Hands-on Phylogenetics

OKEE Workshop

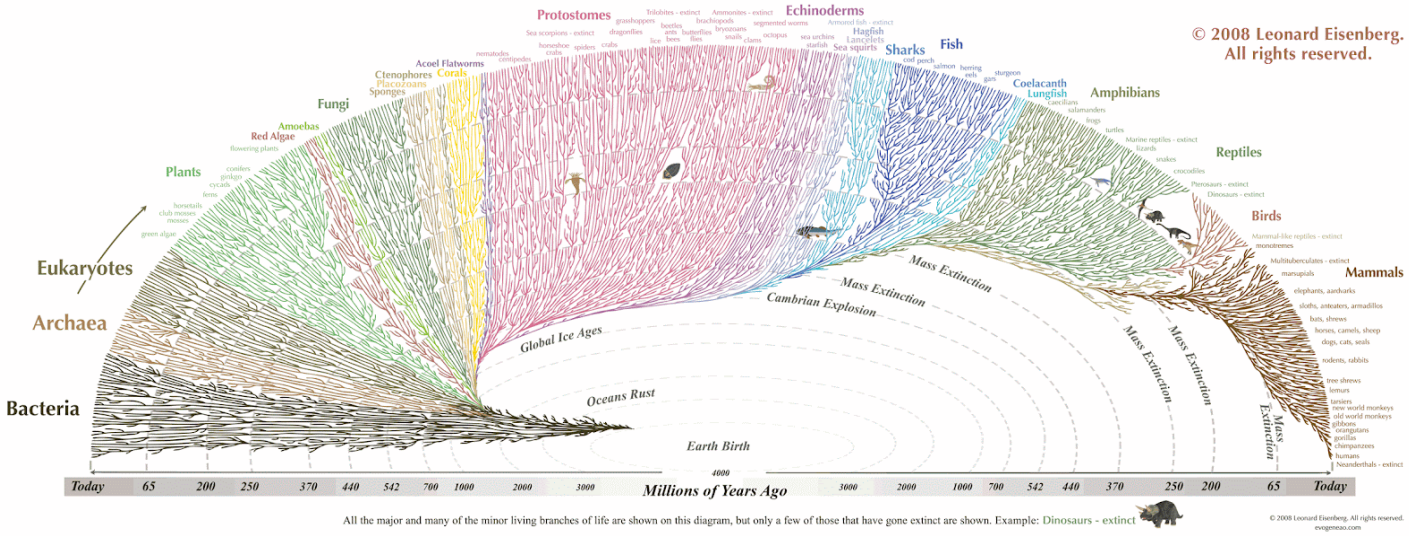
Orlando Schwery

Jan 4. 2023 ETHZ

1 – Basics & Main Concepts

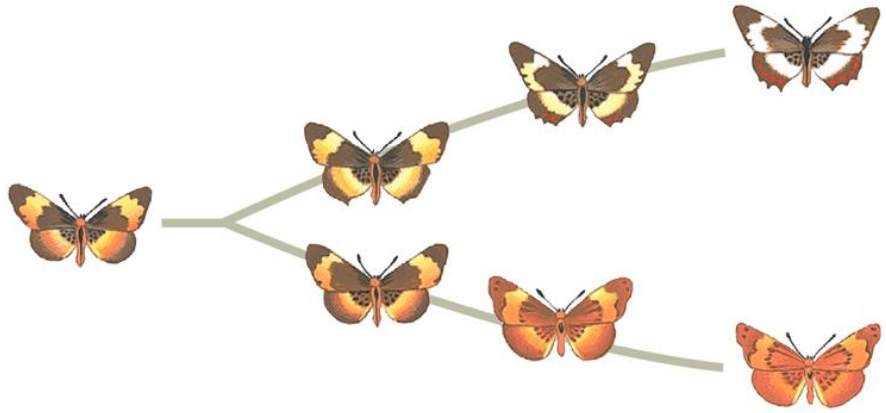
Systematics

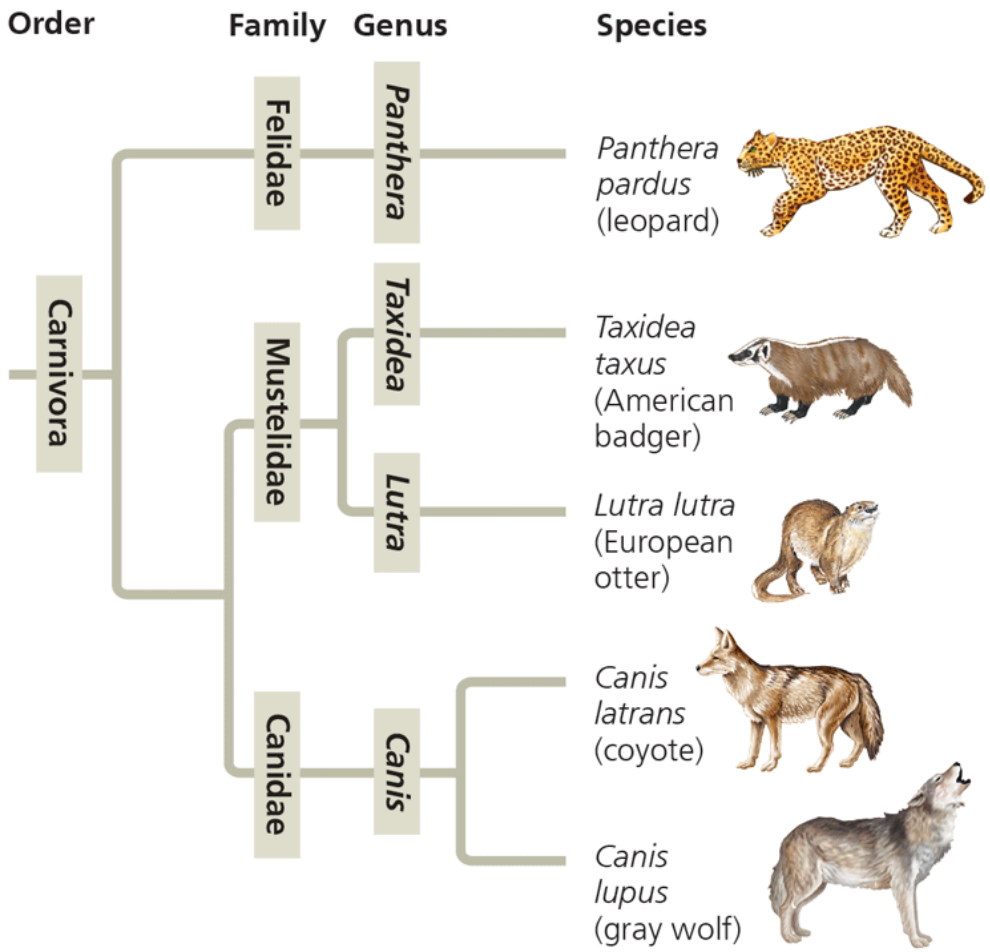
- We want to describe and understand the diversity on earth
- **Systematics** – biological discipline which includes…
  - **Taxonomy**
    - **Classification**
      Describing and organizing organisms
    - **Nomenclature**
      Naming organisms
  - **Phylogenetics**
    Determining evolutionary relationships



Phylogeny

- Speciation turns 1 lineage into 2 (mostly)
- Through evolutionary processes, all life is related that way
- Taxonomy does not fully reflect this (and is sometimes misled)
- Phylogenetic tree / Phylogeny to depict evolutionary relatedness (hierarchical diagram)

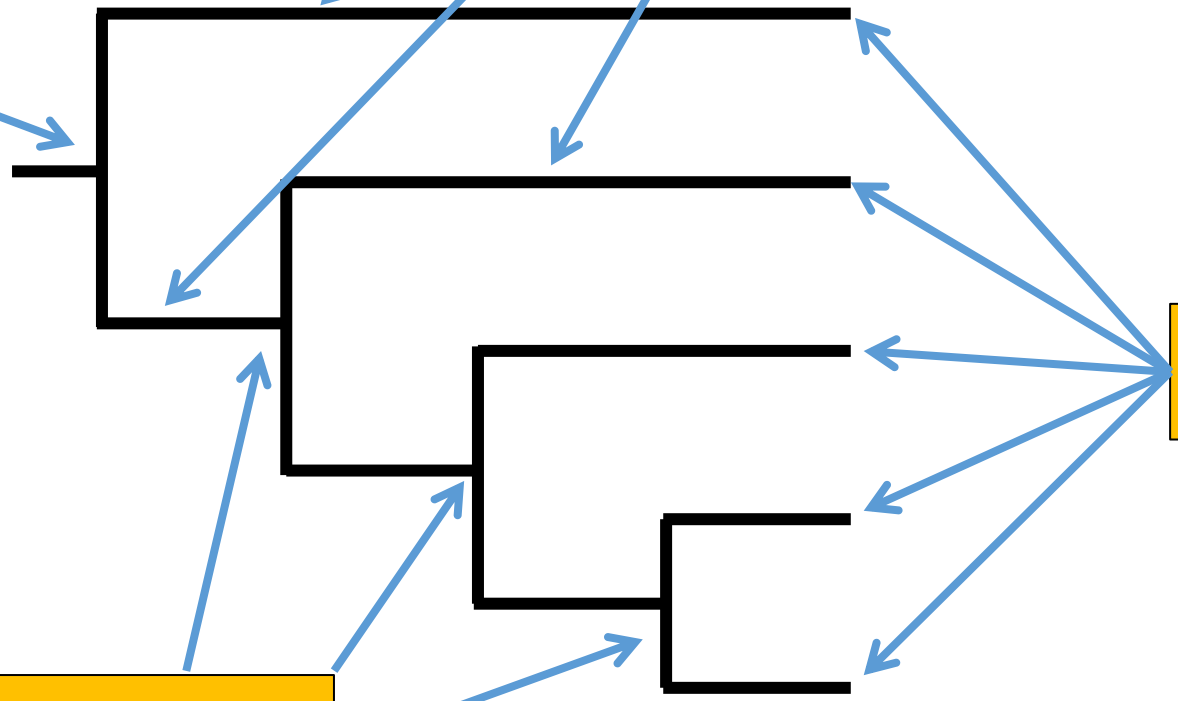




Phylogeny

Branches/Edges
Represents evolutionary lineage, each point on it is an ancestor leading to the species at the tips
→ Length means nothing unless specified to mean time or evolutionary distance (i.e. #mutations)

Root/Root Node
Represents common ancestor of all taxa in tree, lies farthest in the past

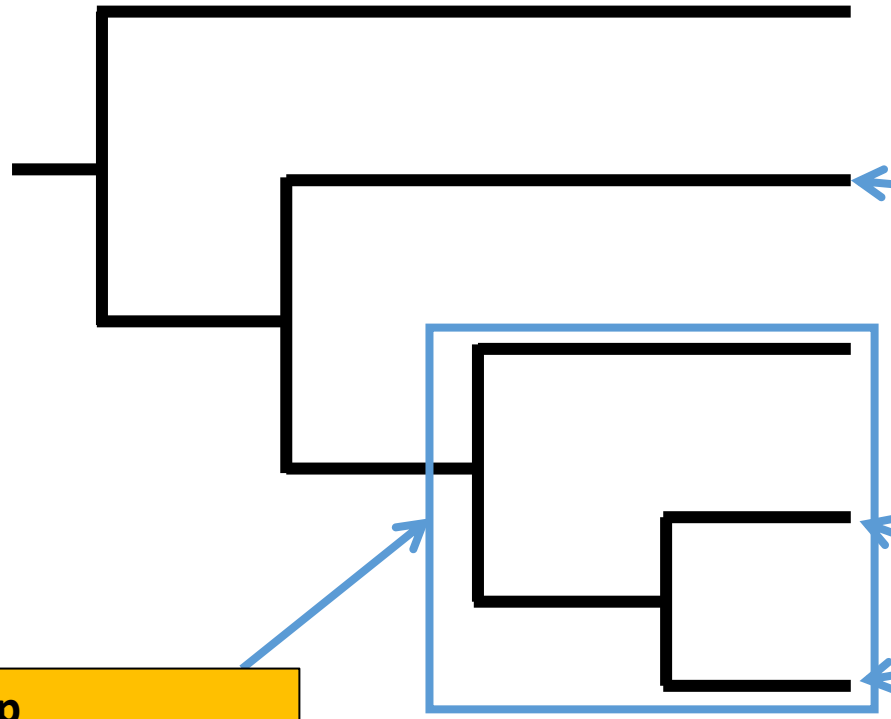


Tips (leaves)
Usually species

Nodes
Branching point, presumed speciation event, common ancestor of the taxa that sprout from it

Phylogeny

There is no such thing as a Basal Taxon!
Despite what the textbook says…



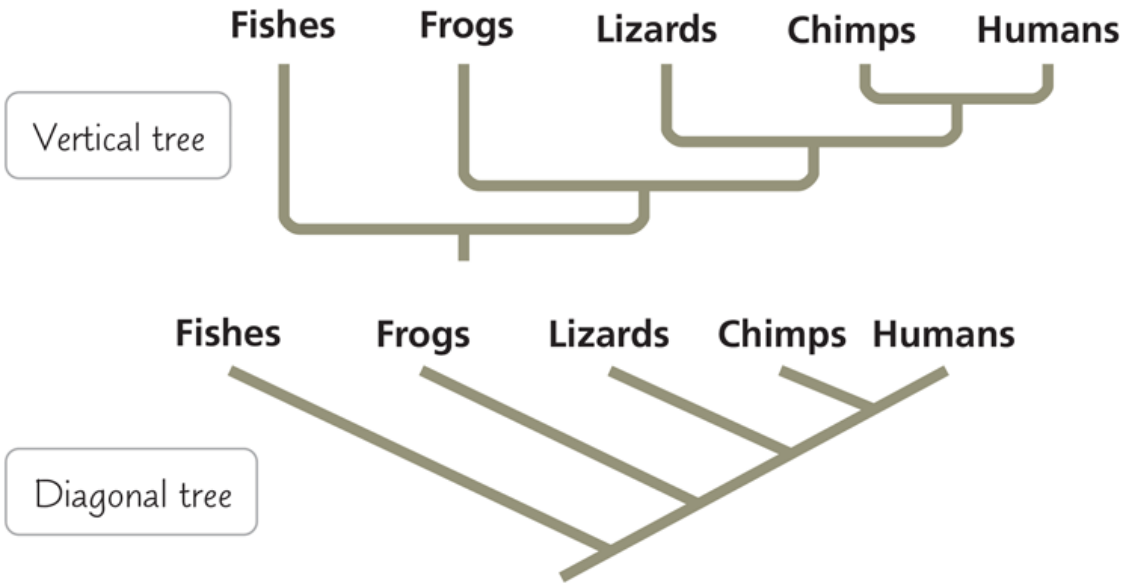
Sister Groups
Concept also works for larger taxa

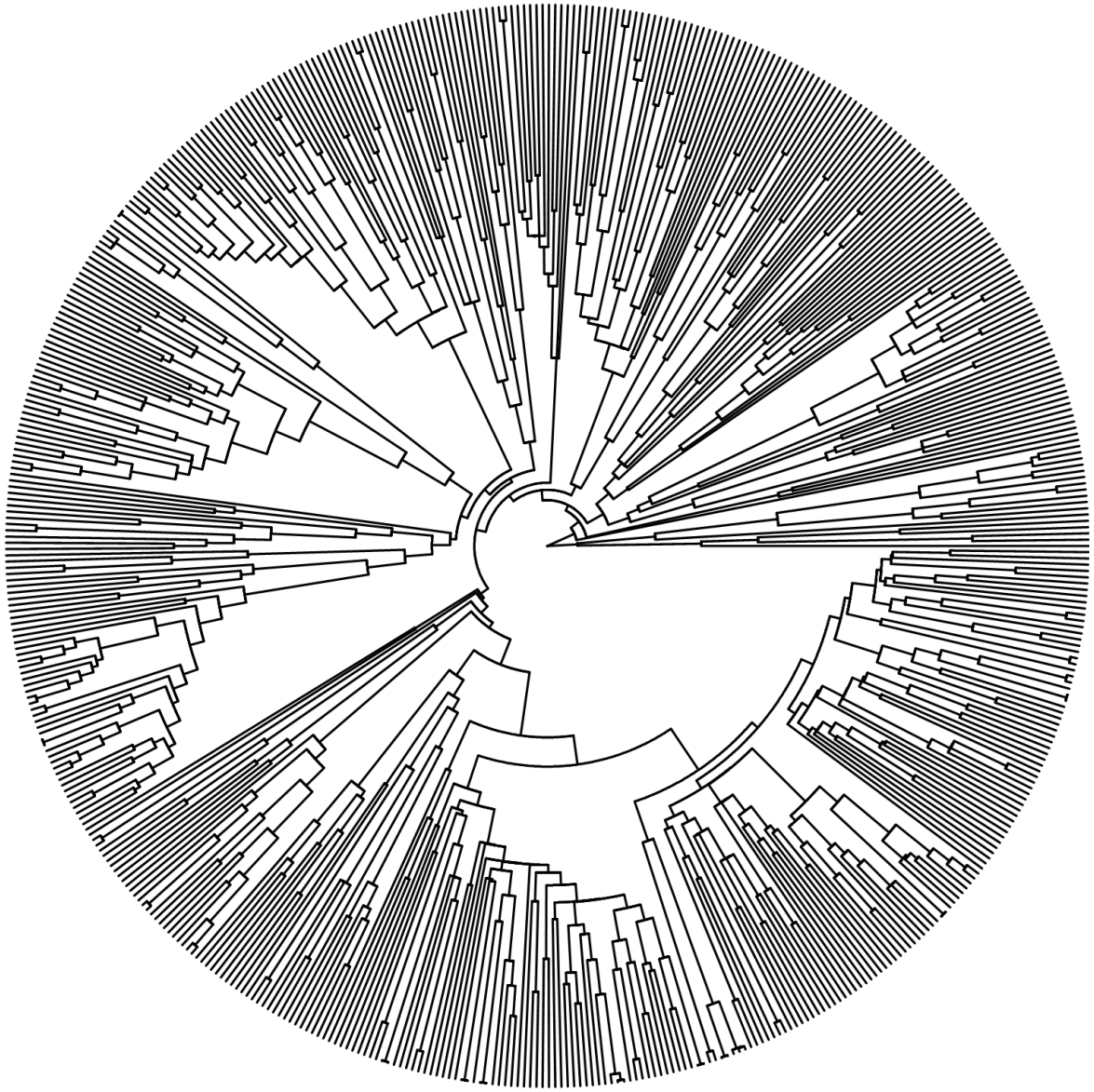
Sister Taxa/Species
Share common ancestor only with each other, are each other's closest relative

Clade / Monophyletic Group
Set that includes the ***Most Recent Common Ancestor*** and all its descendants

Phylogeny

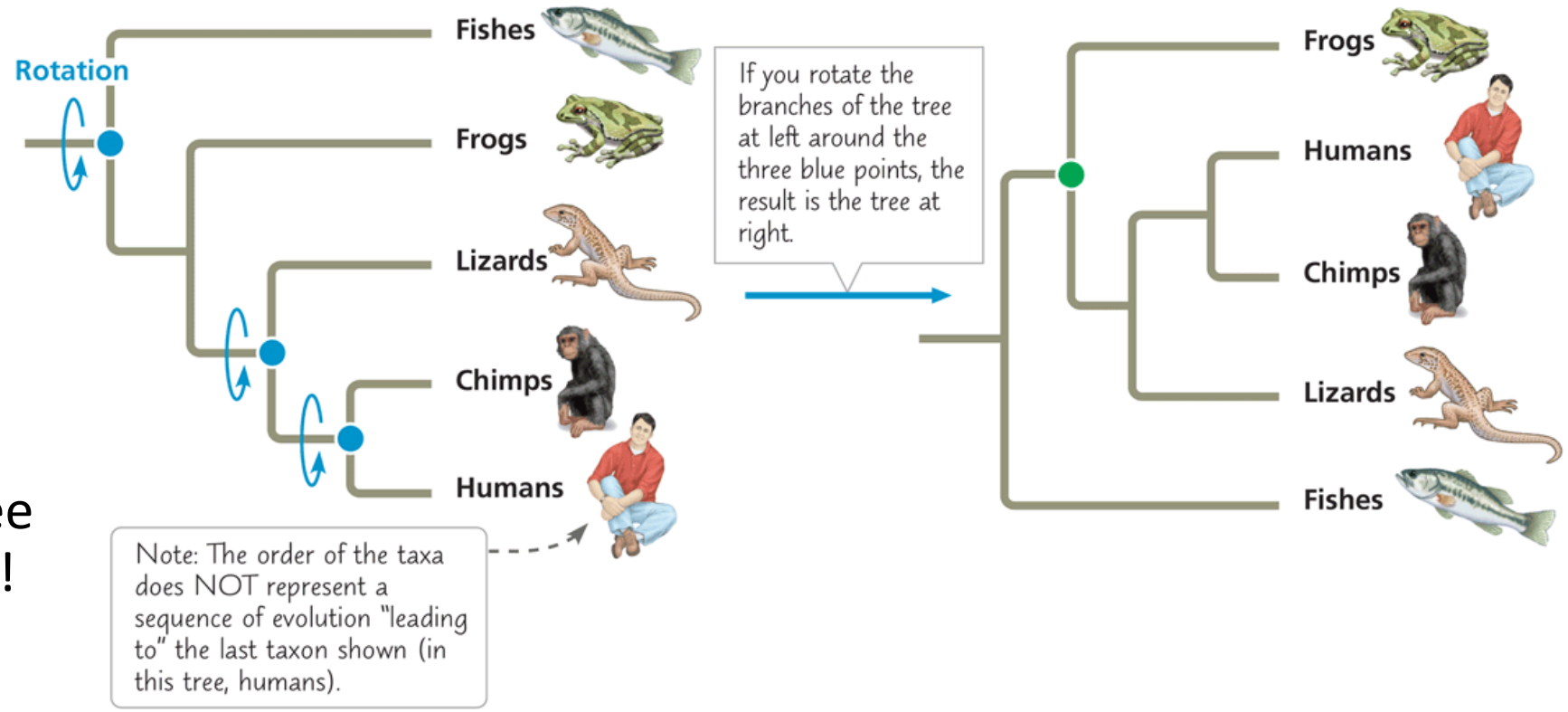
- Different graphical styles represent the same tree





Phylogeny

- We can rotate the branches and the tree still means the same!

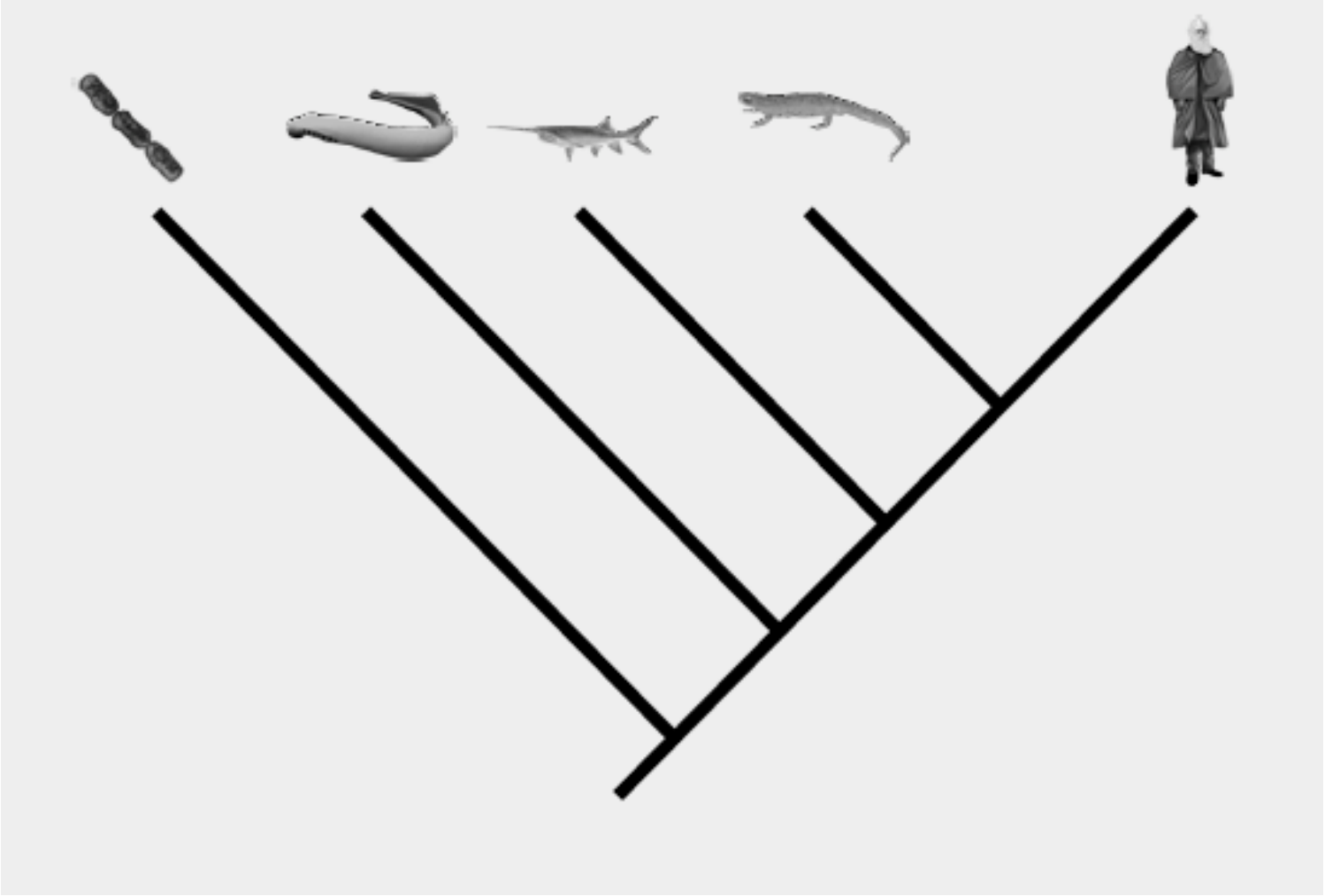


→Relatedness has to do with the Most Recent Common Ancestor (MRCA) and NOT with which tips are next to each other!

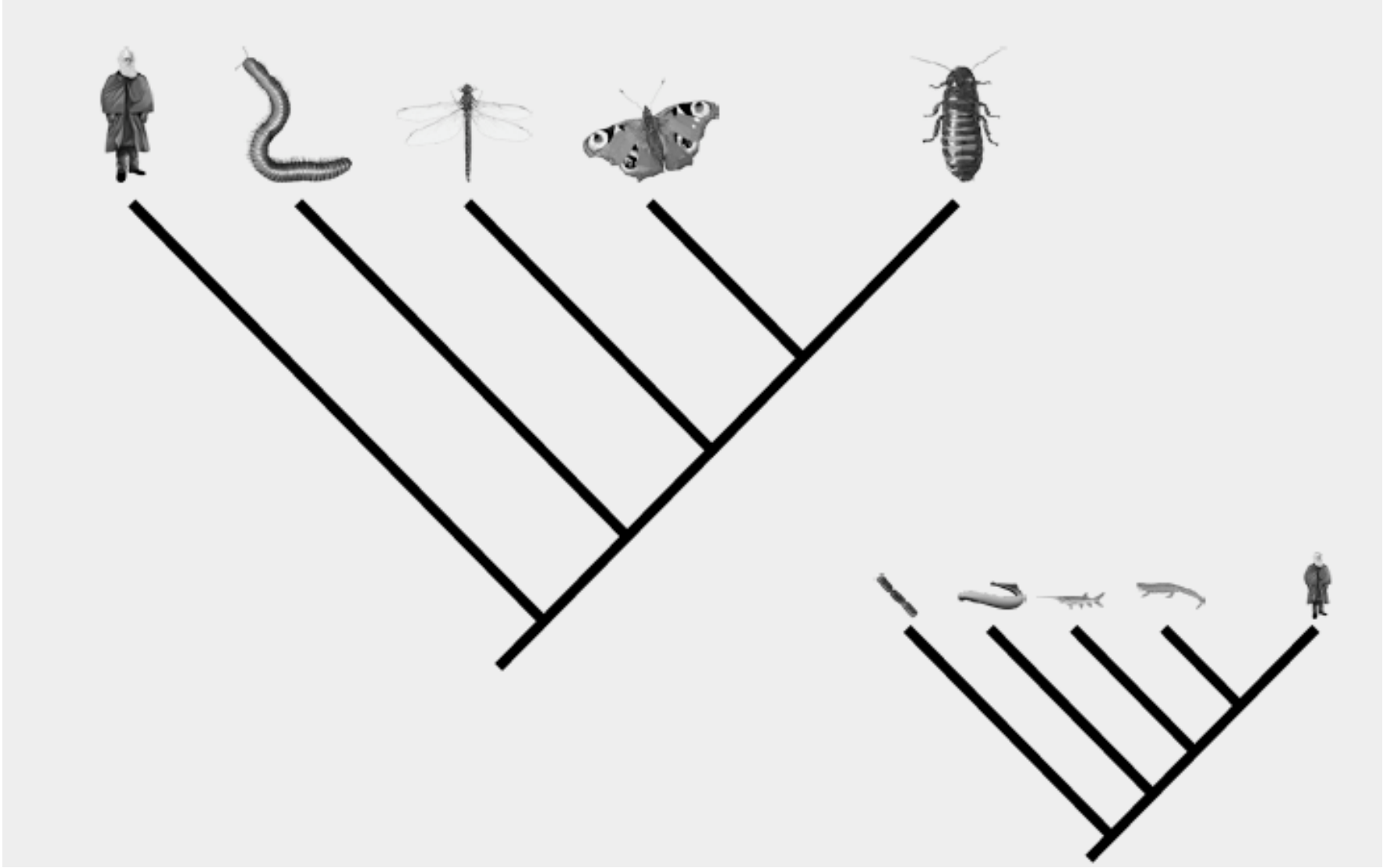
→Living species are always at the tips, never at nodes or along branches!

Phylogeny

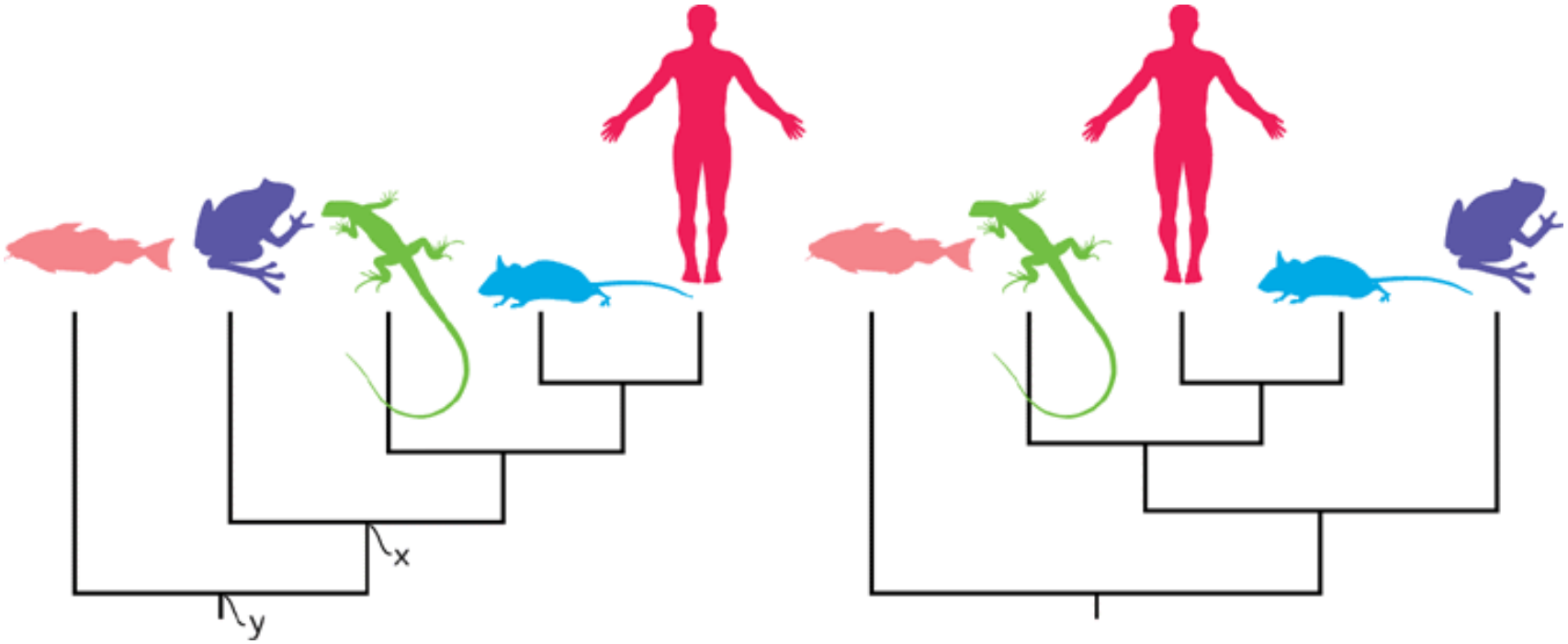
- Anthropogenic "conceit of hindsight"
- Perception that most "evolutionary divine" species at top of trees



Phylogeny



Phylogeny



Which one is ‘true’?

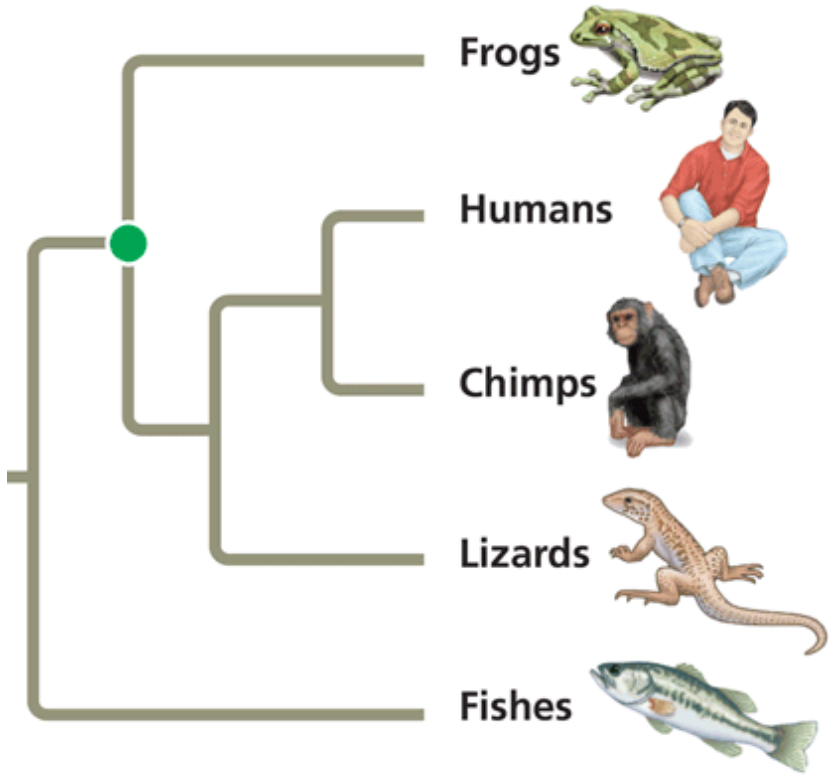
Phylogeny

What this tells us:

- Relatedness of taxa
  - Sister taxa
  - Larger clades
  - Degree of relatedness

→ But keep in mind this is a **hypothesis** (and likely incomplete)

(*please also note that this example tree does not represent reality very well…*)

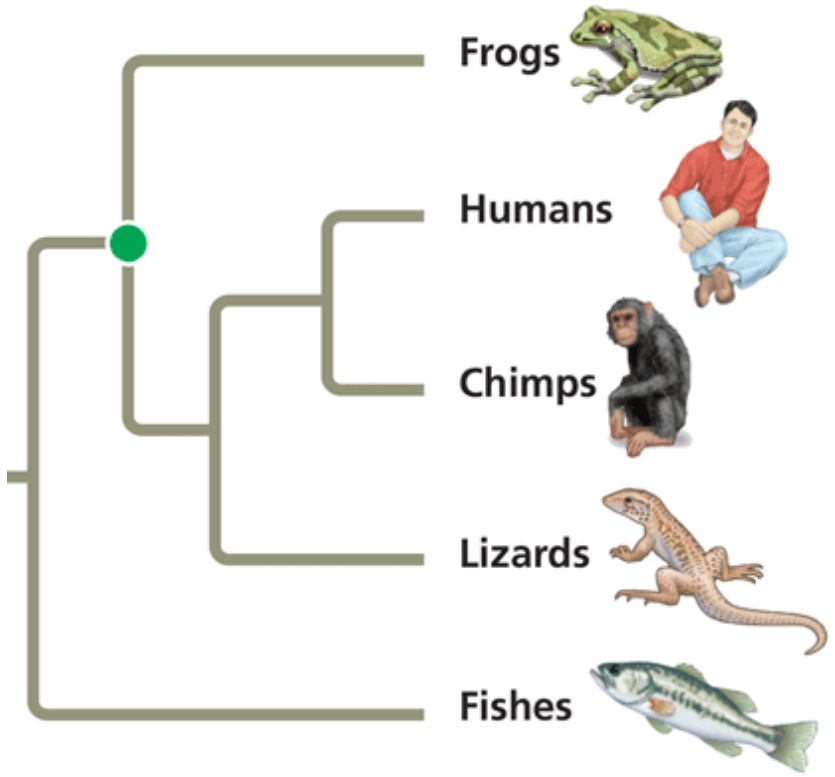


Phylogeny

(*please also note that this example tree does not represent reality very well…*)

What this does NOT tell us:

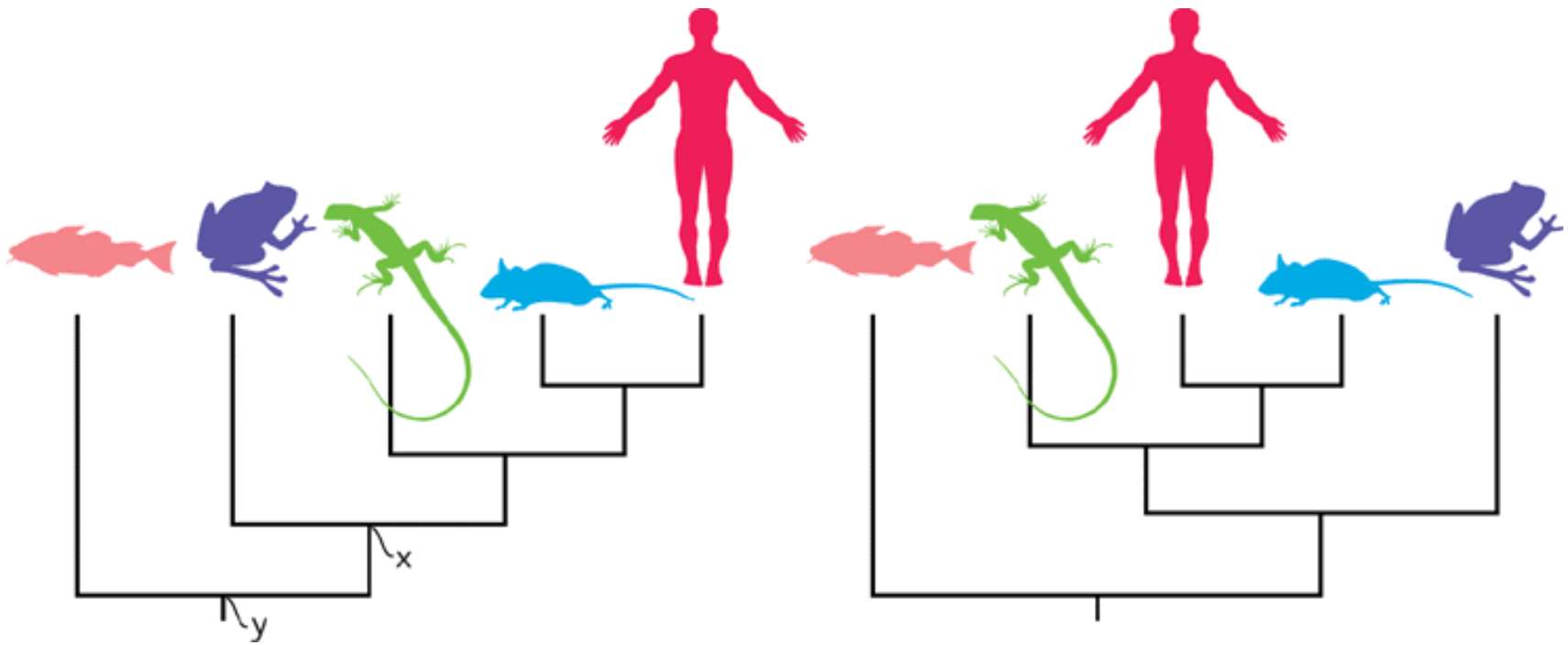
- Phenotypic similarity
  (though generally expected)
- Ages of taxa
  (unless branch lengths are meant to represent time)
- Taxa didn’t evolve from the ones next to it!
- Branch near node represents ancestor that can be quite different from the descendant at the tip it leads to!



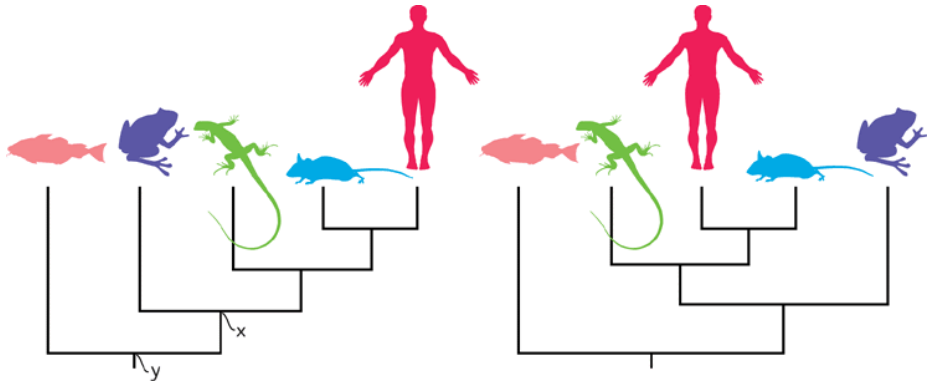
EVOLUTION

The Tree-Thinking Challenge

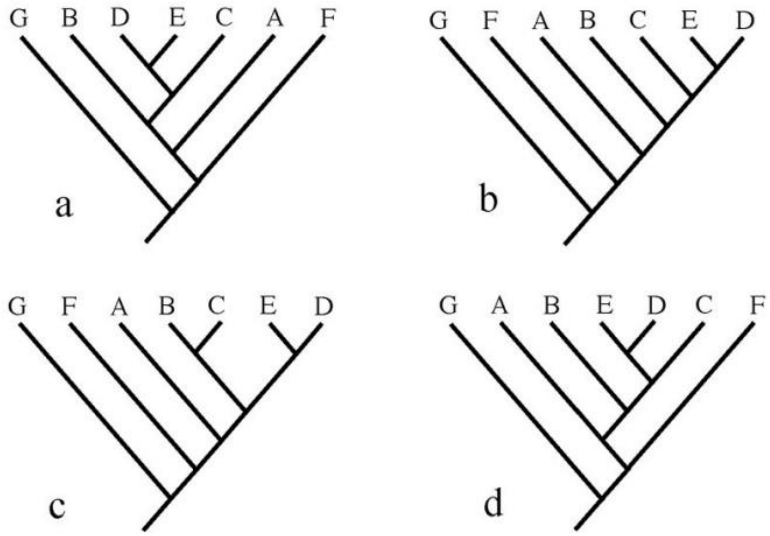
David A. Baum, Stacey DeWitt Smith, Samuel S. S. Donovan



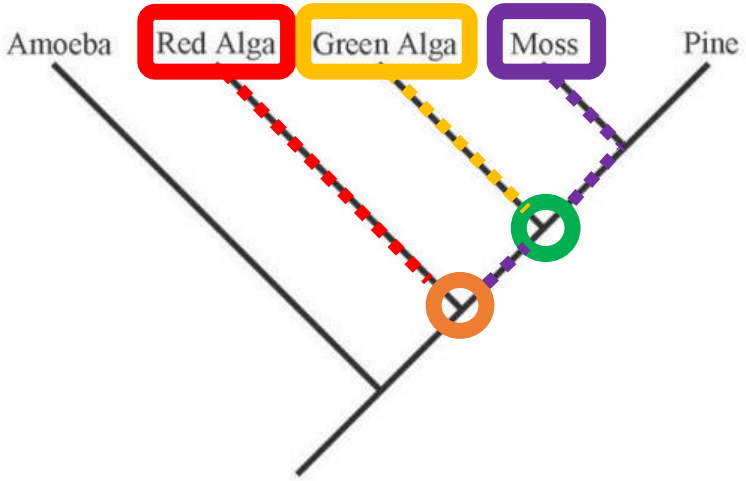
Recap - Tree Thinking



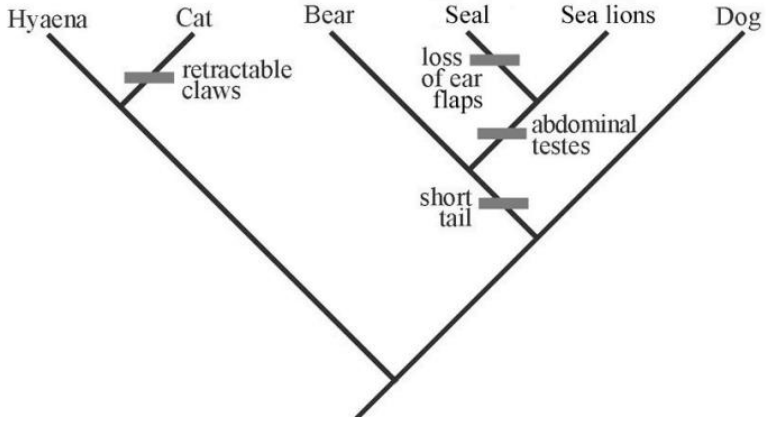
Rotated branches don't change the relationships!



Tell different trees by investigating sister relations



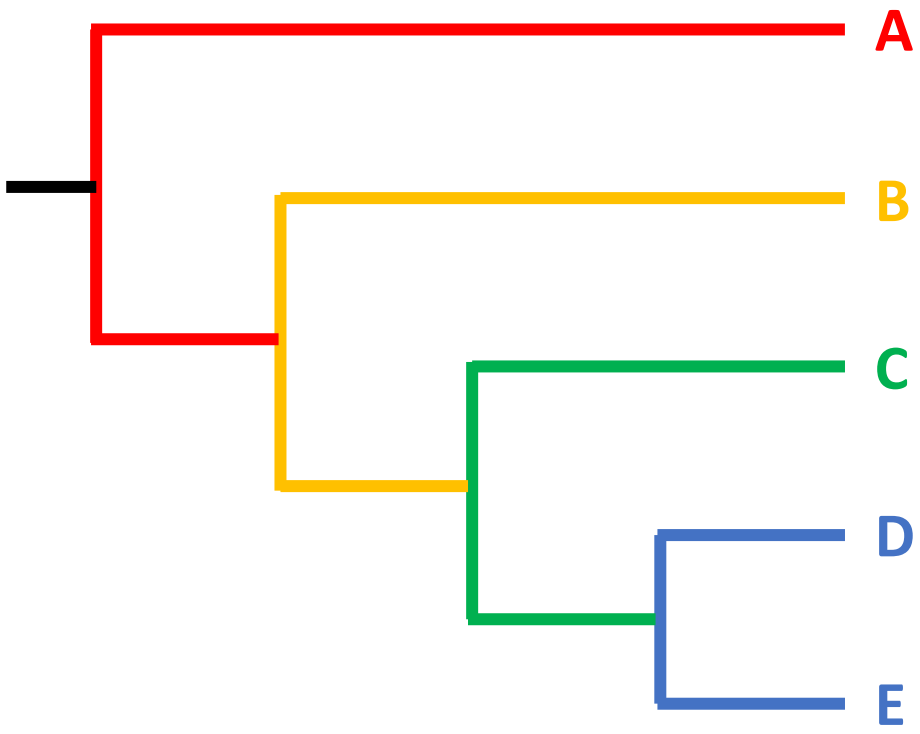
Tell relatedness by comparing MRCA



Trace traits from ancestors to descendant species

2 – Tree-Files and Phylogenies in R

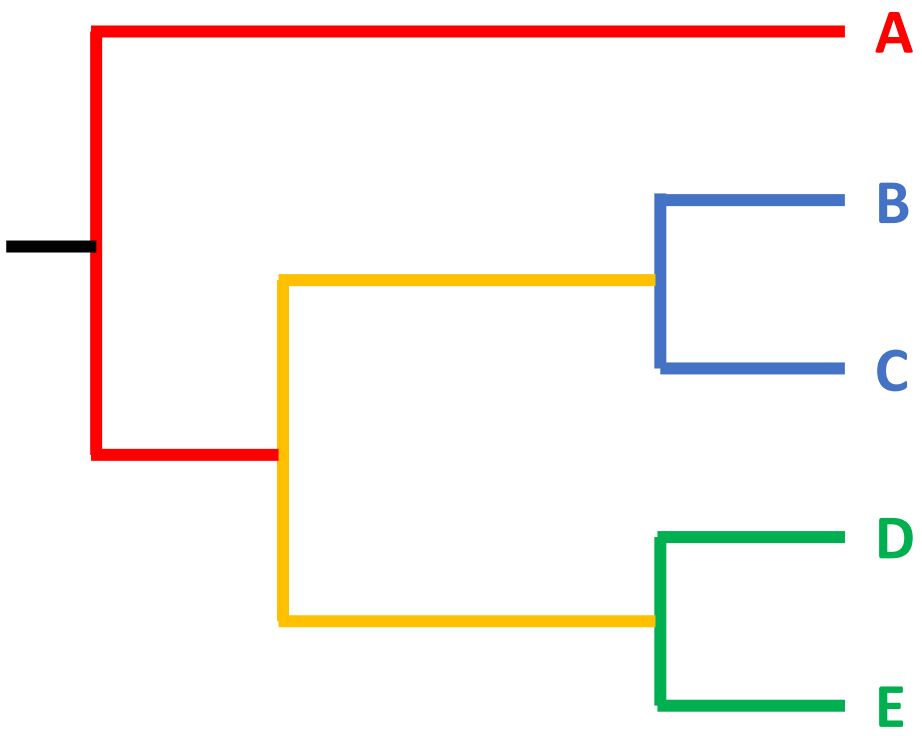
How to encode a phylogeny as text for a computer to read?



- Brackets to encapsulate closest relatives:

(A (B (C (D E))))

How to encode a phylogeny as text for a computer to read?



- Brackets to encapsulate closest relatives:

What would this one look like?

(A ((B C) (D E)))

Main Formats include additional Information:

Newick

```
(((erHomoC:0.28006,erCaelC:0.22089):0.40
998,(erHomoA:0.32304,(erpCaelC:0.58815,(
(erHomoB:0.5807,erCaelB:0.23569):0.03586
,erCaelA:0.38272):0.06516):0.03492):0.14
265):0.63594,(TRXHomo:0.65866,TRXSacch:0
.38791):0.32147,TRXEcoli:0.57336);
```

- Parentheses with tip labels and (here) branch lengths
- Nexus translates and lists tip labels, and can include data and other info on the tree

https://evomics.org/resources/tree-formats/

Nexus

```
#NEXUS

Begin trees; [Treefile saved Wed Jul 26 19:40:41 2000]
[output from your data run] Translate

1 TRXEcoli,

2 TRXHomo,

3 TRXSacch,

4 erCaelA,

5 erCaelB,

6 erCaelC,

7 erHomoA,

8 erHomoB,

9 erHomoC,

10 erpCaelC

;

tree PAUP_1 = [&U]
(1,((2,3),((((4,10),(5,8)),(6,9)),7)));

End;
```

How R deals with phylogenies…

BREAK

3 – Overview on Getting a Tree

The easy way: use an existing phylogeny!

- Making data available is more common
- Search for studies on your group and look for tree in
  - Supplemental material
  - Links to repositories
- Old fashioned: “available upon request…”
  - Just email the authors and ask (and hope they didn’t lose it)

- Data repositories like TreeBase etc.

The still kinda easy way: use aggregated/synthetic trees

- The Open Tree of Life Project
  - Giant project of permanently updating phylogenetic data
  - Build giant tree of existing phylogenies/data
  - Add missing taxa based on taxonomic knowledge
  - Ever growing features (e.g. node ages)

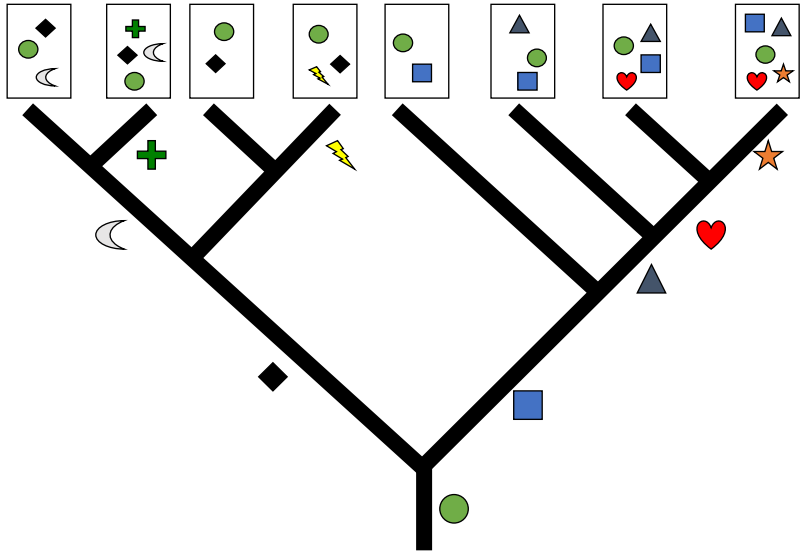
- Check out their website!

- R (and other) interfaces to access data

How to get a tree using `rotl`…

How to build phylogenies?

- Use trait data from extant species
  - Morphological
  - Molecular (DNA/RNA/AS/…)
- Must be result of common ancestry to be useful!
  - Homologies, not Analogies!
  - ***Homology***: similar due to shared ancestry – inherited
  - ***Analogy***: similar due to similar selective pressures – convergent evolution
- Must not be the same in all species or only occur in one to be informative…



- **Character:** Distinguishing feature (morphology, behaviour, molecular)

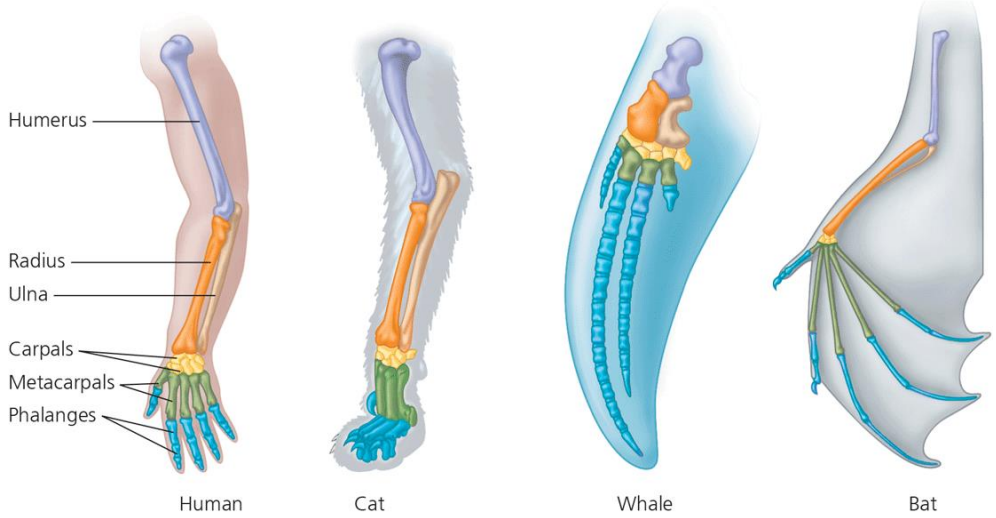
  e.g. eye colour

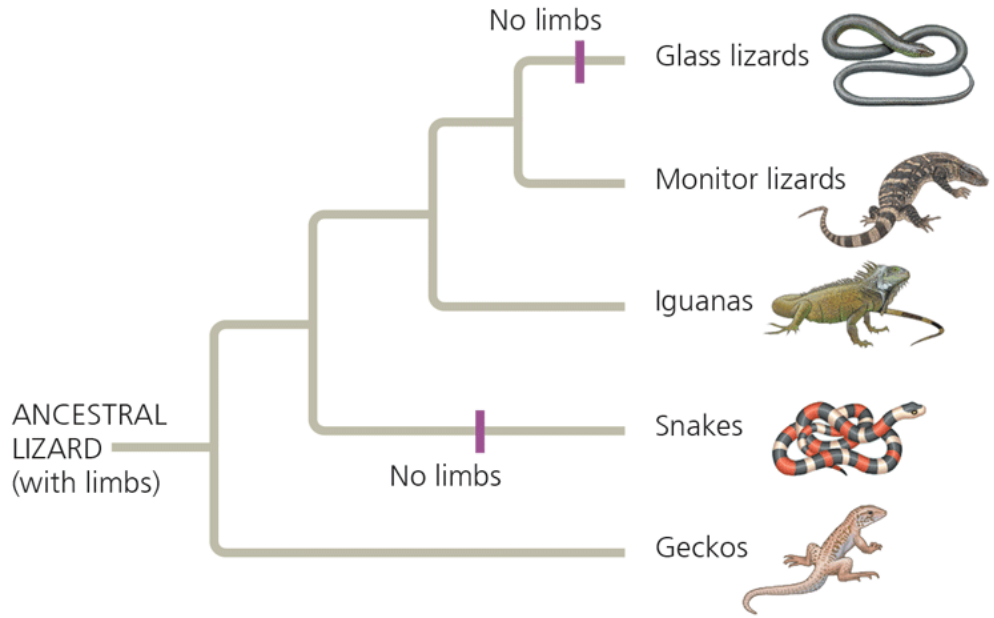
- **Character state:** actually present variant/expression of this feature

  e.g. blue, green, brown

How to build phylogenies?

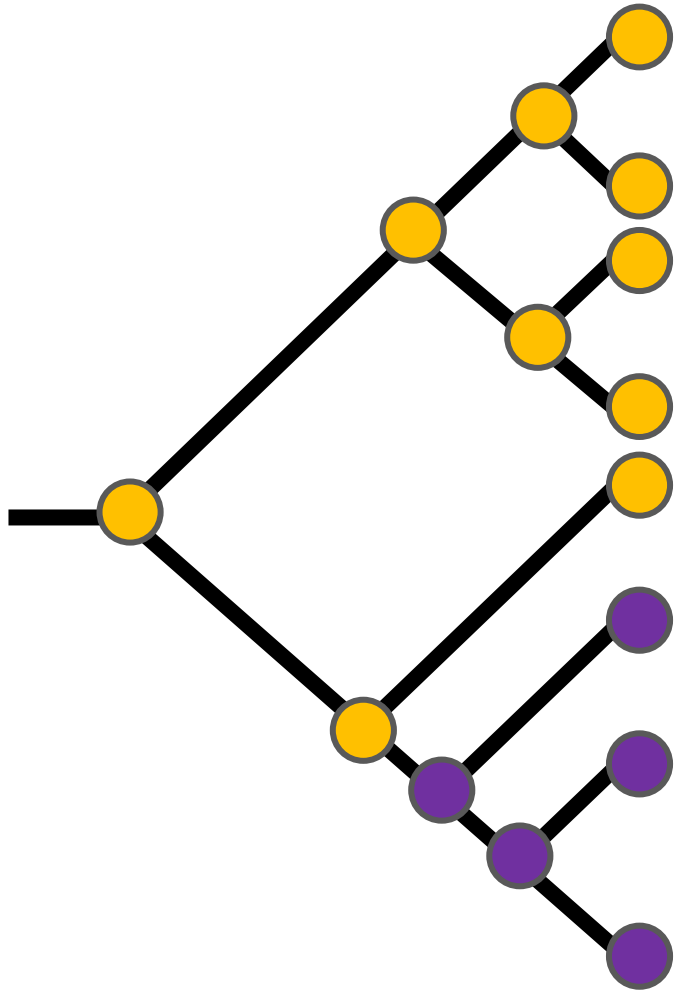
- Use trait data from extant species
  - Morphological
  - Molecular (DNA/RNA/AS/…)
- Must be result of common ancestry to be useful!
  - Homologies, not Analogies!
  - ***Homology***: similar due to shared ancestry – inherited
  - ***Analogy***: similar due to similar selective pressures – convergent evolution
- Must not be the same in all species or only occur in one to be informative…





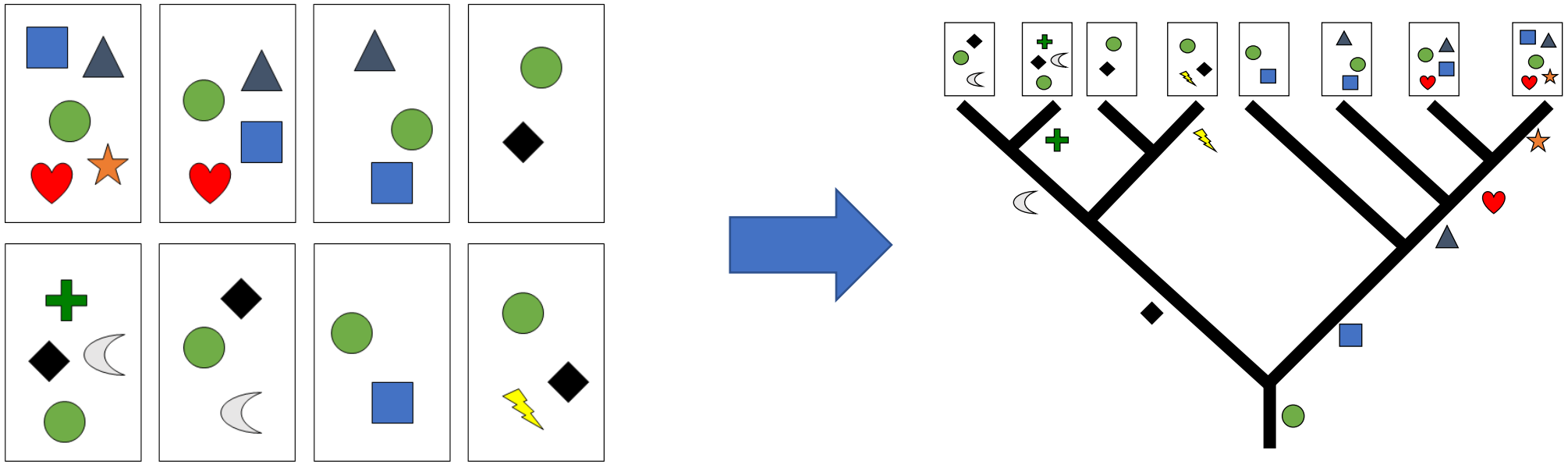
Evolutionary Order of Character States

- Homologous characters can be:
  - **basal** (ancestral, ‘primitive’)
    The older/’original’ state of a character, found in the ancestor of a lineage and all its descendants.
    → ***plesiomorphy***
  - **derived** (evolved, novel, ‘modern’)
    The younger state of a character that evolved from the basal state, only present in some particular lineages and their descendants.
    → can also be the *loss* of a character
    → ***apomorphy***

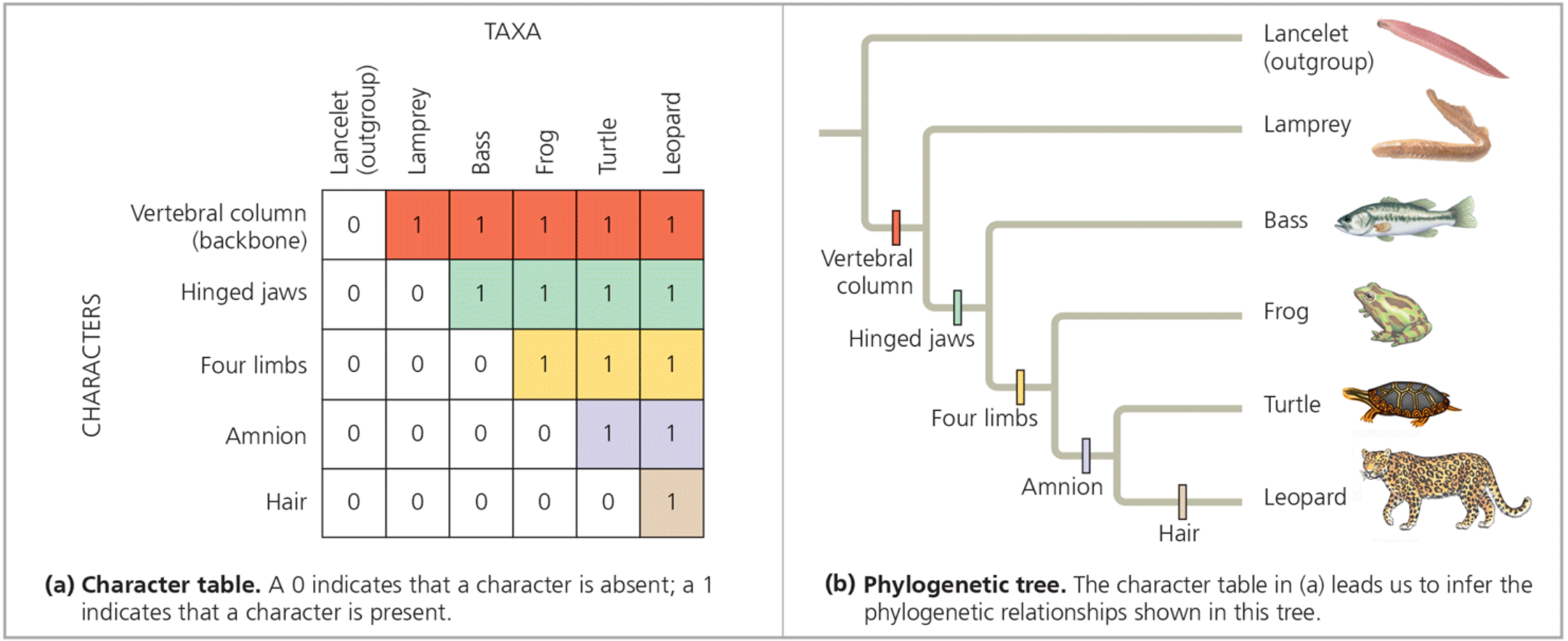


Back to the making of trees…

Knowing all this, how do we get from character data to a phylogeny with meaningful clades?



Trees from Morphological Character Matrices



(a) Character table. A 0 indicates that a character is absent; a 1 indicates that a character is present.

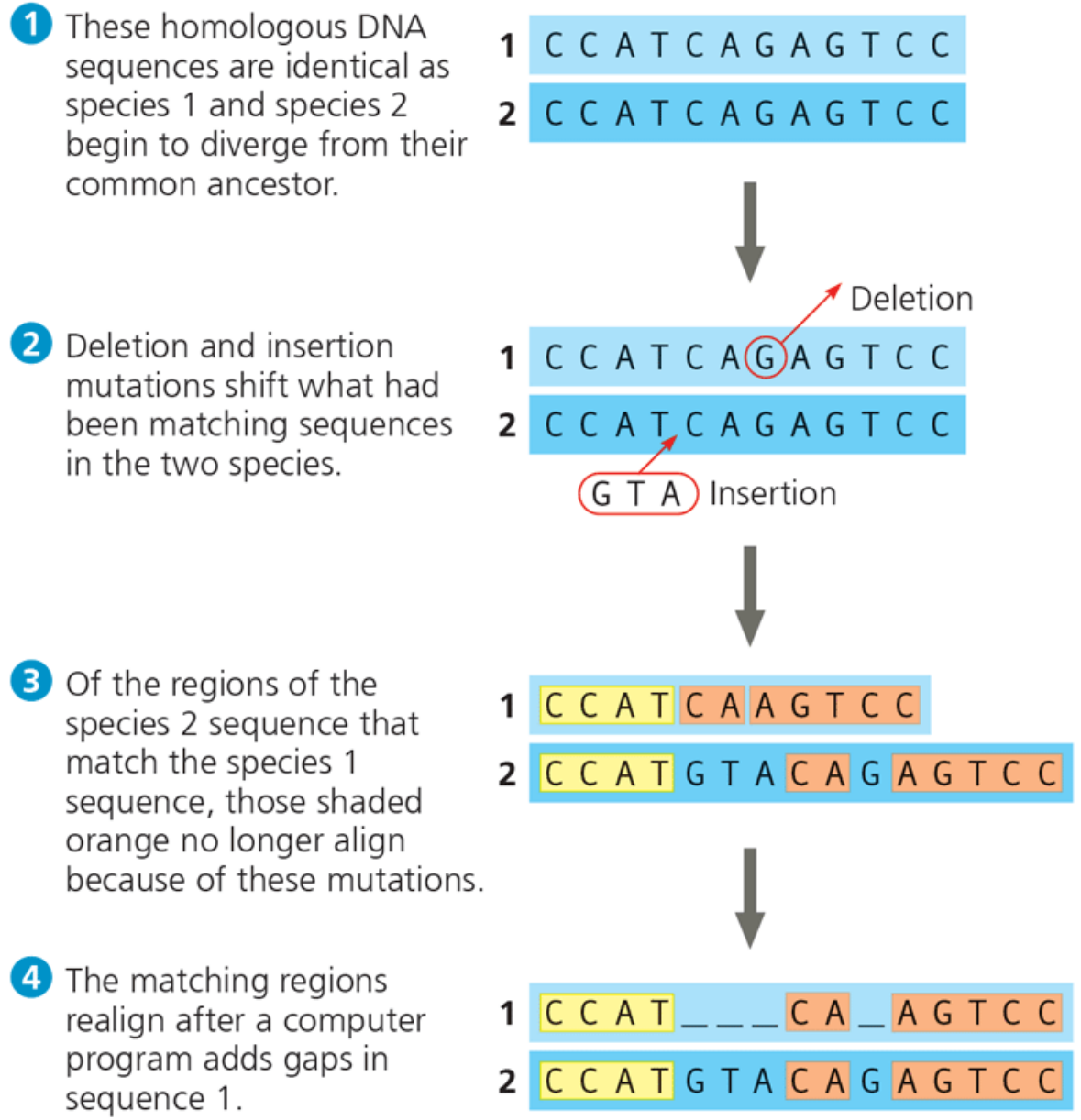
(b) Phylogenetic tree. The character table in (a) leads us to infer the phylogenetic relationships shown in this tree.

→ Usually not as clean cut as this

- Formal criterion: Maximum **Parsimony**
  - Prefer tree with least number of character changes!
  - # of possible trees goes up exponentially with # taxa

Molecular Data

- We can sequence DNA (and other molecular sequences)
- Alignment of sequences to match up homologous bases
- Take loss and gain of new base pairs into account
- In principle infer tree same way as with morphological data before…
- Models of sequence evolution and Maximum Likelihood or Bayesian Methods for more sophisticated inference than mere Parsimony



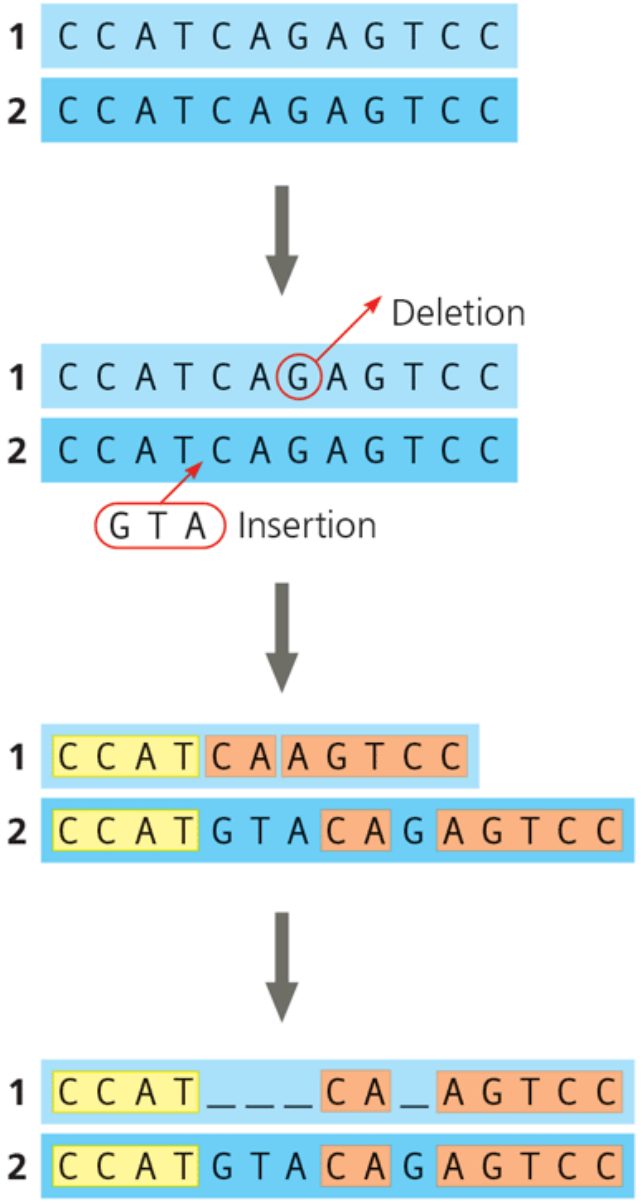
Molecular Data

Advantages of Molecular data?

- DNA is the inherited material (*i.e.* the thing that mutates and gets passed down – though selection works on traits)
- Definition of traits slightly more objective (just ACGT-sequences)
- Discrete changes in characters are easier to quantify
- Knowledge on mechanisms of mutation informs models of sequence evolution (and leads to more accurate trees)
- HUGE amount of data → each base pair is a character!

→Has confirmed and overturned many previously assumed taxonomic relationships!

→Still also came with its own set of problems and challenges (*e.g.* establishing homology, horizontal gene transfer, …)

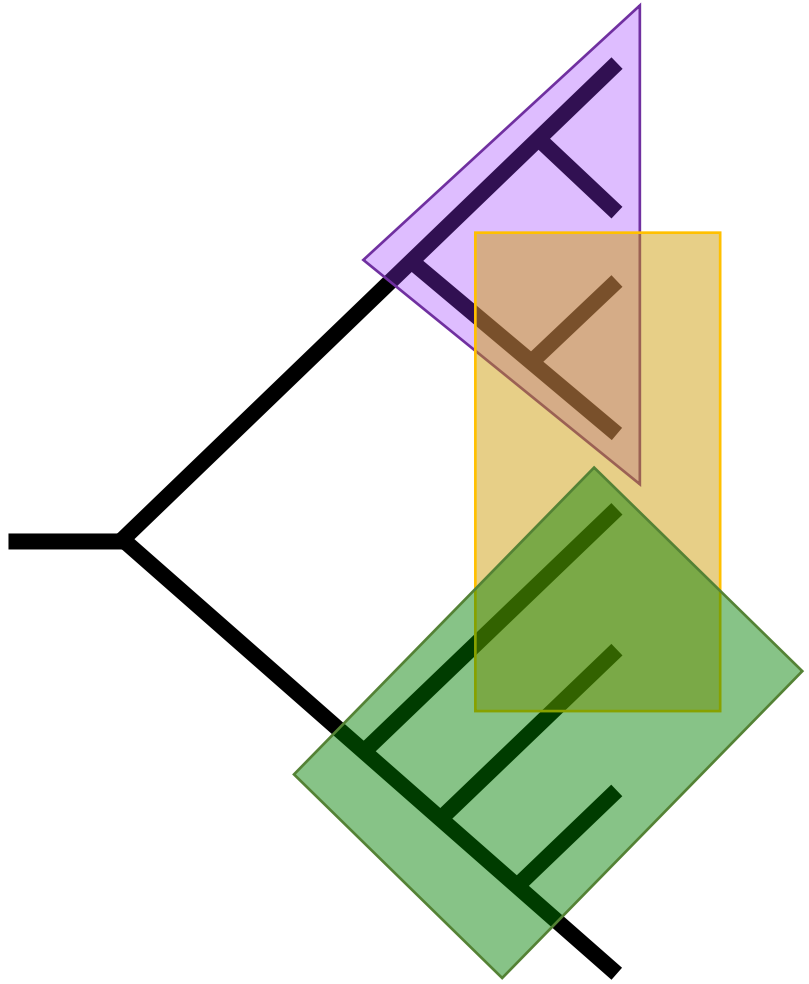


4 – Trees in R and General Checks

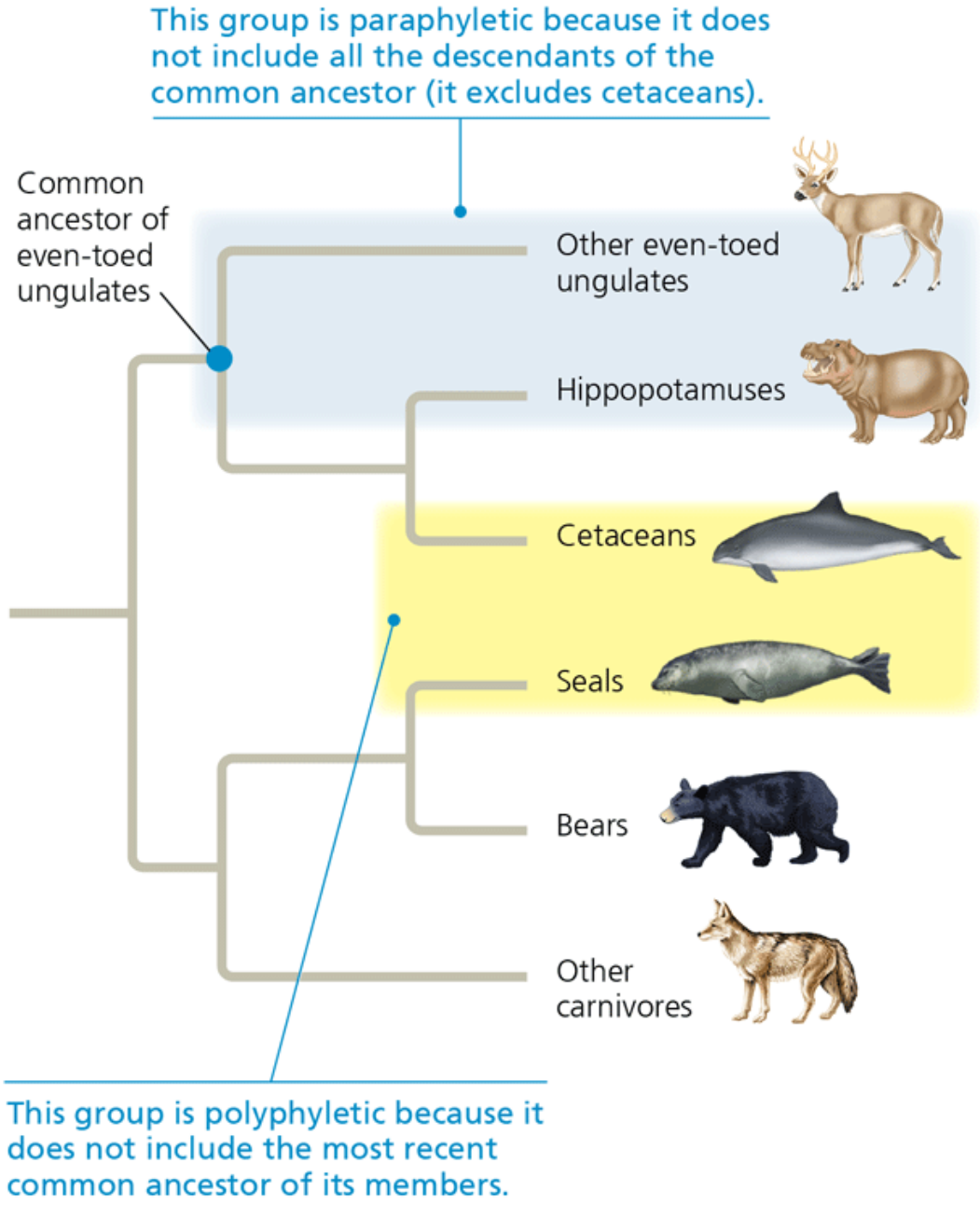
Evolutionarily Meaningful Taxa

- **Monophyletic Group (Clade):**
  Includes single ancestor and all its descendants
- **Polyphyletic Group:**
  Does not include common ancestor / has multiple origins
- **Paraphyletic Group:**
  Does not include all descendants

→Which kind is the most evolutionarily meaningful?

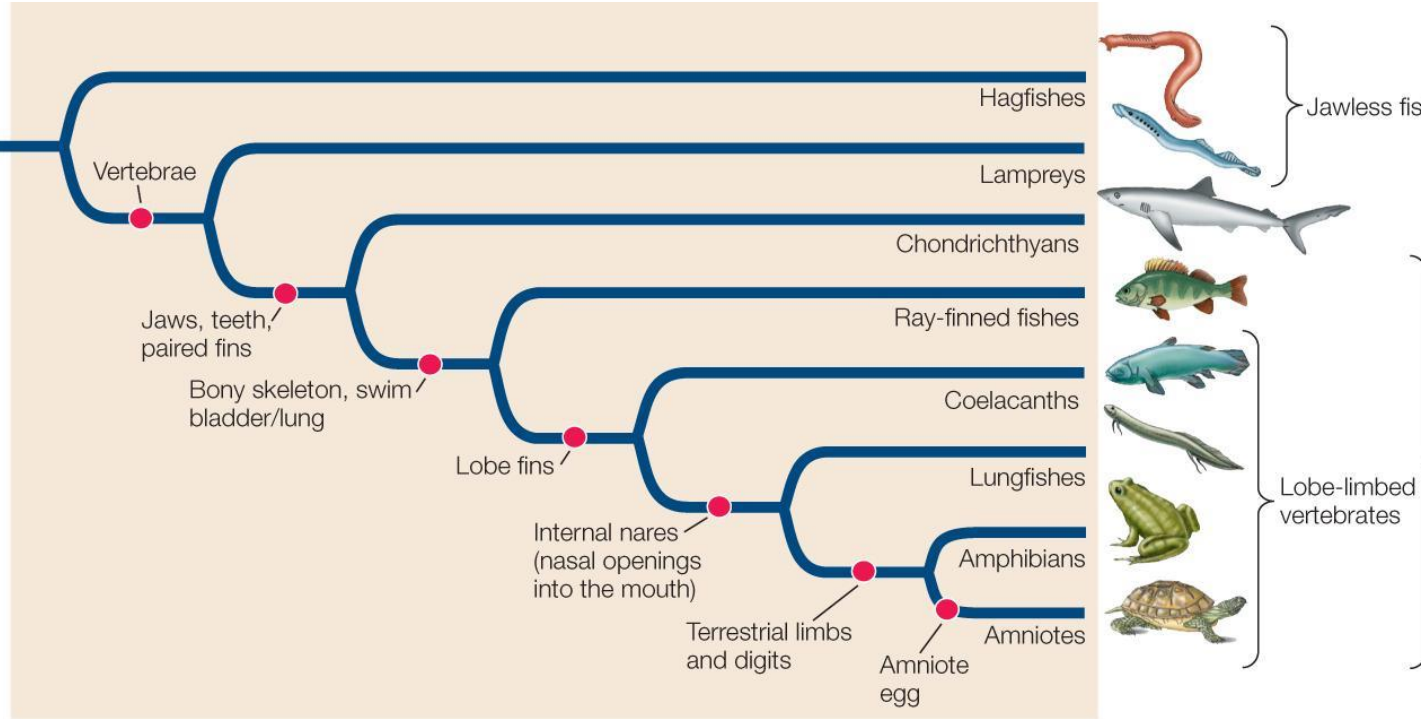


Polyphyly & Convergence

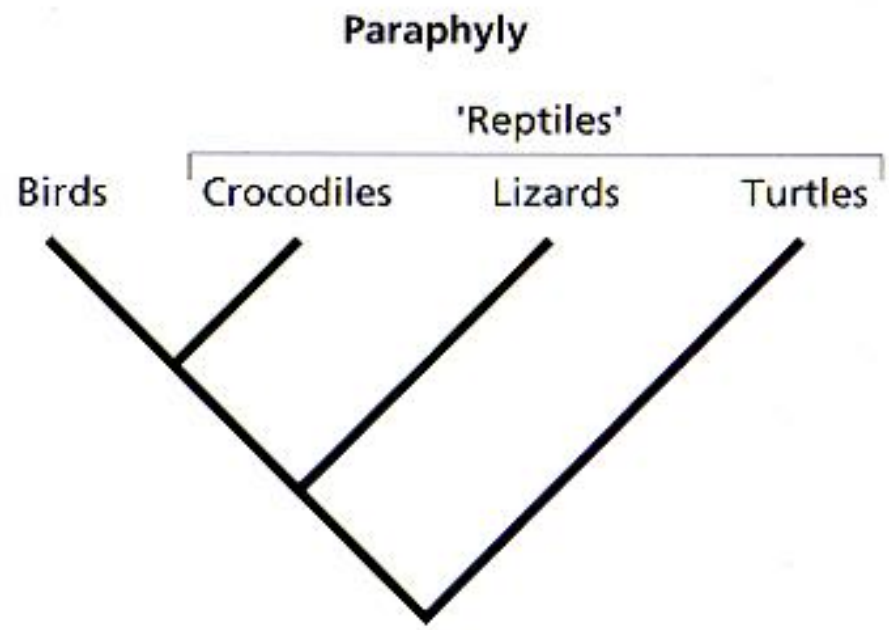


Paraphyly in Conventional Taxonomy

"There's no such thing as a fish…"



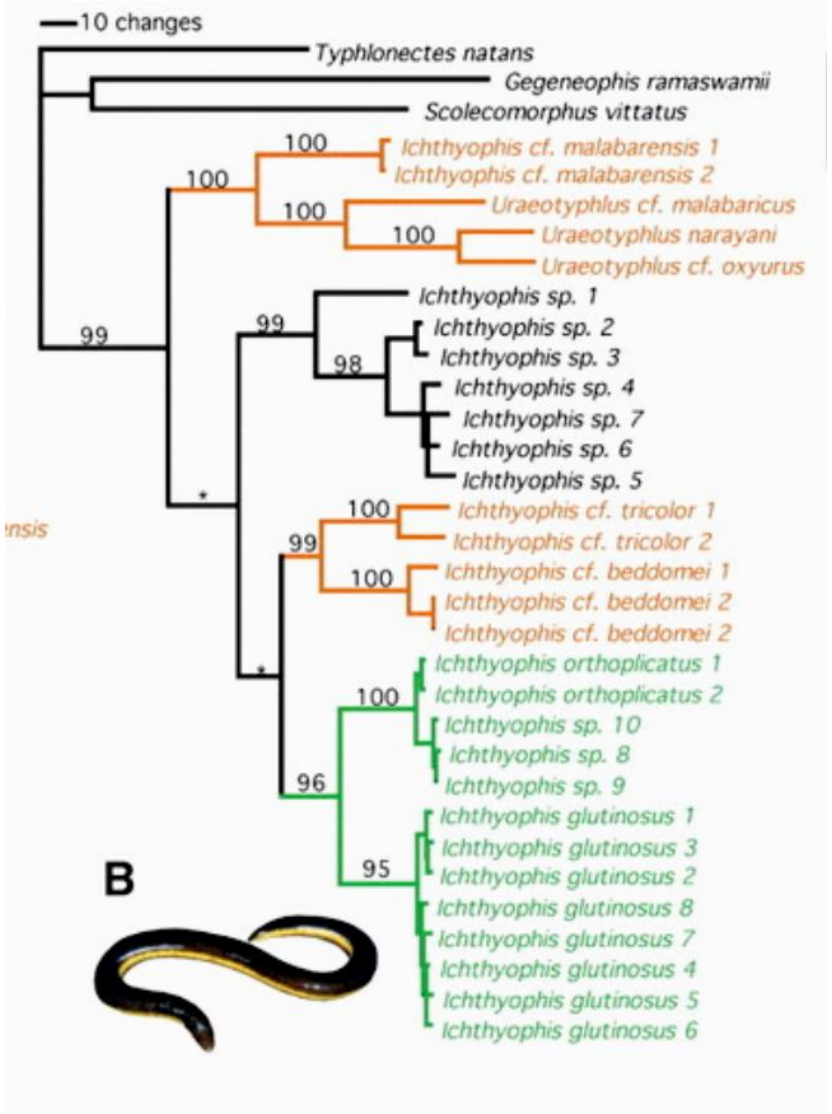
Birds are Reptiles?



Molecular Data & Branch Lengths:
Evolutionary Change

- We know the #changes (*i.e.* mutations) that separate two species
- Pairwise changes can be a measure of relatedness
- Branch lengths relative to #changes
- shows us 'how much evolution happened' since two lineages separated

→ But all tips are species alive today!

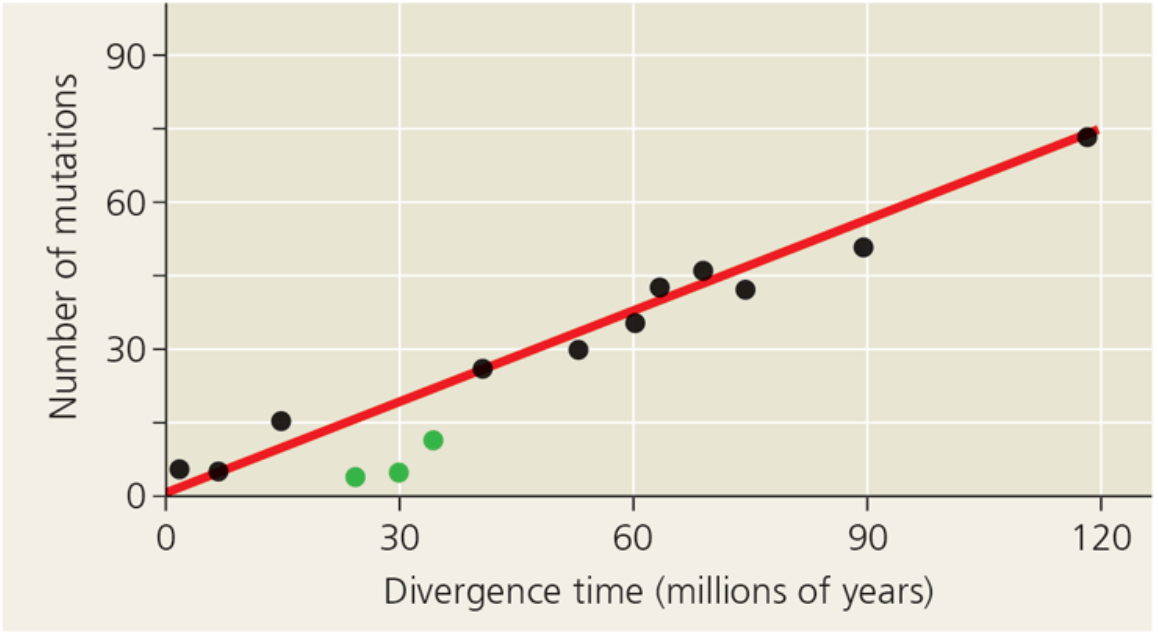


Molecular Data & Branch Lengths

Evolutionary Time Scales

→We know evolutionary distance in terms of mutations, but what about time?

- Molecular Clocks!
- In some genes/parts of the genome, mutations happen at a relatively constant rate
- Calculate rates (can vary by animal size, reproduction rates, …)
- Assumption: if rate constant, #mutations since two lineages split up is proportional to time that elapsed



Time Trees

Molecular Phylogeny (topology)

+

Number of Changes

+

Molecular Clock

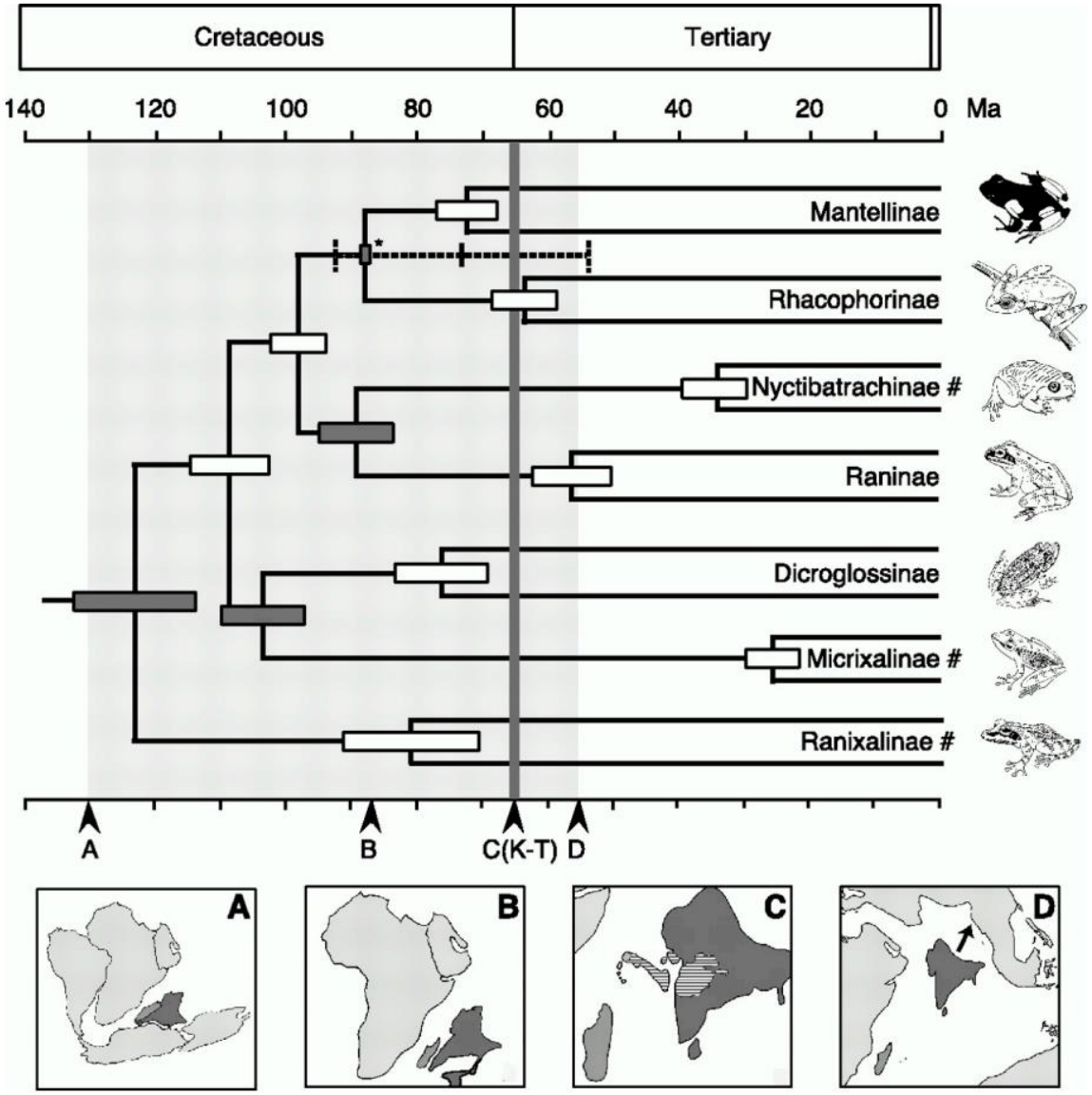
+

Fossil Calibrations

=

Time-Calibrated Phylogeny

(Time tree / ultrametric tree)



F. Bossuyt, M. C. Milinkovitch. Amphibians as indicators of early tertiary "out-of-India" dispersal of vertebrates. *Science* **292**, 93 (2001).

Tree Checking and How to Date a Tree

Some more stringent ways to date trees:

- Using a molecular clock model (with a ‘proper’ DNA-based tree)
  - possibly with partitioning etc. (jmodeltest)
- Using fossil age ranges instead of fix ages
  - Perhaps even a prior distribution in a Bayesian analysis
- Joint analysis of tree and ages
  - MCMC, e.g. in BEAST
- Tip-Dating, Fossilized-Birth-Death, Total-Evidence, …

LUNCH

5 – Methods I: Phylogenetic Signal

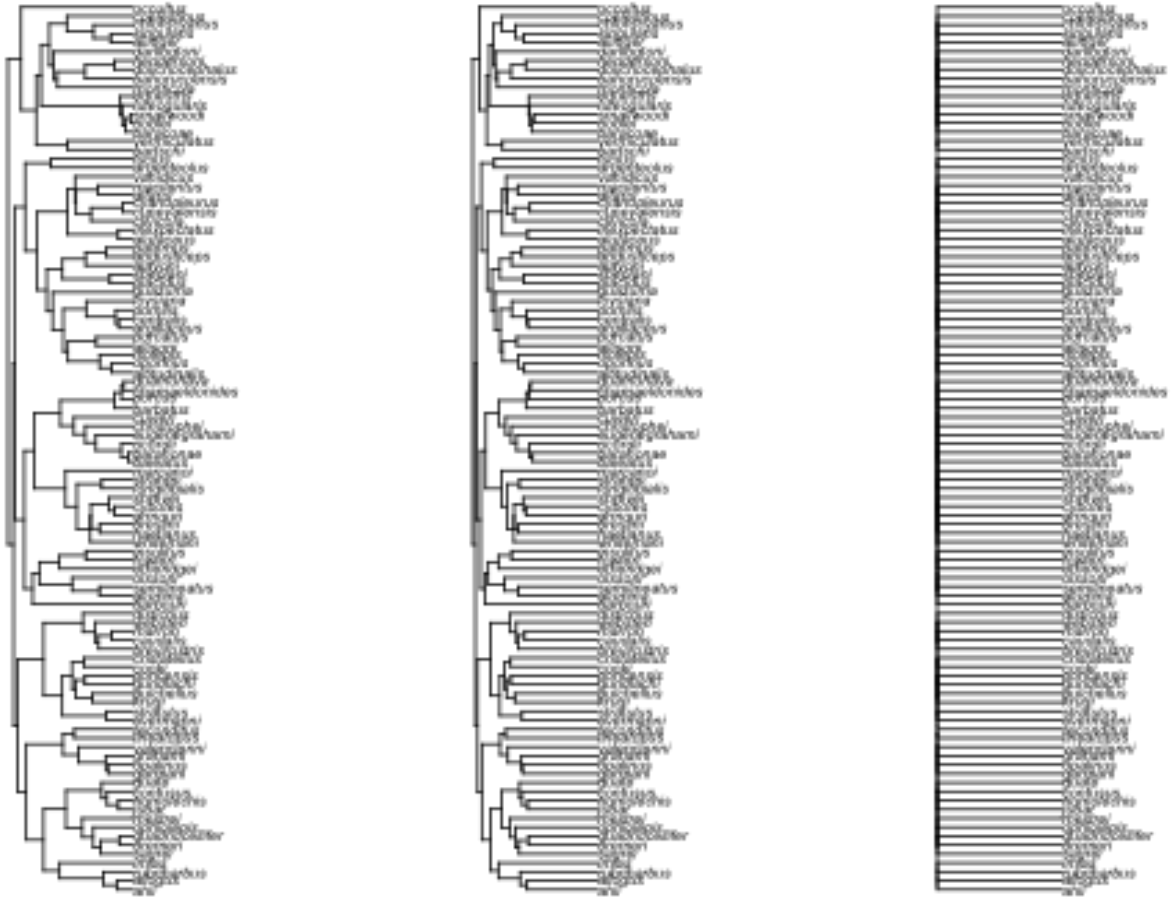
Phylogenetic Signal:

- If a trait is heritable, two related species may have started out with the same trait state/value
- Closely related species may be similar in traits only due to relatedness
- Important to take into account when e.g. investigating the ecological role of this trait!

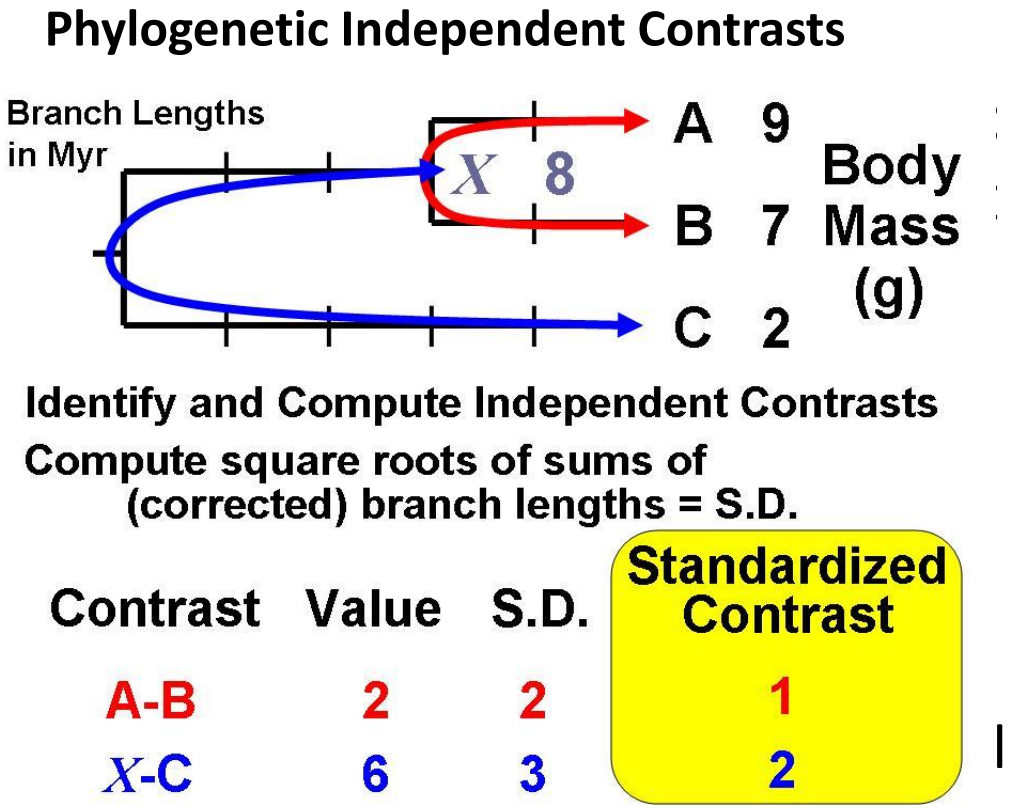
Phylogenetic Signal:

Pagel's Lambda

- Lambda (0-1) transforms tree: Length of tip branches vs. rest
- Lambda = 1:
  - Tree is unchanged
  - Trait as if evolved under BM
- Lambda = 0:
  - Tree almost full polytomy
  - No influence of phylogeny at all



Phylogenetic Signal:



Blomberg's K

- Compares variance in PIC to expectation under BM
- PIC:
  - Phylogenetic independent contrasts
  - a way to transform tip data into statistically independent values
- K < 1: less signal than BM
- K > 1: more signal than BM

Phylogenetic Signal for a Continuous Trait

Phylogenetic Signal for a Discrete Trait

6 – Methods II: Trait Evolution

Models of trait Evolution: Categorical/Discrete

- **Parsimony**:
  - Simplest, but biologically questionable
  - Reconstruction to minimize # of trait changes on tree

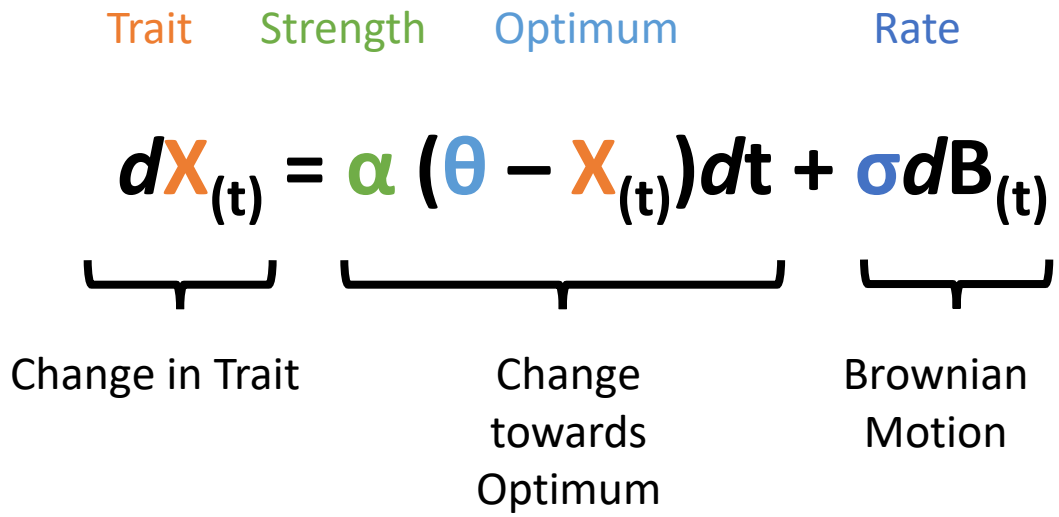
- **Mk Models (Markov k-state):**
  - Derived from DNA models (e.g. Jukes-Cantor)
  - Instantaneous rate matrix gives probabilities of state changes
  - Maximise likelihood of rates given trait
  - Equal Rates, Symmetrical, All Rates Different, …

	0	1	2	3
0	-	q	q	q
1	q	-	q	q
2	q	q	-	q
3	q	q	q	-

Models of trait Evolution: Continuous

- Brownian Motion (BM)
  - Named after molecular motion
  - Trait state ‘wiggles’ up and down randomly along branches
  - *Variants*: Simple, relaxed, multivariate, state-dependent, …

- Ornstein-Uhlenbeck (OU)
  - Evolution towards ‘trait optimum’
  - *Variants*: Simple, relaxed



Models of Trait Evolution