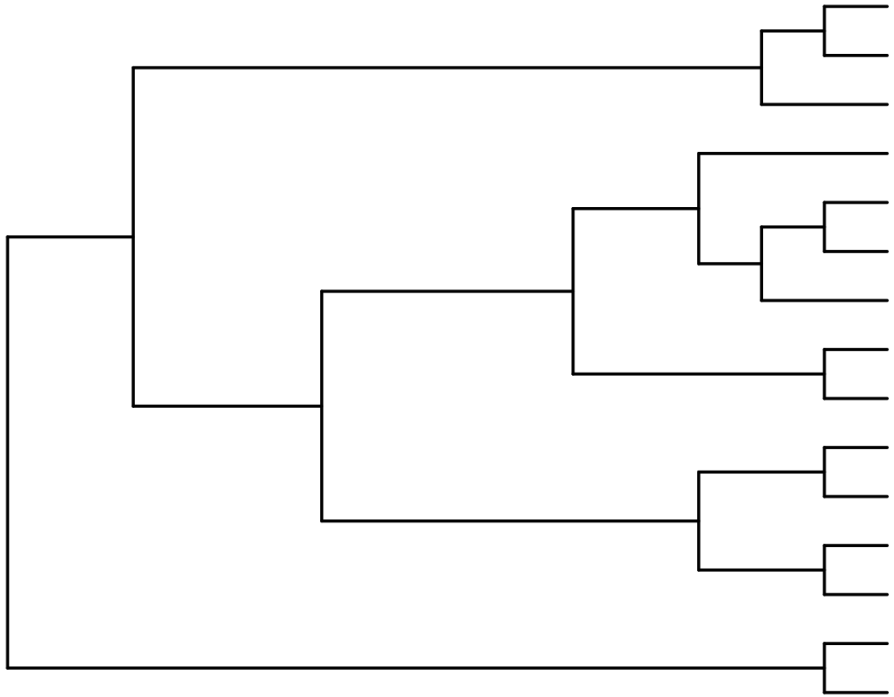
Basics of phylogenetics

Dom Bennett (03/2015)

Objectives

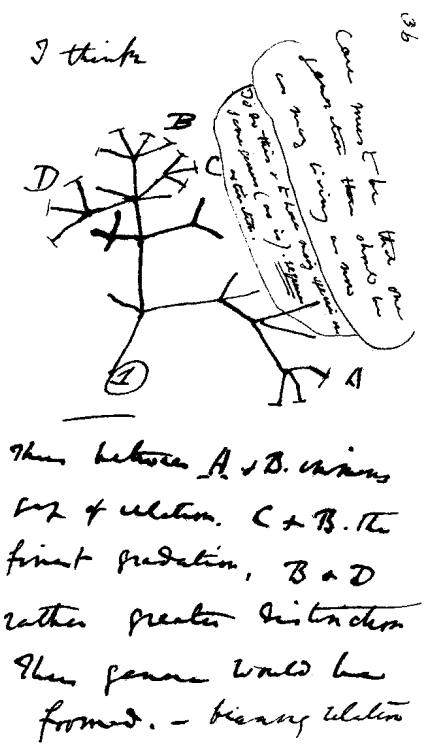
- *Know* what a phylogenetic tree is
- *Know* phylogenetic terminology
- *Understand* conceptually how to build a phylogenetic tree
- *Understand* the steps to create a molecular phylogeny

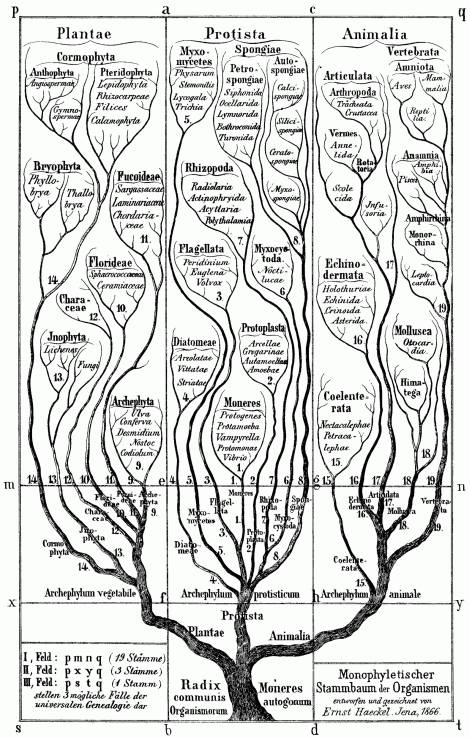
What is a phylogenetic tree?



- The diagrammatic representation of the relationships between taxa
  - not the same as a distance matrix!
- Etymology:
  - *phylos* (Greek race) + *geny* (Greek origin)
- Related terms exist:
  - cladogram
  - dendrogram
  - chronogram
  - hierarchical cluster

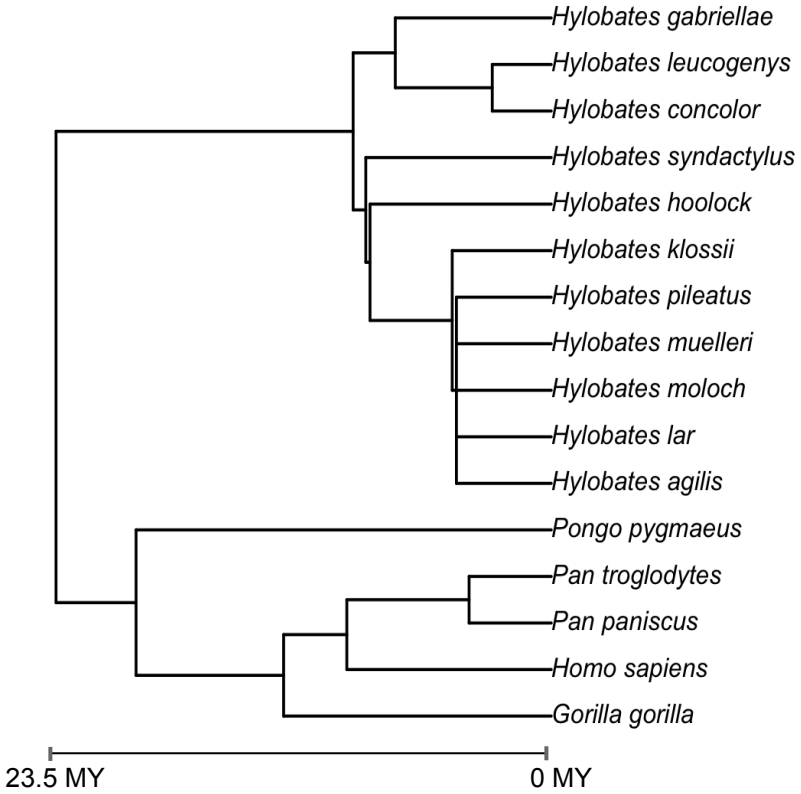
Phylogenetics: a brief history





- ‘First started’ with Darwin as a conceptual sketch
- Ernst Haeckel was the first to formalise the process with the ‘theory of recapitulation’
- 1950s+, the process switched to morphological matrices
- 1990s+, the process moved into using molecular methods

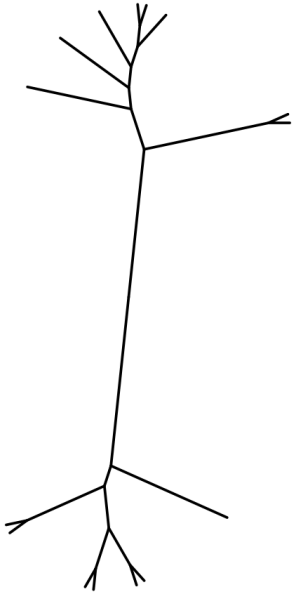
Comparison with taxonomy



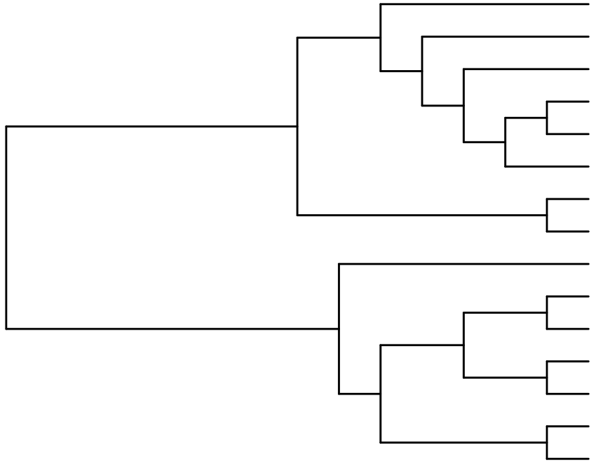
- Phylogeny is NOT the same as taxonomy
- Taxonomy is the arbitrary assortment of organisms into groups
- Phylogeny the objective grouping, often time-calibrated.

~30 MYA

Terminology: rooted

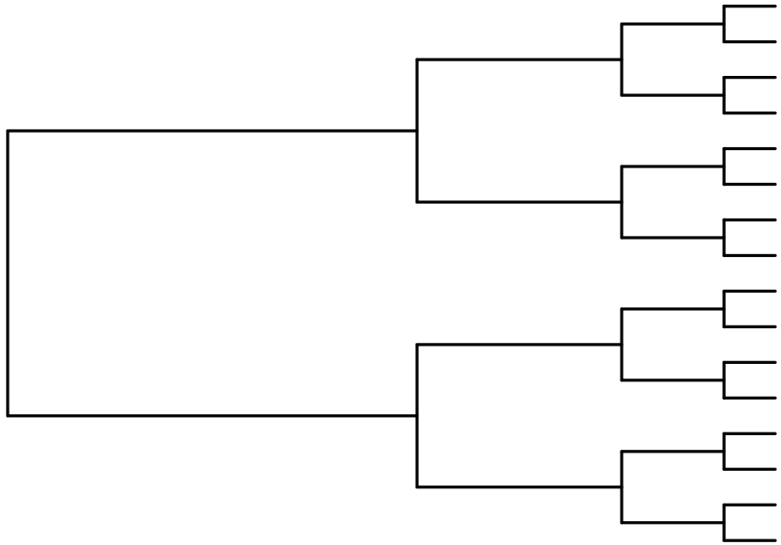


Unrooted

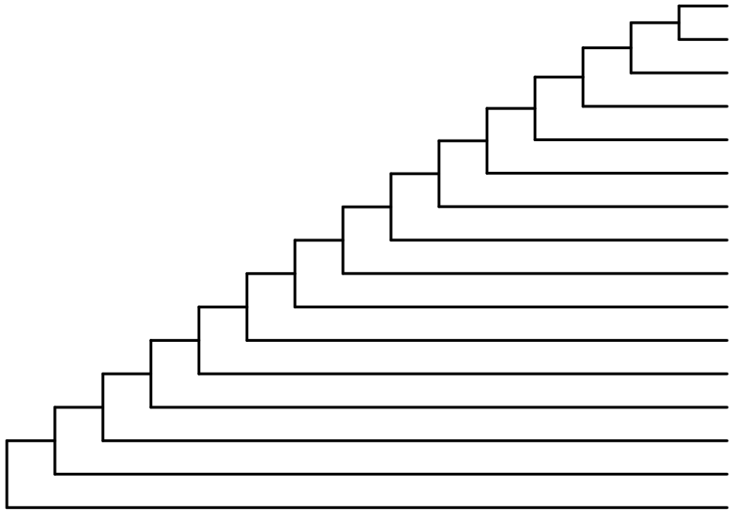


Rooted

Terminology: balance

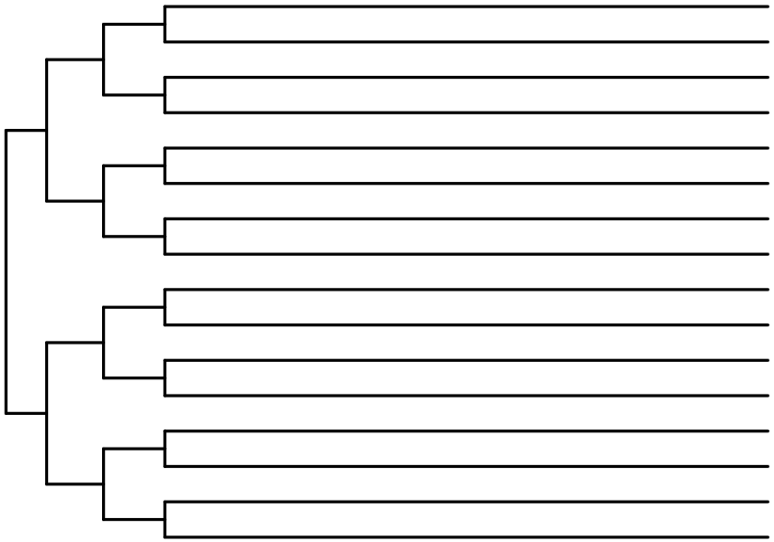


Balanced

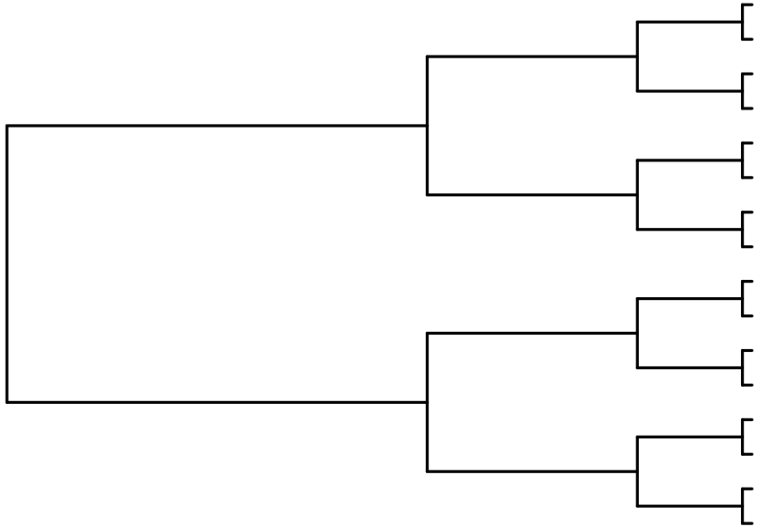


Unbalanced

Terminology: loading



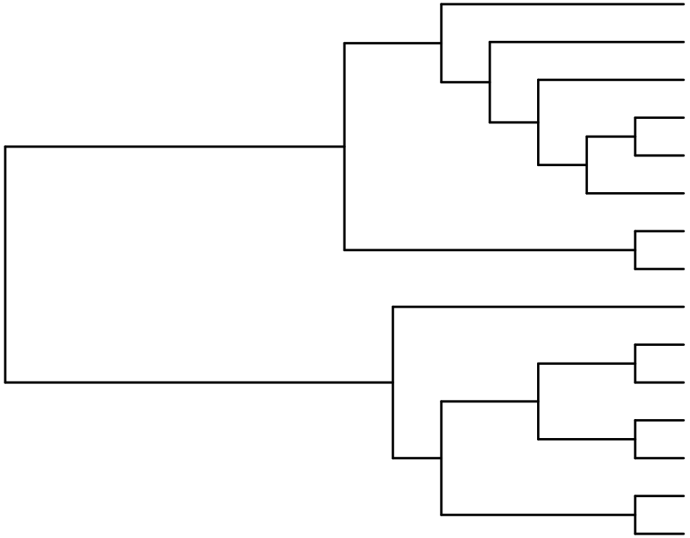
High loading/Early burst



Low loading/Late burst

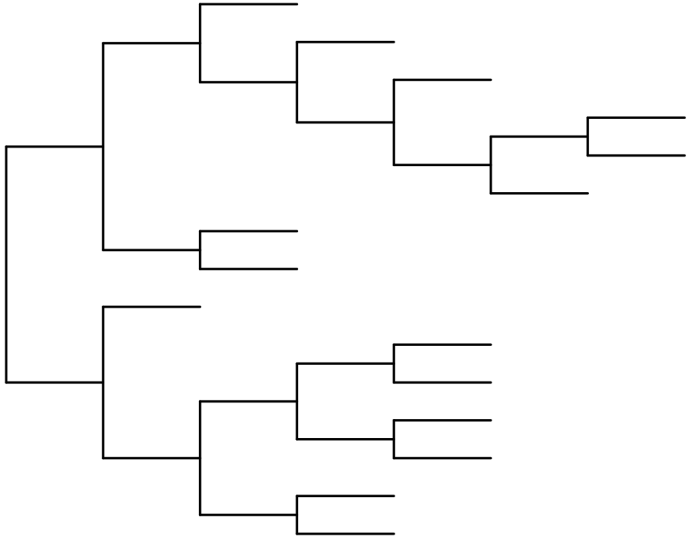
Terminology: binariness

Multifurcating/polytomous

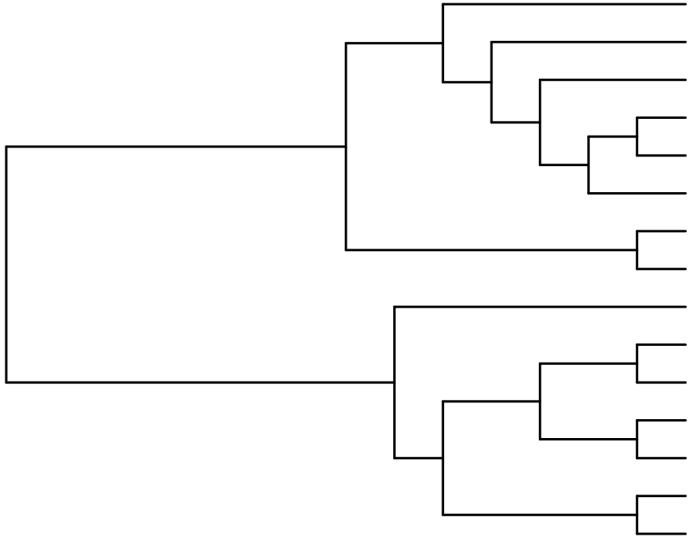


Bifurcating/dichotomous

Terminology: ultrametrecy

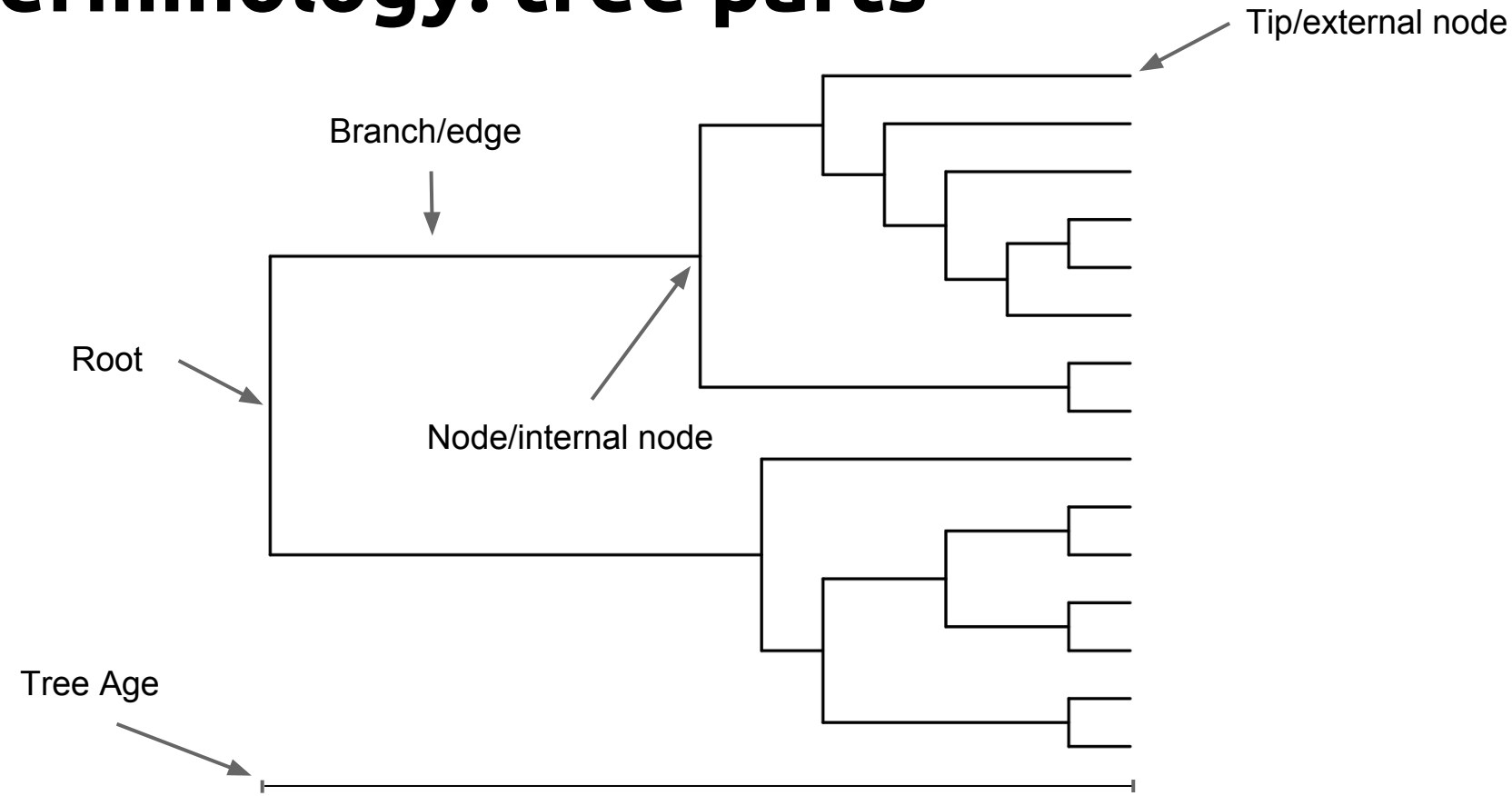


Non-ultrametric

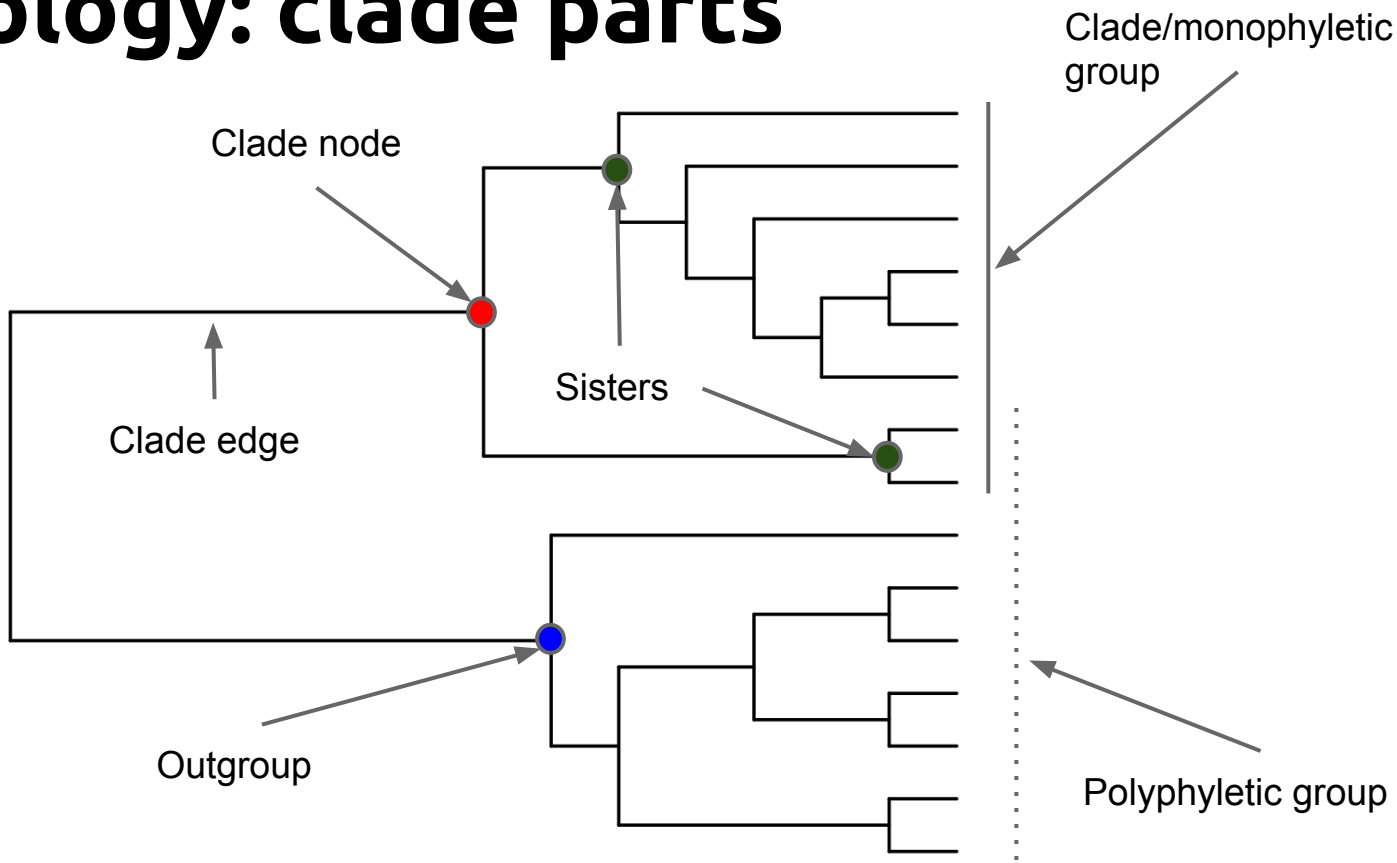


Ultrametric

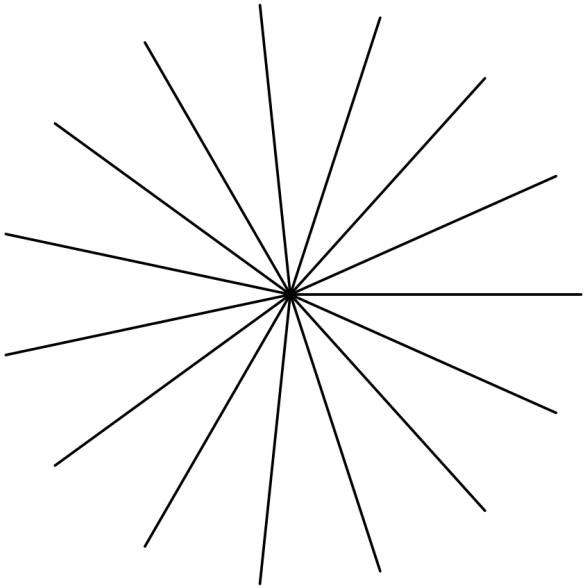
Terminology: tree parts



Terminology: clade parts



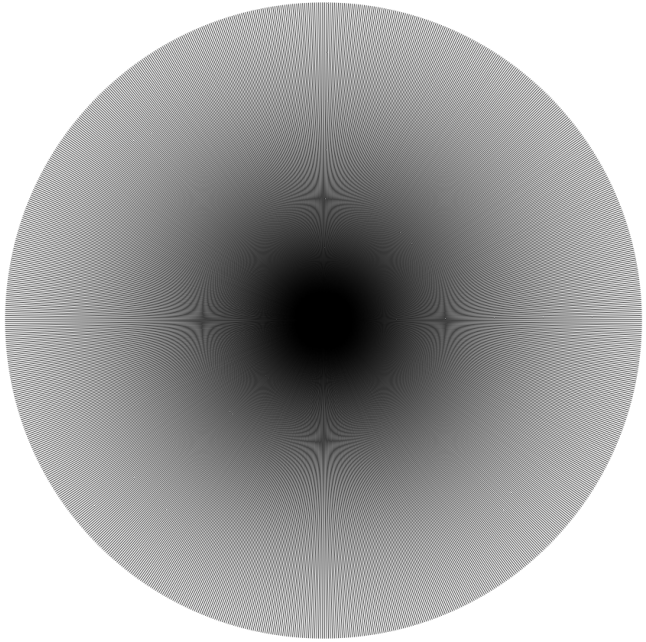
Terminology: poor trees



Star phylogeny: all species are equally distant/related

Long branch attraction: one species is very distant from others

Terminology: really poor trees



Black hole (+1000 species
equally distant/related)

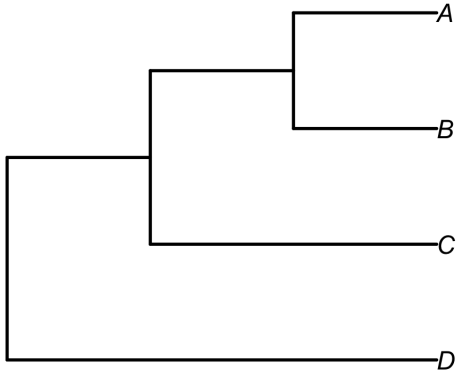
How to build a phylogeny 101

- Create a matrix of characters for your taxa of choice
- Create some hypothetical trees
- Count the number of changes required for each hypothetical tree
- Choose the tree with the fewest changes

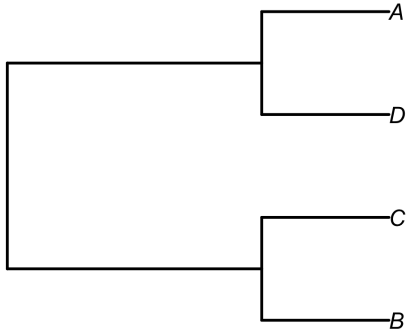
Trivial example ....

Character	*A*	*B*	*C*	*D*
1	1	1	1	0
2	1	1	0	0
3	0	0	0	1
4	1	1	0	0
5	0	0	0	1

Count minimum number of changes



5



7

Big problem ....

N. tips	*N. trees*
2	1
3	3
4	15
5	105
6	945
7	10,395
8	135,135

N. tips	*N. trees*
9	2,027,025
10	34,459,425
20	8.2008×10^{21}
30	4.9518×10^{38}
40	1.00985×10^{57}
50	2.75292×10^{76}

The number of trees increases at a phenomenal rate:

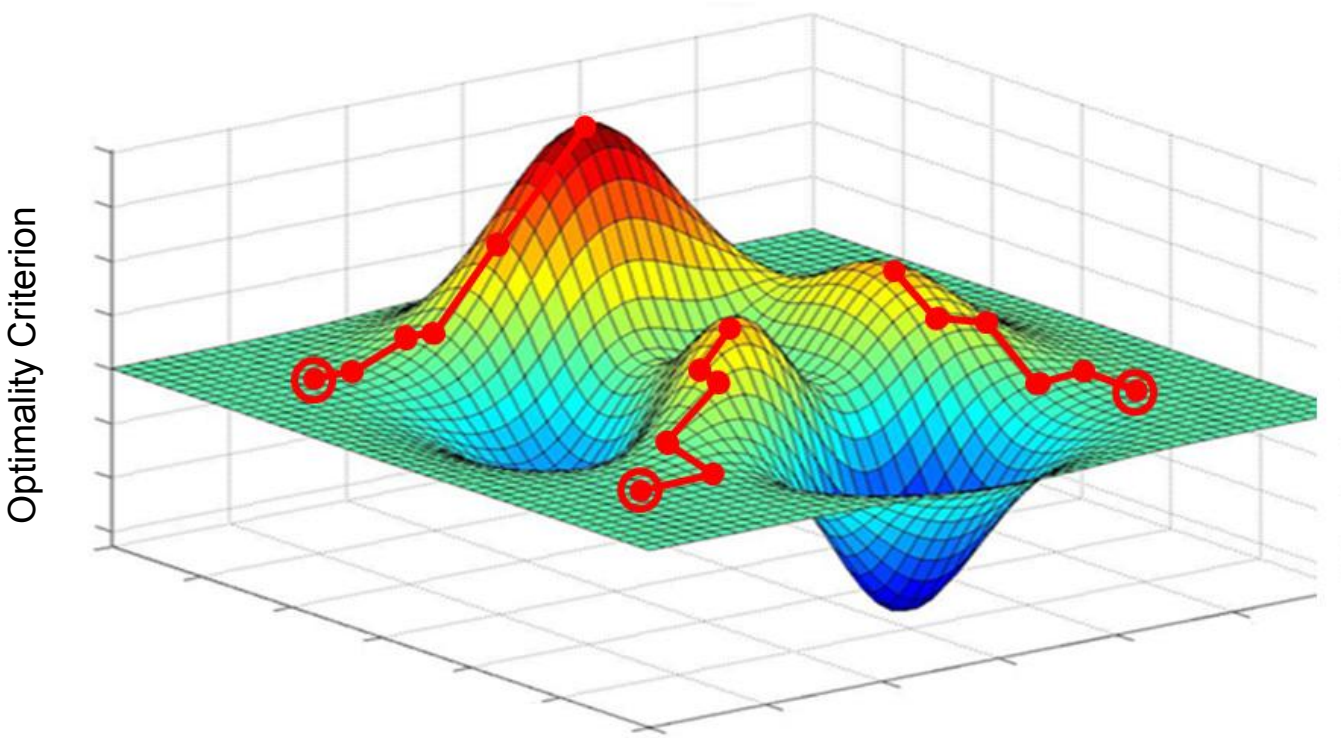
$$\frac{(2n - 3)!}{2^{n-2}(n - 2)!}$$

It is just not possible to try out all possible solutions.

$\sim N_{edd}$

Ref: Felsenstein, Inferring Phylogenies, Chapter 3.

Searching tree space



Cannot compute all trees and test.

Must instead use heuristic algorithms that search the range of possible tree shapes in order to find the best given an optimality criterion

Optimality criteria

What about branch lengths?

- Parsimony

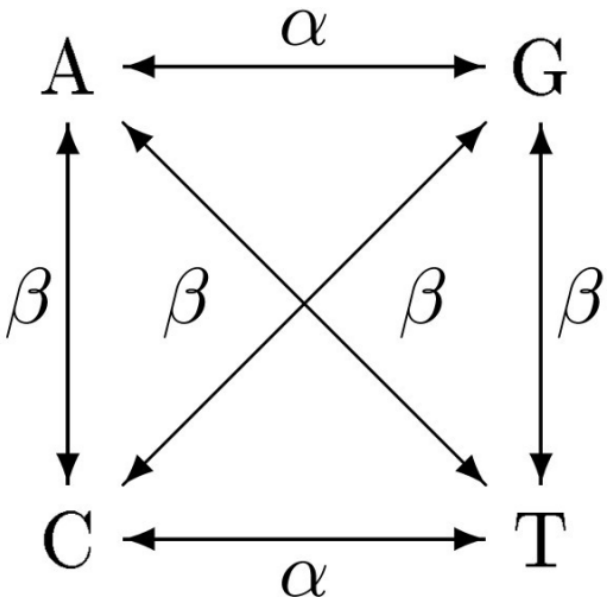
Necessary for datasets for which no appropriate model is available. e.g. morphological.

- Model based
  - Minimum evolution
  - Maximum Likelihood
  - Bayesian inference

Most common methods

What substitution model?

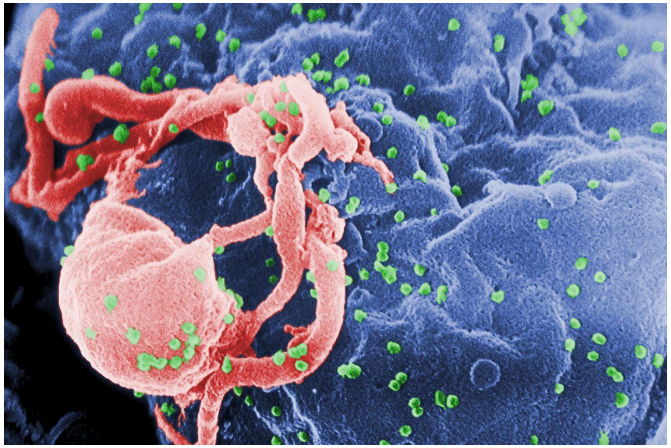
- Nucleotide models:
  - Jukes-Cantor
  - Kimura
  - *General Time Reversible (GTR)*
- Amino acid models:
  - Dayhoff
  - BLOSUM



Jukes-Cantor model, transitions and transervations have different rates.

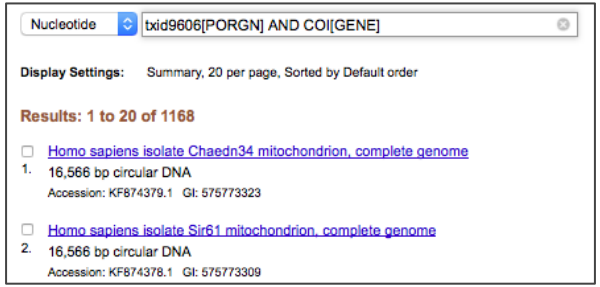
Genomics revolution

- With increasing data, phylogenetics might become easier
- Less reliance on models, more on presence absence
- Bigger character space
- Less character exhaustion
- e.g. viral fingerprints

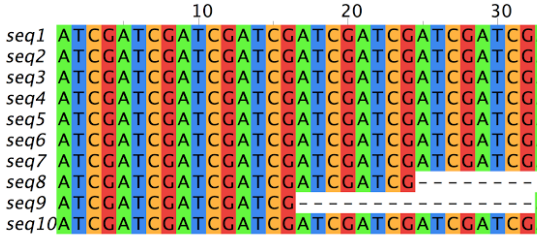


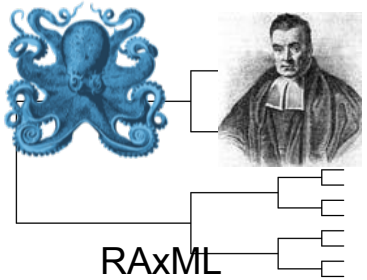
How to build a molecular phylogeny

Sequence retrieval → Sequence alignment → Phylogeny construction



NCBI GenBank





Sequence retrieval

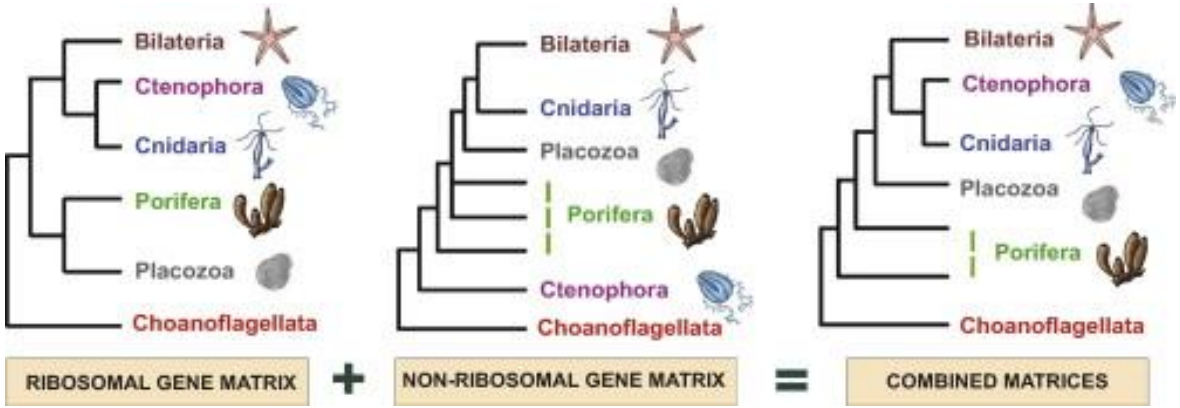
- What sequences are best?
  - Non-recombining -- *easy to align*
  - Constrained/core-gene -- *changes are neutral*
  - Single copies -- *easy to extract*
- Different genes change at different rates
  - nuclear recombinations > nuclear introns > nuclear exons > mitochondrial protein-coding > mitochondrial RNA > nuclear RNA

Sequence retrieval

Common examples:

- COI (mitochondrial protein, DNA barcode)
- CytB (mitochondrial protein)
- 12S (mitochondrial RNA)
- 28S (nuclear RNA)
- rbcL (chloroplast protein) ← Informs at multiple levels

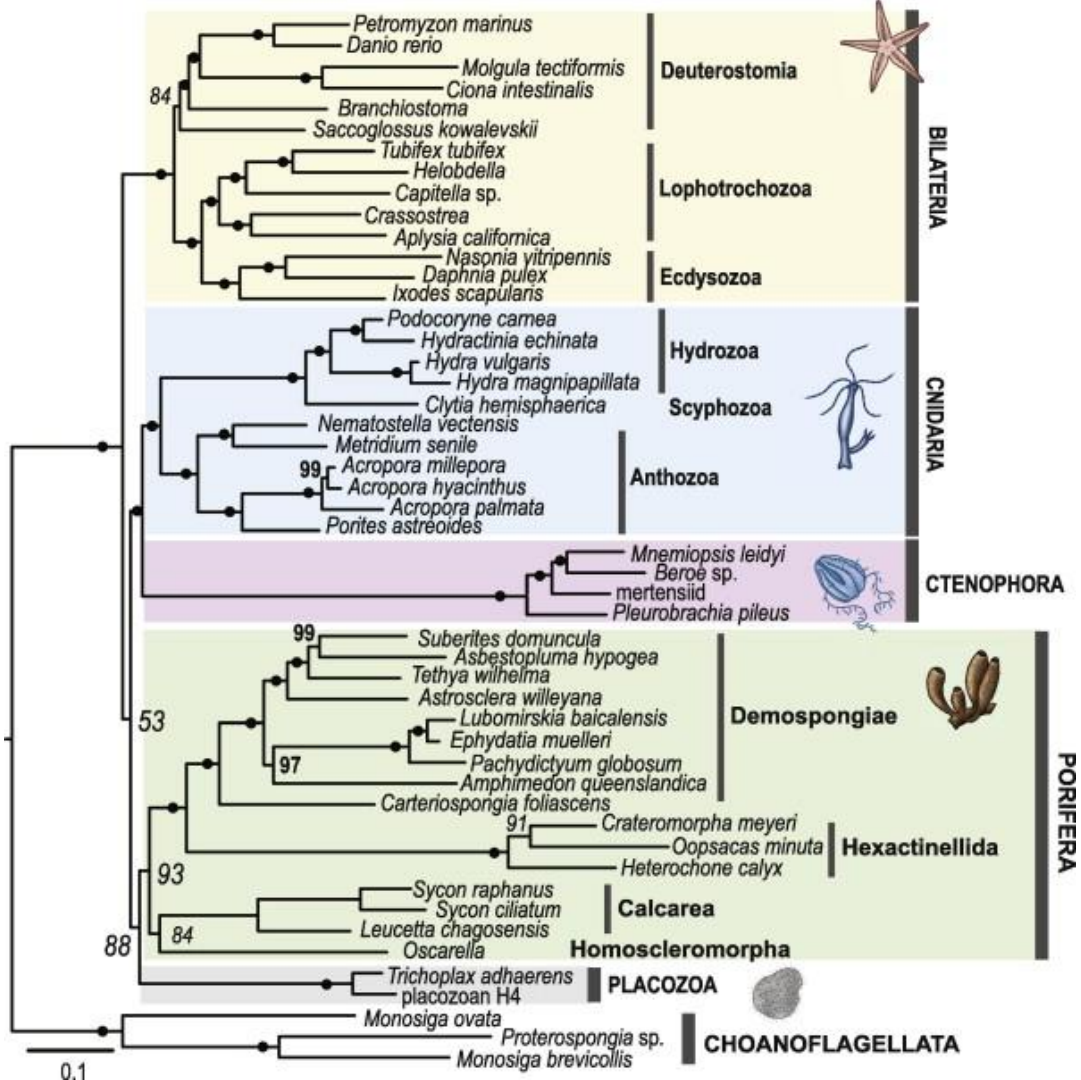
Genes inform at different levels



- Choose your sequences at the level at which you want to be informed
- To be informed across multiple levels, using a combination

Ref: Tetyana Nosenko, Fabian Schreiber, Maja Adamska, Marcin Adamski, Michael Eitel, Jörg Hammel, Manuel Maldonado, Werner E.G. Müller, Michael Nickel, Bernd Schierwater, Jean Vacelet, Matthias Wiens, Gert Wörheide, Deep metazoan phylogeny: When different genes tell different stories, Molecular Phylogenetics and Evolution, Volume 67, Issue 1, April 2013, Pages 223-233

For example, this metazoan tree needs genes that can inform across time spans of ~600MY

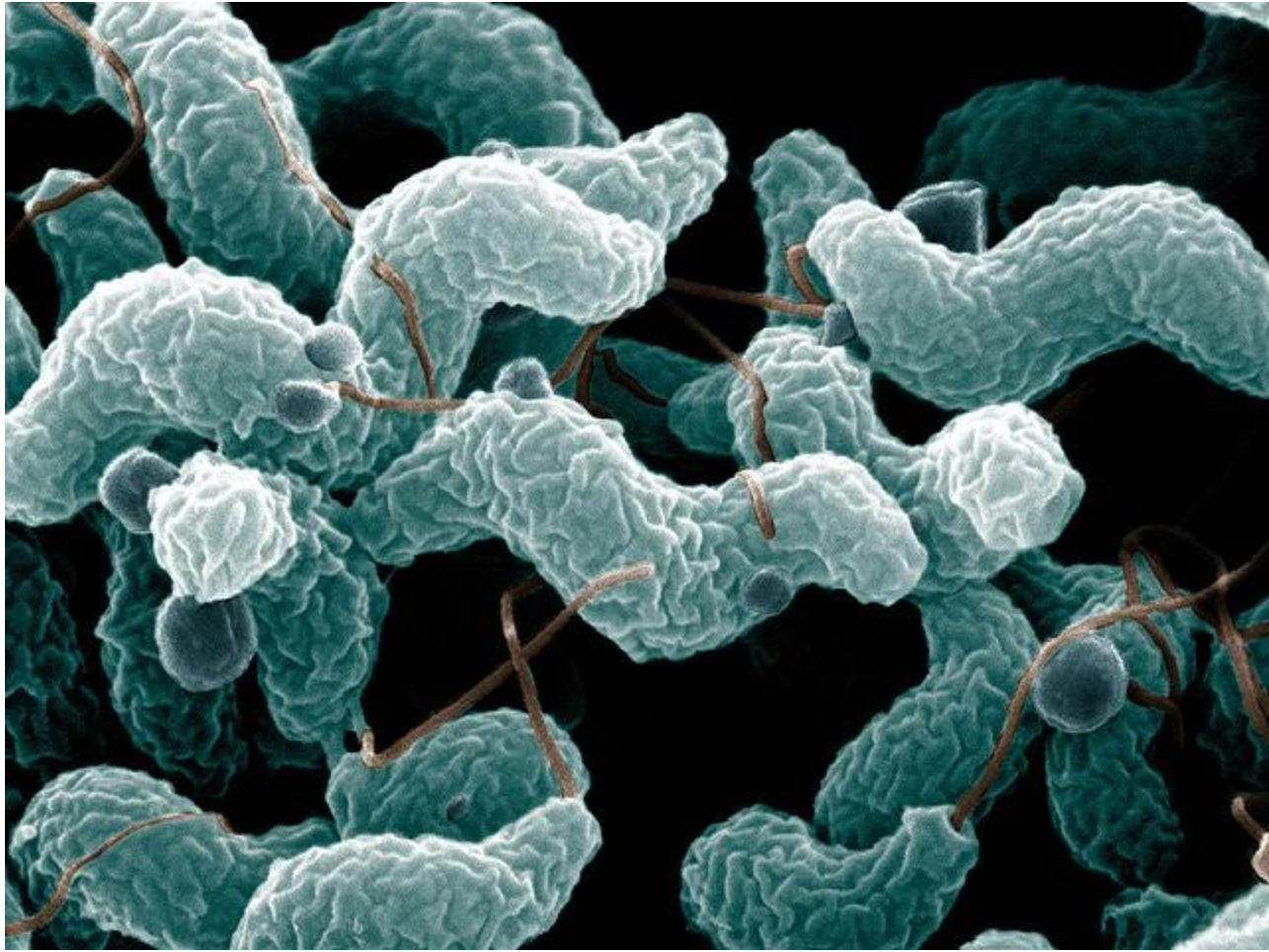


Ref: Tetyana Nosenko, Fabian Schreiber, Maja Adamska, Marcin Adamski, Michael Eitel, Jörg Hammel, Manuel Maldonado, Werner E. G. Müller, Michael Nickel, Bernd Schierwater, Jean Vacelet, Matthias Wiens, Gert Wörheide, Deep metazoan phylogeny: When different genes tell different stories, Molecular Phylogenetics and Evolution, Volume 67, Issue 1, April 2013, Pages 223-233

Eukaryotes are **hard**.

Archaea, bacteria and viruses are even **harder**.

The limiting factor for organisms with small genomes and fast rates of evolution is *character space* and *character exhaustion*.



Sequence alignment

- *Multiple* sequence alignment
- Often the hardest step of your analysis
- Models are no good if your sequences are not homologous

A range of methods

Different processes for different sequence types:

- *DNA* -- gap penalties, forwards and backwards translations which takes into account the wobble base pair
- *RNA* -- must take into account higher level folding structure of RNA
- *Amino acid* -- likewise with amino acid, these methods take into account the molecular properties of the amino acid sequences

MUSCLE

Clustal-Omega/W

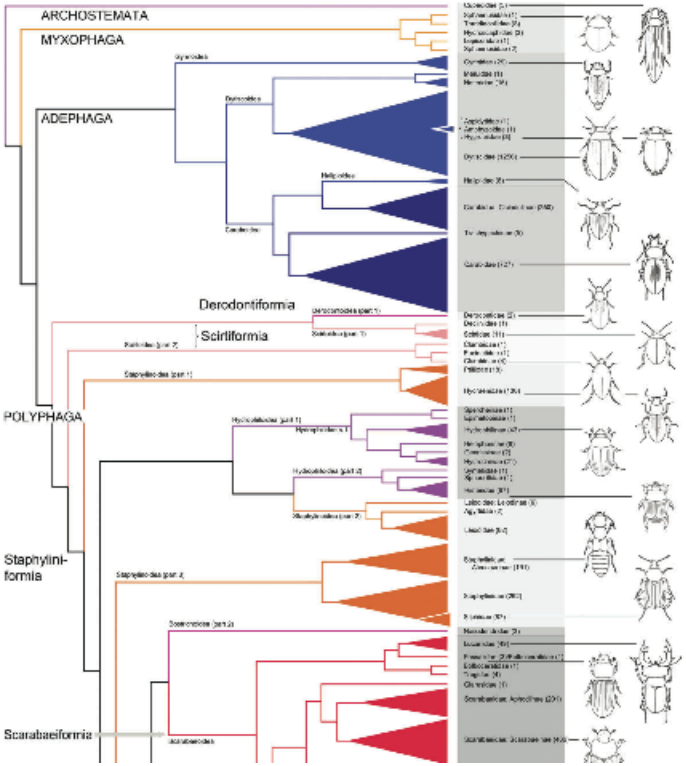
T-Coffee

Transalign

MAFFT

Supermatrices and rate-partitioning

- Allows you to use multiple genes
- Allows you to control for different rates within genes e.g. COI

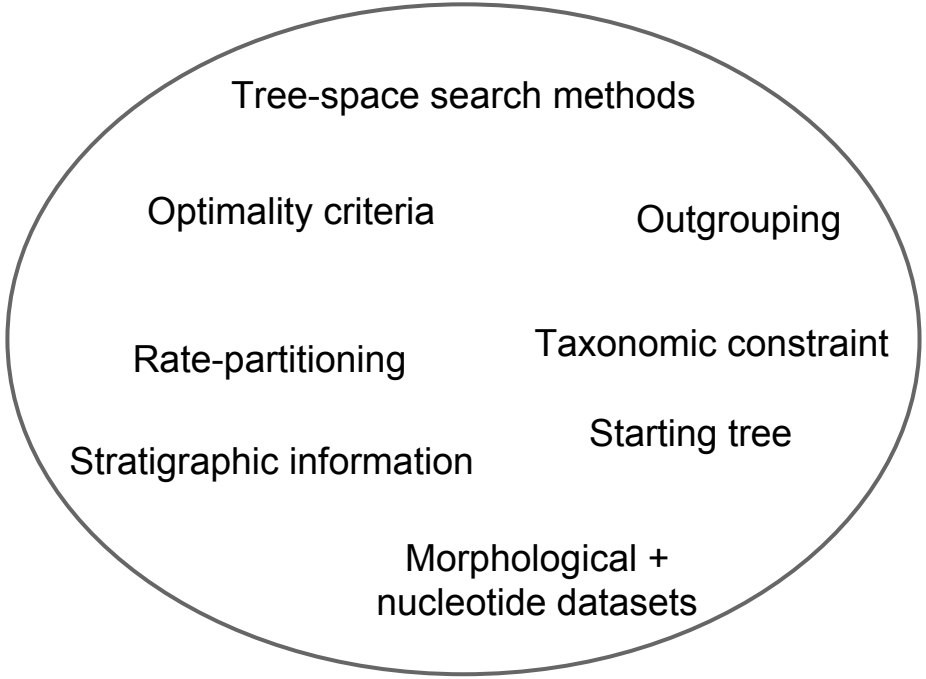


Ref: Bocak et al. (2014) Building the Coleoptera tree-of-life for >8000 species: composition of public DNA data and fit with Linnaean classification. *Systematic Entomology*

Tree construction

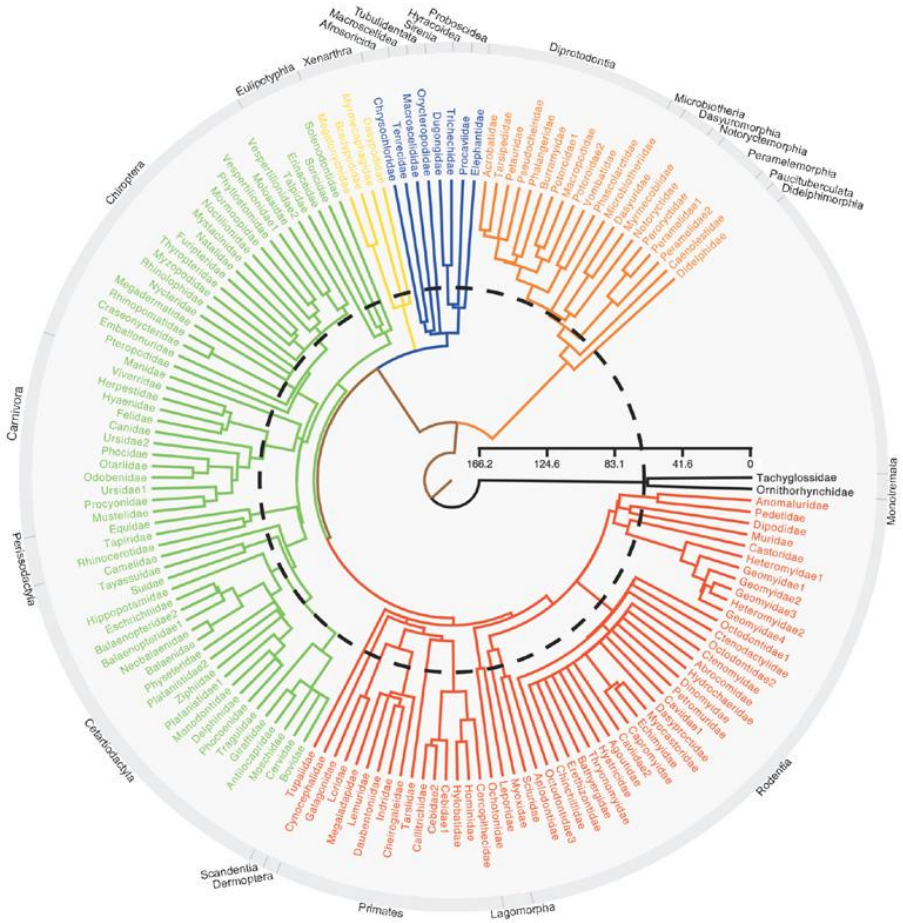
- Likelihood
  - PhyML
  - RAxML
- Bayesian
  - BEAST
  - MrBayes

A range of tools and methods



Further steps

- Tree testing
- Rate smoothing
  - BEAST or PATHD8
- Fossil calibration
  - PaleoDB
  - Fossil calibration database



Ref: Bininda-Edmonds et al. 2007. *Nature* 446, 507-512

Useful resources

- Felsenstein 2004 ‘Inferring Phylogenies’ -- *the* book on how-to phylogenetics
- www.timetree.org -- online resource of divergence dates between taxa
- www.treebase.org -- repository of trees and alignments
- blog.opentreeoflife.org -- project for the construction of the whole tree of life