

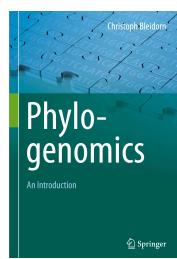
Molecular phylogenetics

Niklas Wahlberg, Jadranka Rota, Leidys Murillo, Hamid Ghanavi

Molecular Phylogenetics

- Lectures in the mornings: Niklas Wahlberg and Jadranka Rota
- Practicals in the afternoon: led by Hamid Ghanavi and Leidys Murillo-Ramos
- https://github.com/leimurillo/Mol_Phyl_2021

- Useful eBook:



Introduction to the course

- Aims of course
- Overview - why molecular phylogenetics?
- Some basic concepts e.g. phylogeny, monophyly, homology & analogy
- Exploring patterns in sequence data

Aims of the course

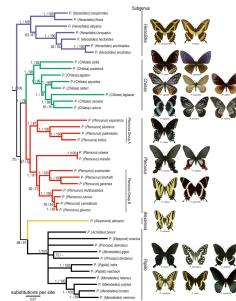
- To introduce the theory of phylogenetic inference from molecular data
- To provide an introduction to some of the most used methods (and computer programs)

The questions

- What is a phylogeny?
 - Why are we interested in phylogenies?
 - Why should we use molecular data (sequences) to infer phylogenies?

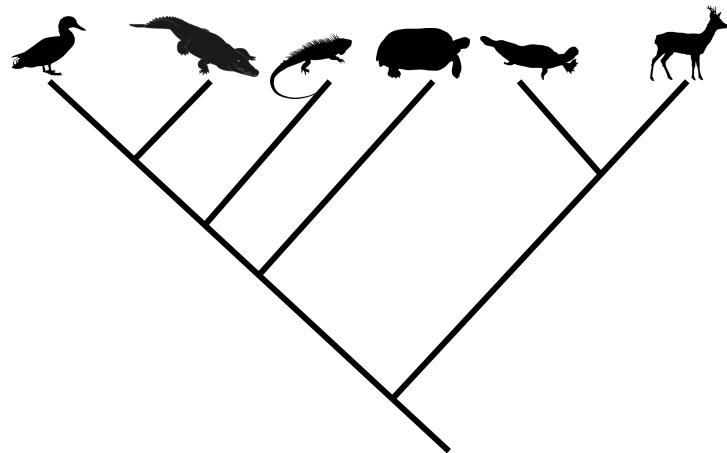
The very basic facts

- What we see today in nature is the outcome of what has happened in the past
 - Ecology and evolution are inseparable
 - “Species” or “genes” are not individual entities without any connections to other species or genes
 - phylogeny



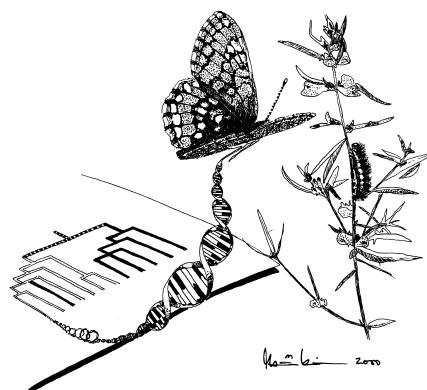
What is a phylogeny?

- A phylogeny is the historical genealogy of a group of species



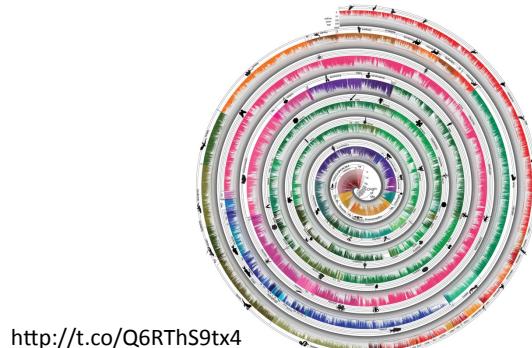
A phylogeny is an inference

- Envisioned as a dichotomously branching tree
- A phylogeny cannot be observed
- A phylogenetic hypothesis can be inferred from observed data



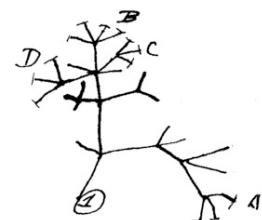
What we are after

- Phylogenies – the Tree of Life
- With phylogenies we are attempting to get a good working framework for Life
- Getting to the root of how evolution has worked



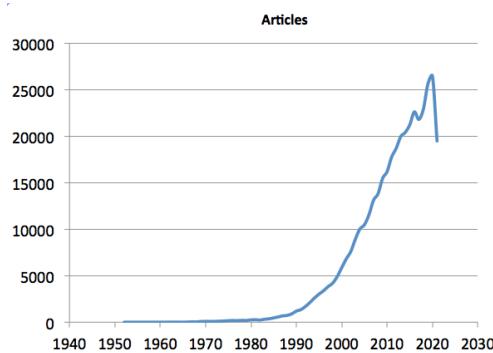
- "Nothing makes sense in biology except in the light of evolution"
- Dobzhansky 1973
- "Nothing in evolution makes sense except in the light of phylogeny"
- Savage 1997

I think



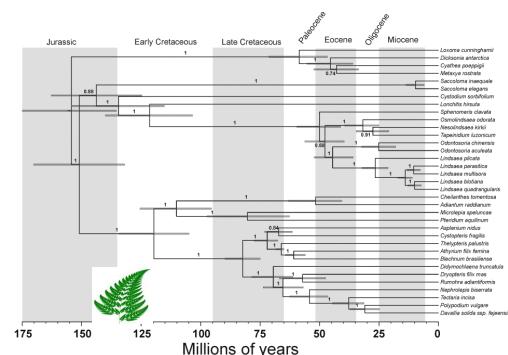
The rise of systematics

- Within the last 25 years the number of phylogenetic studies has skyrocketed
- Largely due to the advent of easy DNA sequencing methods
- Is helping us understand biodiversity and evolutionary processes better



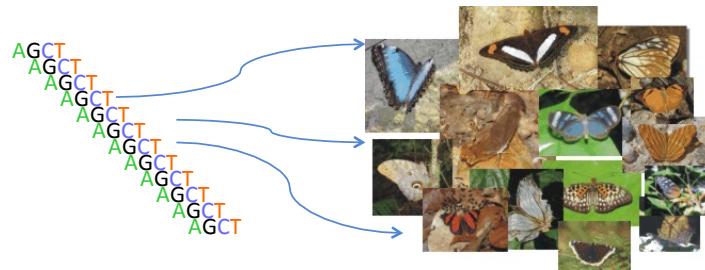
Systematics is...

- The study of the kinds and diversity of life
- The study of character evolution
- The study of historical biogeography
- The study of the temporal framework of evolution
- The study of molecular evolution



Why *molecular* systematics?

- Ease of data generation for large numbers of taxa
- Ease of generating a large number of independent data sets for given taxa
- Molecular characters behind the morphological characters we see

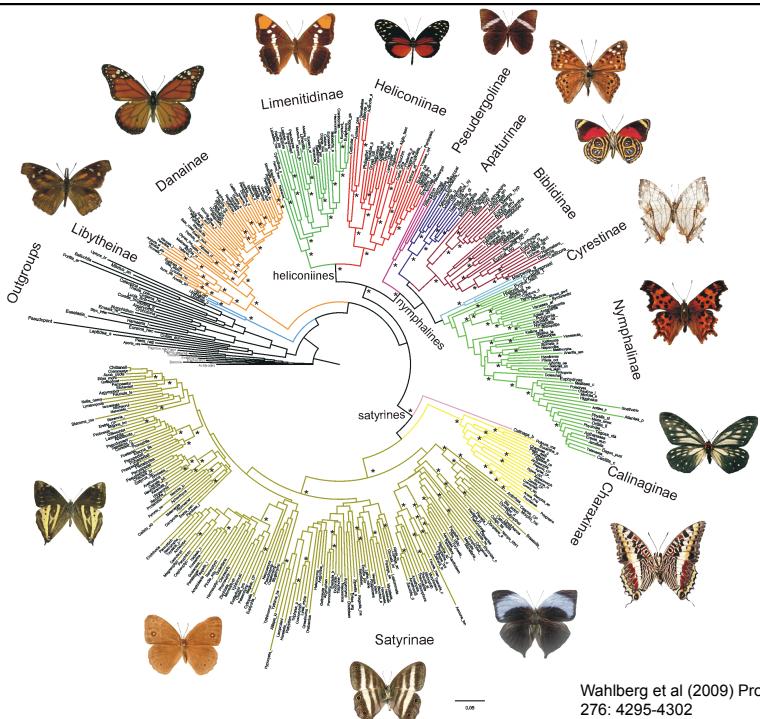
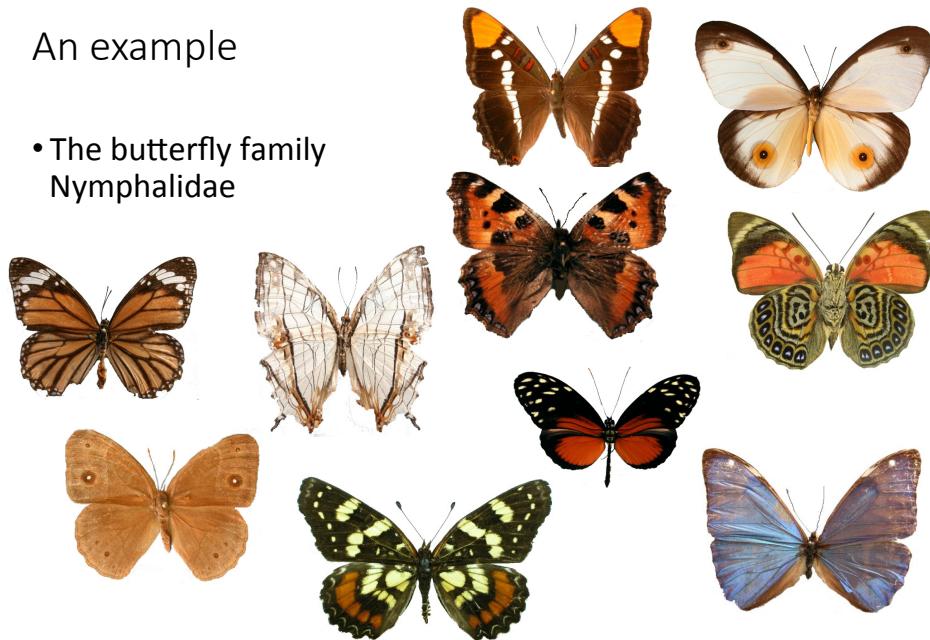


Molecular systematics as a part of understanding evolution

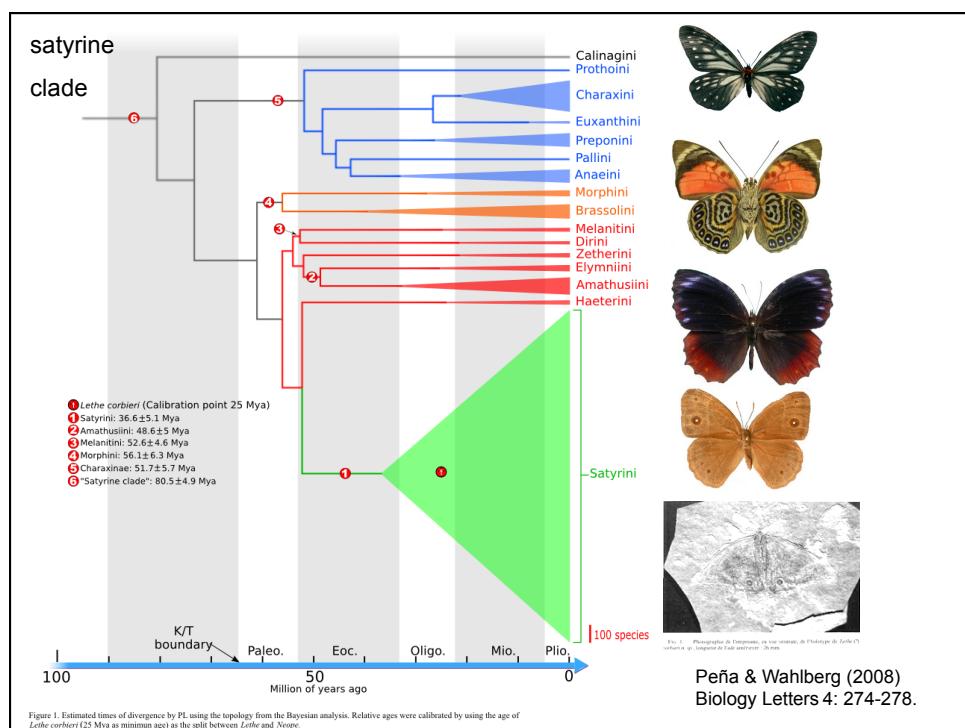
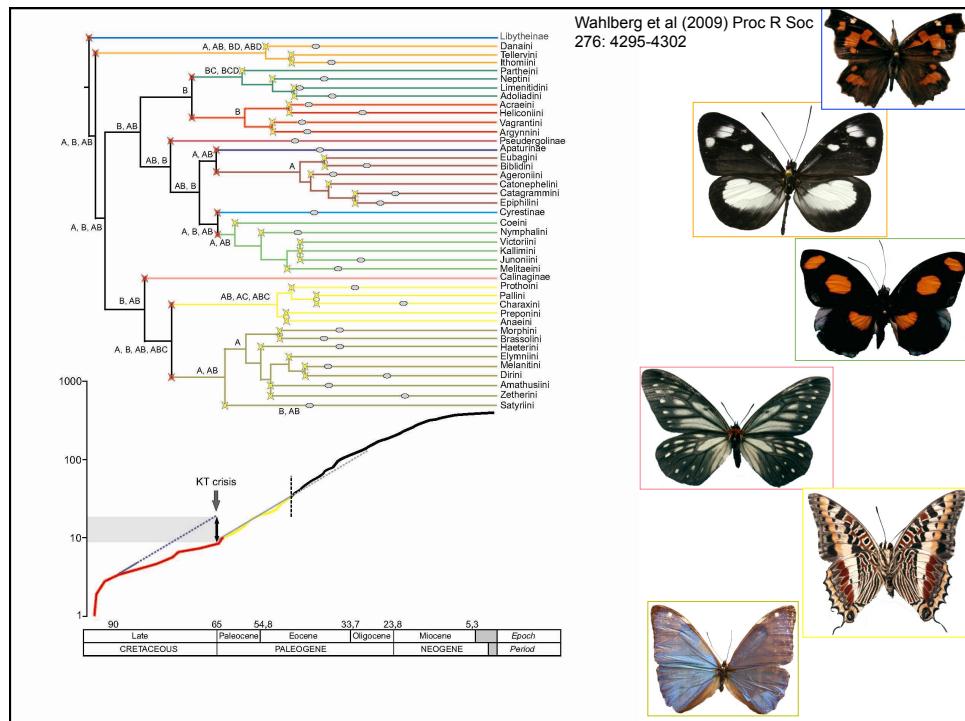
- **Biochemistry** — basic low-level processes (e.g., nucleotide substitution, amino acid interactions)
- **Molecular genetics** — fundamental genetic processes (e.g., DNA replication, recombination)
- **Population genetics** — micro-evolutionary processes
- **Systematics** — macro-evolutionary processes

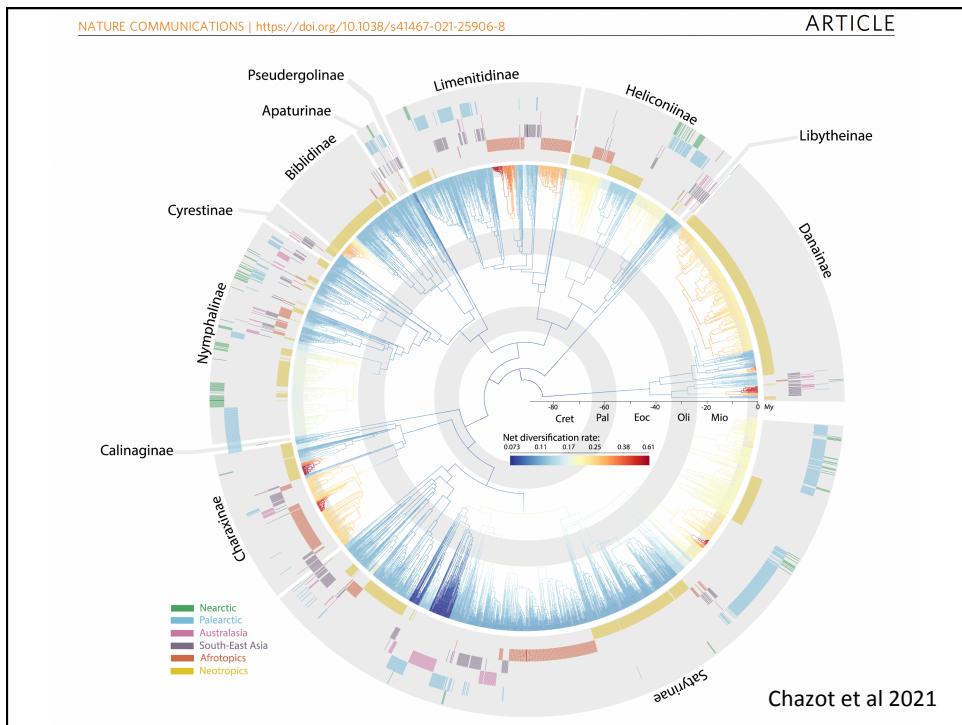
An example

- The butterfly family Nymphalidae



Wahlberg et al (2009) Proc R Soc
276: 4295-4302

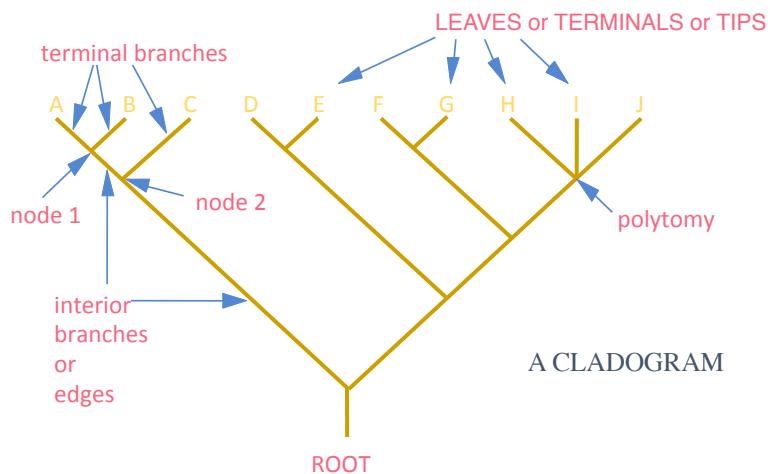




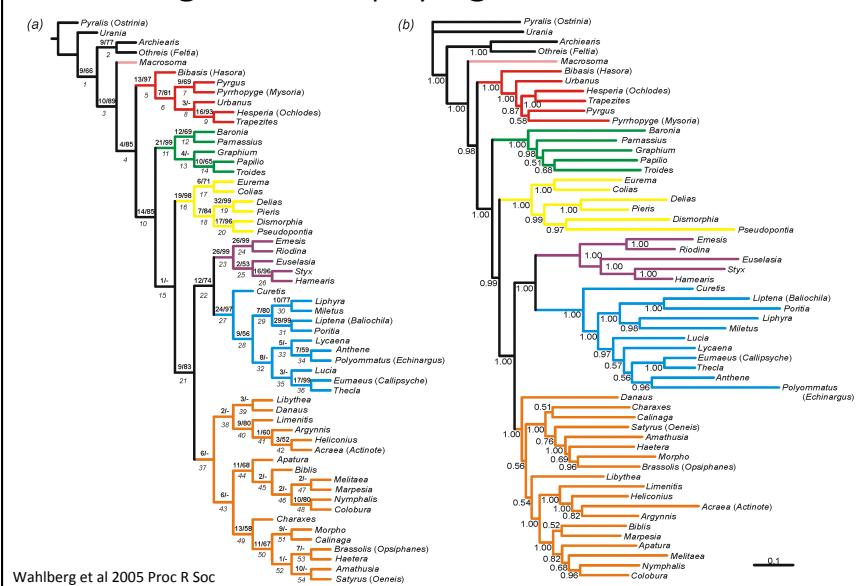
Some basic concepts

- **Cladogram** - a tree diagram which depicts a hypothesised evolutionary history
- **Phylogram** - a tree which indicates by branch length the degree of change believed to have occurred along each lineage
- **Chronogram** – a tree in which branch lengths are directly in proportion to time

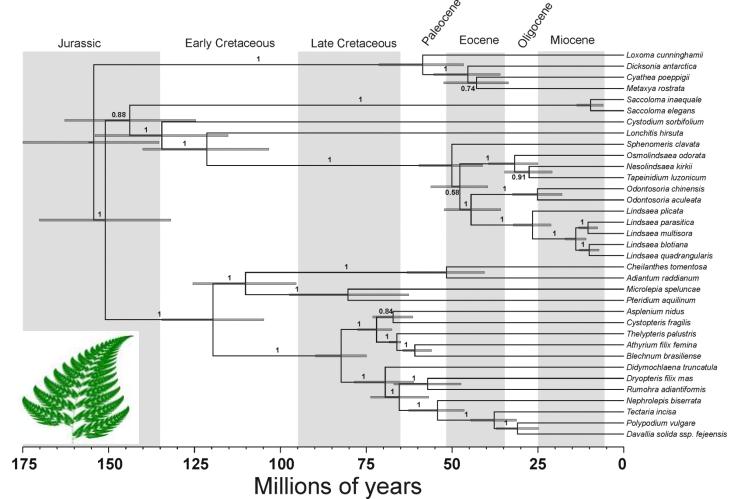
Phylogenetic Trees



Cladograms and phylogenograms

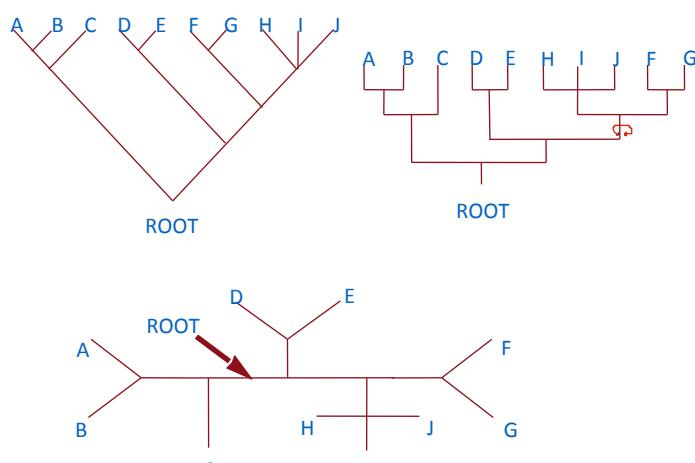


Chronogram



Lehtonen et al. 2012: Bot J Linn Soc 170:489.

Trees - Rooted and Unrooted



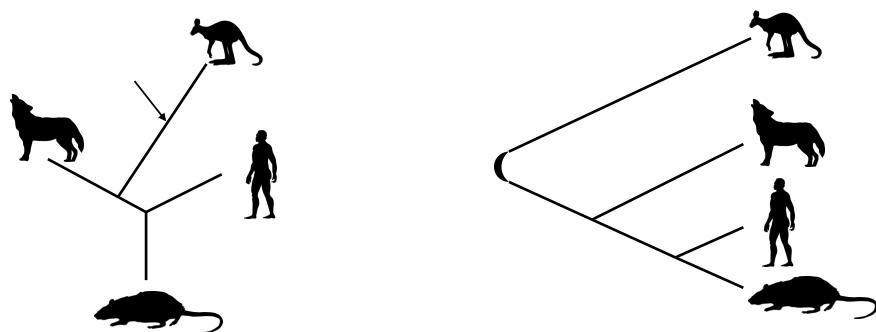
Rooting a tree

- Rooting a tree using outgroups
 - Place the root on the branch leading to the outgroup taxon
 - Use outgroup taxa in the analysis (rarely done)
- Other ways of rooting a tree
 - Assume a molecular clock
 - Midpoint rooting (root on the longest branch)



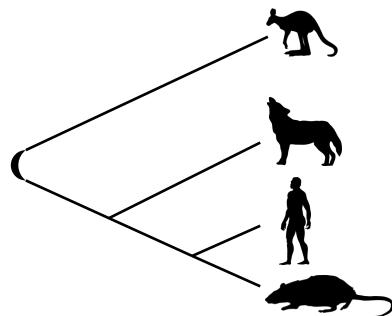
Outgroup rooting of unrooted trees

- Outgroup – related sequence that definitely diverged earlier (paleontological evidence)
- Not too distantly related (tree method becomes unreliable)

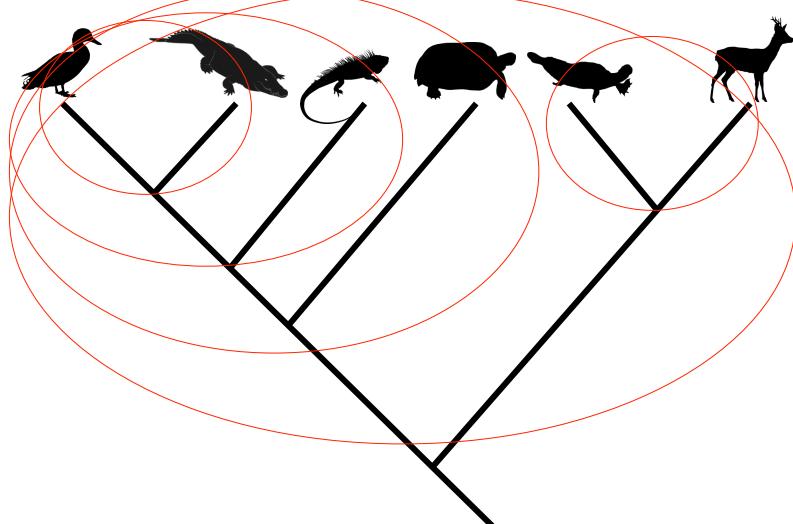


Phylogenetic systematics

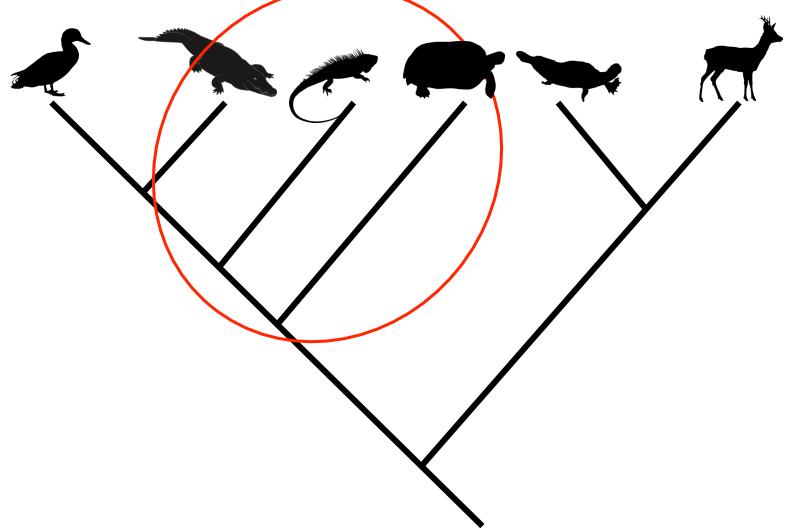
- Uses tree diagrams to portray relationships based upon recency of common ancestry
- **Monophyletic groups (clades)** - contain species which are more closely related to each other than to any outside of the group



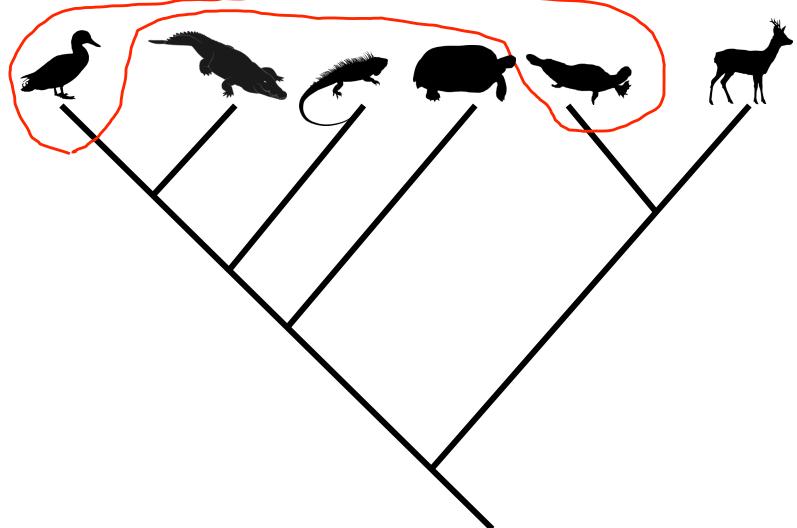
Monophyletic groups



Paraphyletic groups



Polyphyletic groups

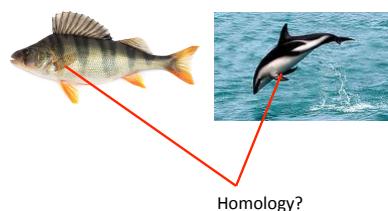


Some premises underlying phylogenetic inferences

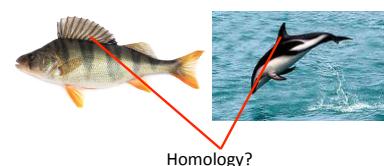
- Phylogenetic inferences are premised on the inheritance of ancestral characters, and on the existence of an evolutionary history defined by changes in these characters
- Sees homology as evidence of common ancestry
- A tree-like model of evolution
 - paralogy and lateral transfer?

Homology

- The most fundamental concept in inferring phylogeny is **homology**
- We need to be sure the characters we are studying are homologous, ie "the same" character in different organisms
- Otherwise our analyses will be misled



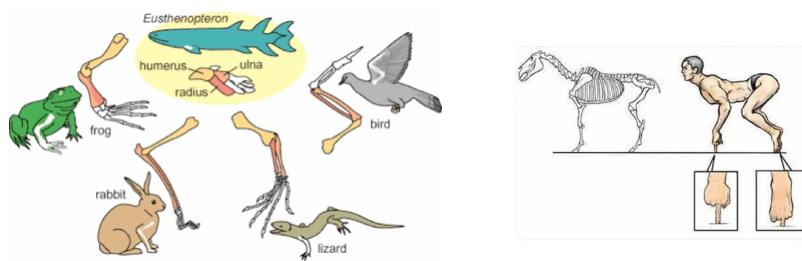
vs.



Owen's definition of homology

Homologue: the same organ under every variety of form and function
(true or essential correspondence)

Richard Owen 1843



Homologies can be:

- Apomorphic
 - Shared derived
- Plesiomorphic
 - Shared ancestral



Character evolution

- Heritable changes (in morphology, gene sequences, etc.) produce different character states
- Similarities and differences in character states provide the basis for inferring phylogeny (i.e. provide evidence of relationships)
- The utility of this evidence depends on how often the evolutionary changes that produce the different character states occur independently

Unique and unreversed characters

- Given a heritable evolutionary change that is **unique** and **unreversed** (e.g. the origin of hair) in an ancestral species, the presence of the novel character state in any taxon must be due to inheritance from the ancestor
- Similarly, absence in any taxon must be because the taxa are not descendants of that ancestor

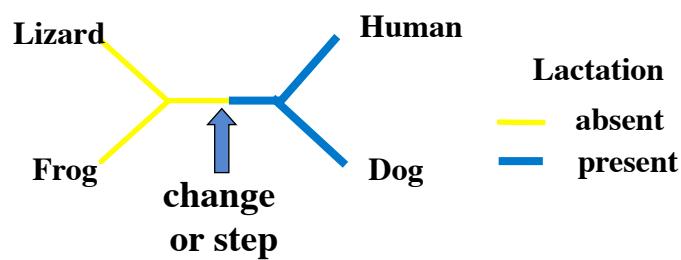
Unique and unreversed characters

- The novelty is a *homology* acting as badge or marker for the descendants of the ancestor
- The taxa with the novelty are a clade (e.g. Mammalia)

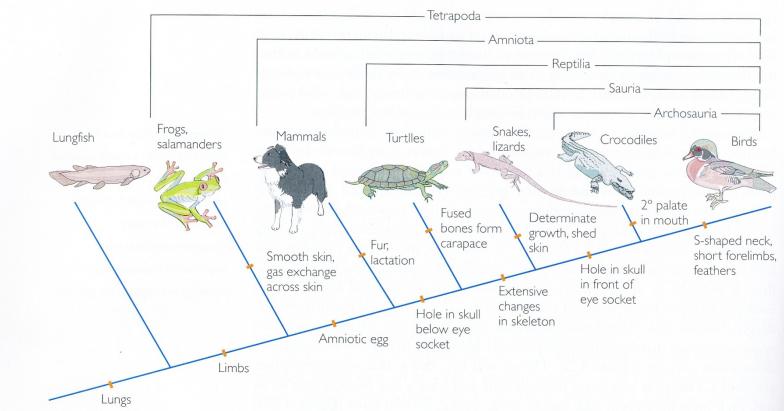


Unique and unreversed characters

- Because lactation evolved only once and is unreversed (not subsequently lost) it is *homologous* and provides unambiguous evidence of relationships

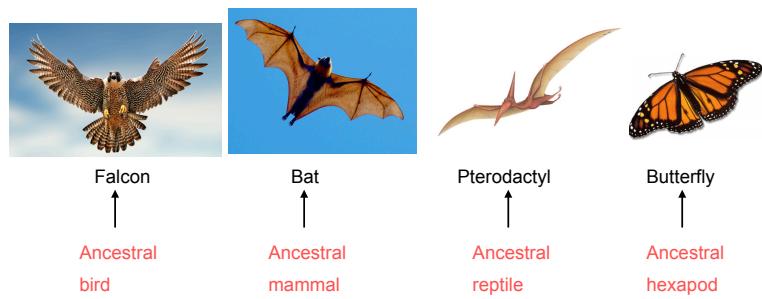


In the ideal world



The converse of homology

- **Analogy:** superficial or misleading similarity
 - Homoplasy



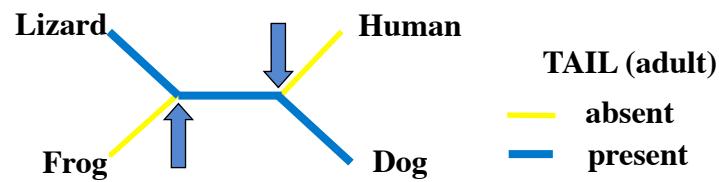
Homoplasy - Independent evolution

- Homoplasy is similarity that is not homologous (not due to common ancestry)
- It is the result of independent evolution (convergence, parallelism, reversal)
- Homoplasy can provide misleading evidence of phylogenetic relationships (if mistakenly interpreted as homology)



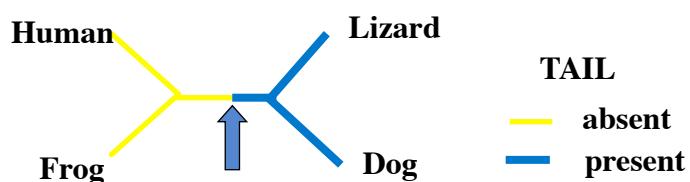
Homoplasy - Independent evolution

- Loss of tails evolved independently in humans and frogs - there are two changes on the true tree



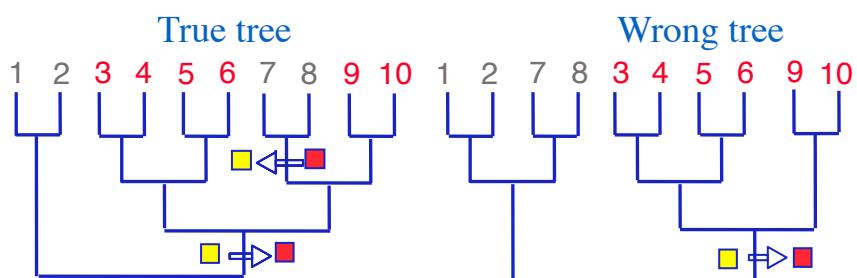
Homoplasy - misleading evidence of phylogeny

- If misinterpreted as homology, the absence of tails would be evidence for a wrong tree: grouping humans with frogs and lizards with dogs



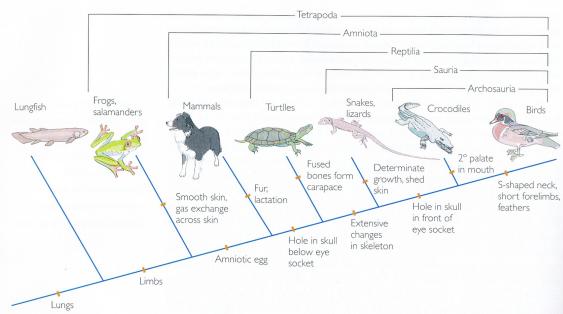
Homoplasy - reversal

- Reversals are evolutionary changes back to an ancestral condition
- As with any homoplasy, reversals can provide misleading evidence of relationships



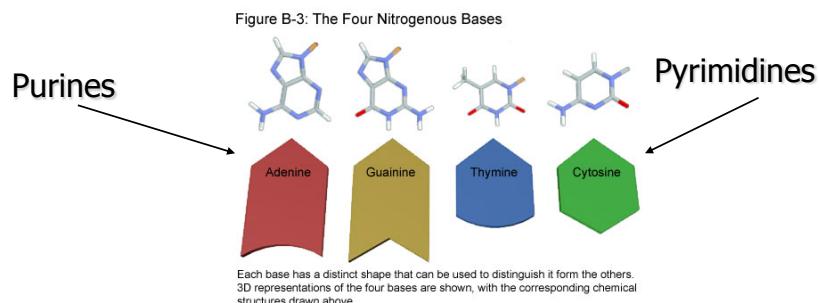
Homoplasy - a fundamental problem of phylogenetic inference

- If there were no homoplastic similarities inferring phylogeny would be easy - all the pieces of the jig-saw would fit together neatly
- Distinguishing the misleading evidence of homoplasy from the reliable evidence of homology is a fundamental problem of phylogenetic inference



Homoplasy in molecular data

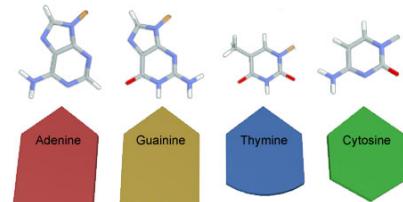
- Incongruence and therefore homoplasy is common in molecular sequence data
 - There are a limited number of alternative character states (e.g. only A, G, C, T and "gap" in DNA)



Homoplasy in molecular data

- Character states are chemically identical
- Homology and homoplasy are equally similar
- Cannot be distinguished by detailed study of similarity and differences

Figure B-3: The Four Nitrogenous Bases



Each base has a distinct shape that can be used to distinguish it from the others.
3D representations of the four bases are shown, with the corresponding chemical
structures drawn above.

Saturation in sequence data

- Saturation is due to [multiple changes](#) at the same site subsequent to lineage splitting
- Models of evolution attempt to infer the missing information through correcting for “multiple hits”
- Most data will contain some fast evolving sites which are potentially saturated (e.g. in proteins often position 3)
- In severe cases the data becomes essentially random and all information about relationships can be lost

Multiple changes at a single site - hidden changes

Ancest **GGCGCG**

Seq 1 **AGCGAG**

Seq 2 **GCGGAC**

Number of changes

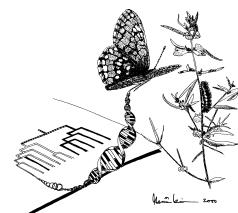
1 2 3

Seq 1 **C** → **G** → **T** → **A**

Seq 2 **C** → → → **A**
 1

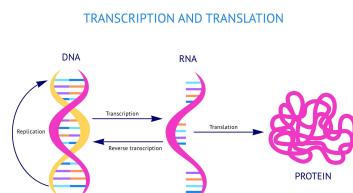
Phylogenetic analysis is an attempt to infer the past

- Inferring a phylogeny is an attempt to produce a best estimate of an evolutionary history based upon incomplete information
- Our direct information about the past is limited
 - Fossil record very incomplete
 - Access to contemporary species and molecules



Why DNA is the Ultimate Source of Information

- Higher levels lack information
 - For example, one can infer protein sequence from DNA sequence data, but not complete DNA sequence from protein sequence data
- Lower levels provide no additional useful information
 - For example, sub-atomic structure does not provide information about historical relationships



What is “Molecular” in Molecular Systematics?

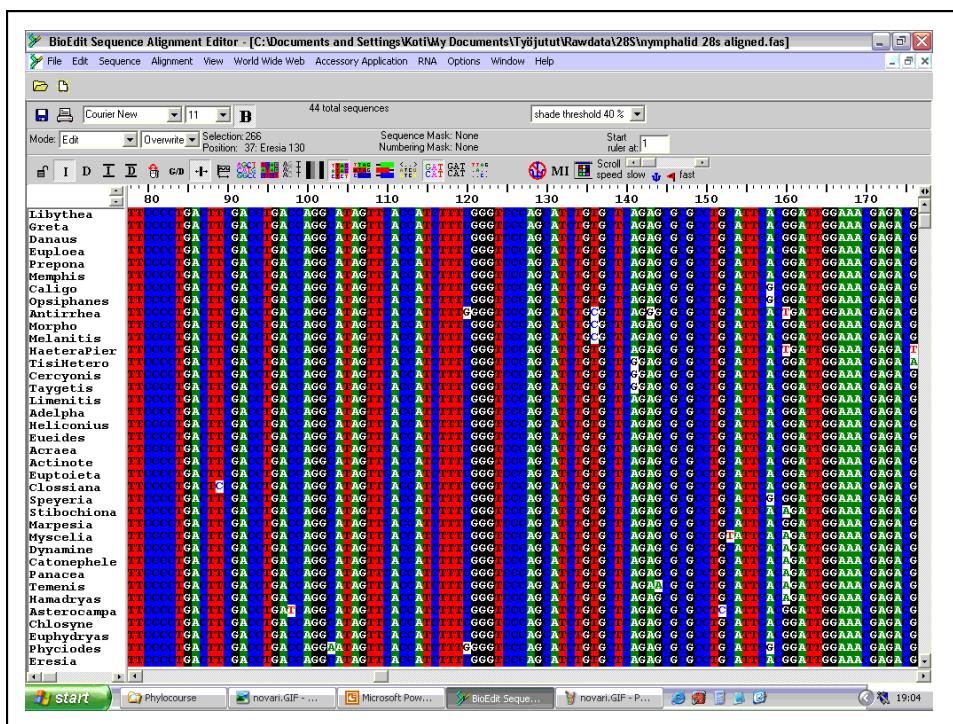
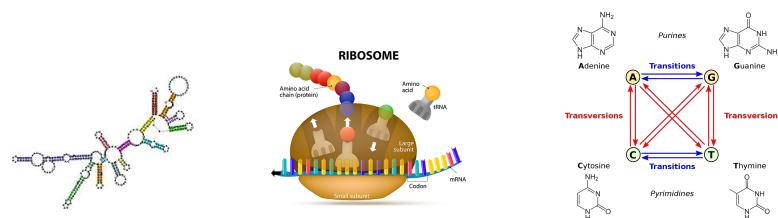
- Carbohydrates — No
- Lipids — No
- Secondary metabolites — No
- Proteins (amino acids) — Yes
- Nucleic Acids (DNA and RNA) — Yes

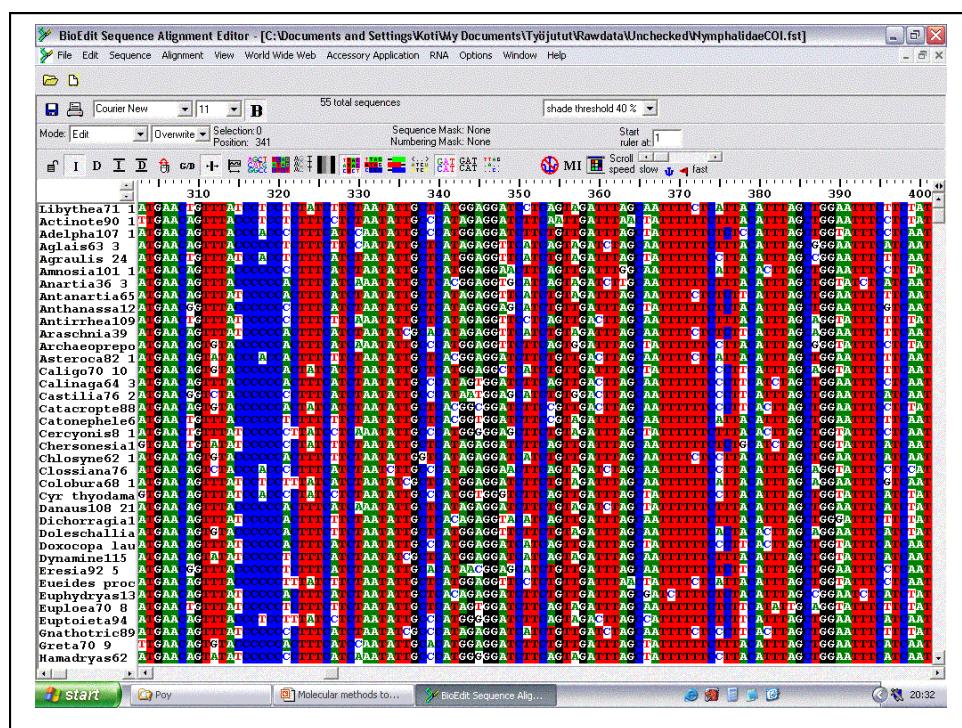
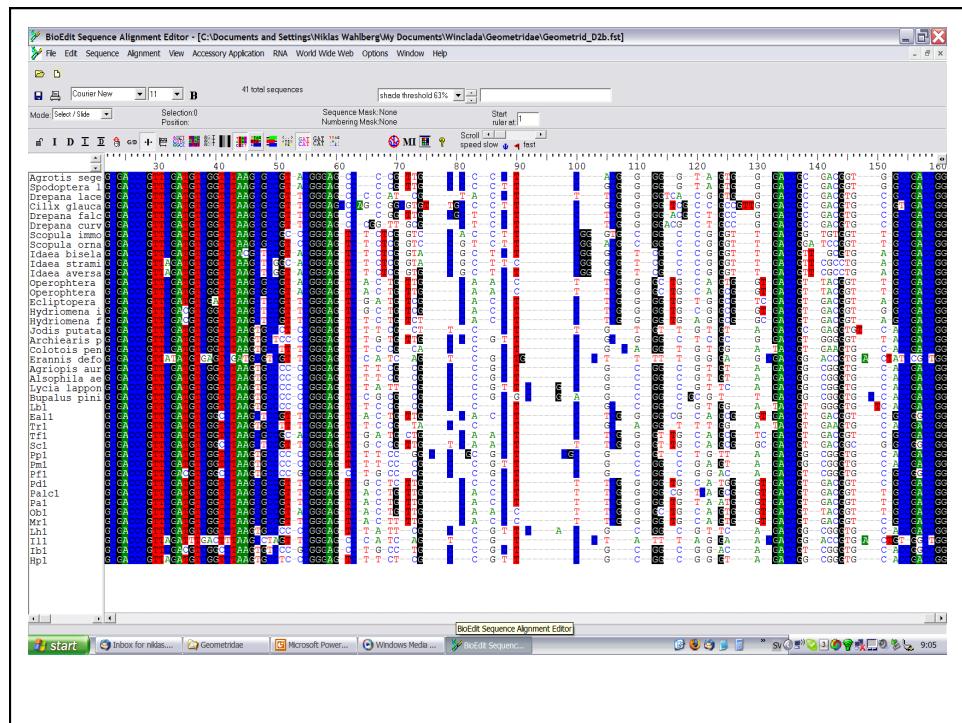


Proteins and DNA are sometimes described as
“informational macromolecules”

What sequences should we use?

- Choice of sequence - appropriate for question (fast or slow evolving - close or distant relationships).
- Many sequences are a mosaic of different rates
 - rRNA different structural regions evolve at different rates
 - Proteins - synonymous (silent) rate (codon position 3) is often faster than nonsynonymous (positions 1 & 2 - changes aa) rate of change
 - Transitions occur more readily than transversions



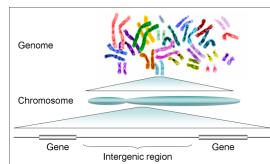


Data

- For long, the field of systematics was restricted by the amount of data
- 10 years ago, datasets comprising 3-5 genes were the norm
- 5 years ago the genomic revolution swung into full effect
- We are now faced with an abundance of potential data, but where can we get it?

The Era of Phylogenomics

- Genomes can now be sequenced relatively easily
- Whole genomes contain a lot of information that is irrelevant for systematics, especially at deep levels
- The field of systematics is still trying to figure out how best to utilize genomic level data



Phylogenomic data

- Transcriptomes
 - All genes that are being expressed in a certain tissue at a certain time
- Ultra-Conserved Elements
 - Probes to pull out UCEs and flanking regions
- Anchored Hybrid Enrichment
 - Probes for e.g. exons of protein coding sequences
- RADseq
 - Restriction site associated DNA markers
- Whole Genome Sequencing