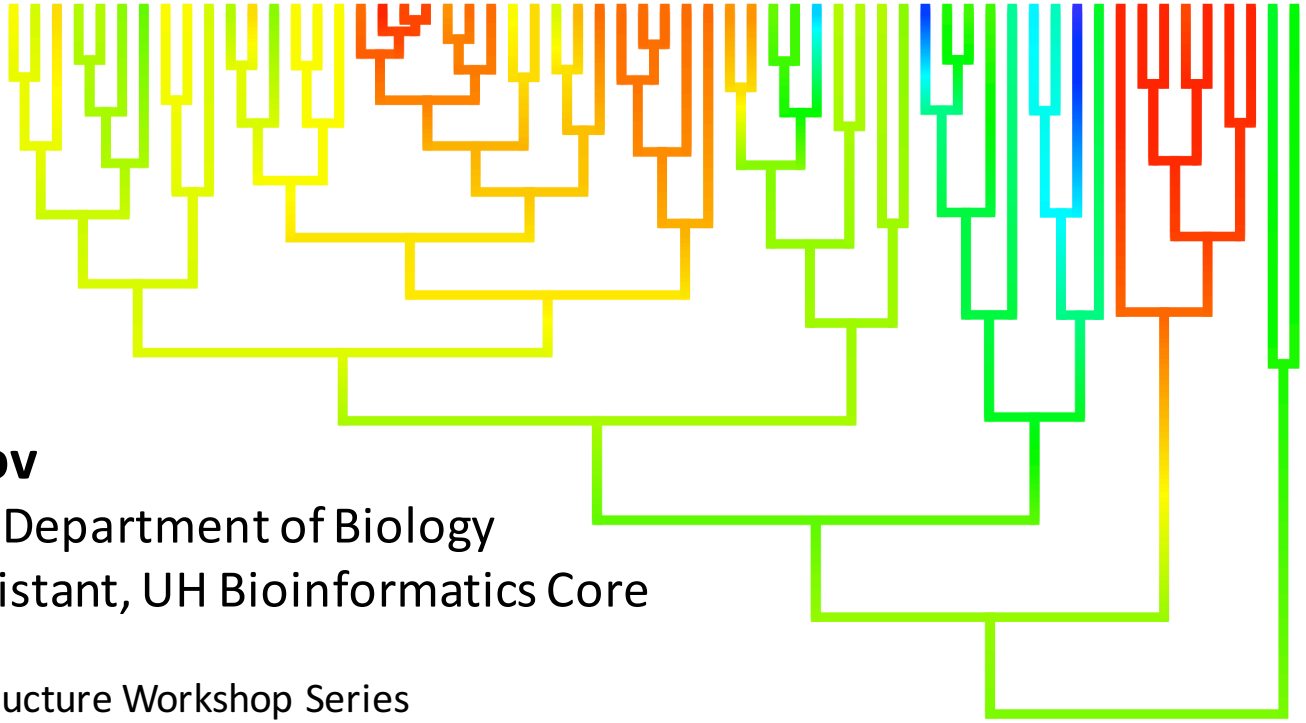
An Introduction to Molecular Phylogenetic Inference



Kirill Vinnikov
PhD Student, Department of Biology
Graduate Assistant, UH Bioinformatics Core

ITS Cyberinfrastructure Workshop Series
September 9, 2016

Goals for today

- How to read phylogenetic trees
- How to access sequence information in online databases
- How and when to use different sequence alignment methods
- How and when to use different phylogenetic methods
- How to visualize your phylogenetic trees

Workshop schedule

9:00 - 9:40 am	Introduction and Basic Concepts
9:40 - 10:20 am	Practical Lab 1
10:20 - 10:30 am	**BREAK**
10:30 - 11:10 am	Introduction to Phylogenetic Methods
11:10 - 12:00 pm	Practical Lab 2
12:00 - 12:10 pm	**BREAK**
12:10 – 1:00 pm	Practical Lab 3
1:00 pm	Q&A

1. Introduction and basic concepts

Outline

- The tree thinking challenge
- Understanding the basics:

  homology, orthology, paralogy, xenology
- Sequence alignment approaches
- Introduction to online sequence databases
- Phylogenetics vs Phylogenomics: what's the difference?

Why use phylogenies?

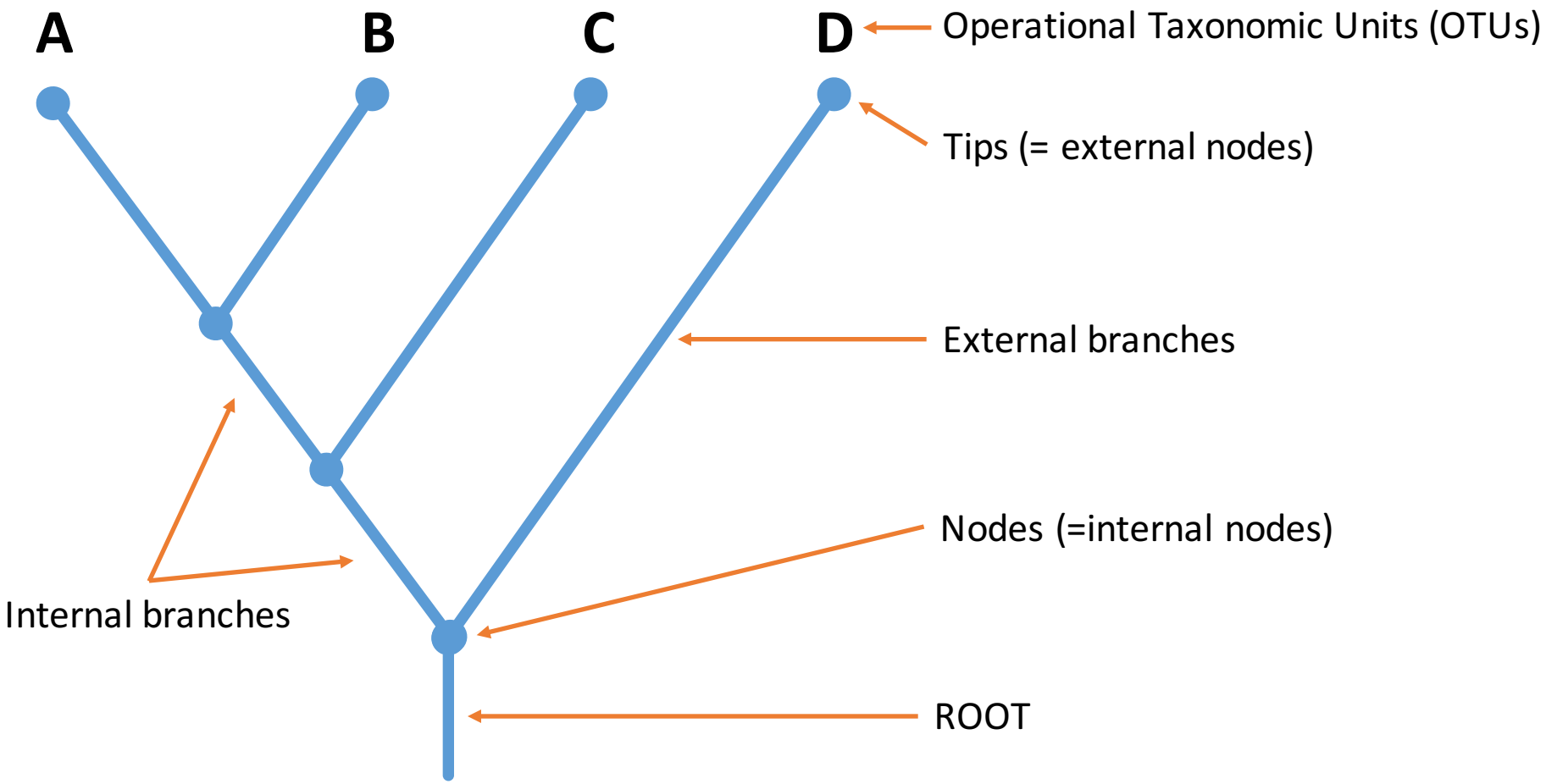
"Nothing in biology makes sense

except in the light of evolution"

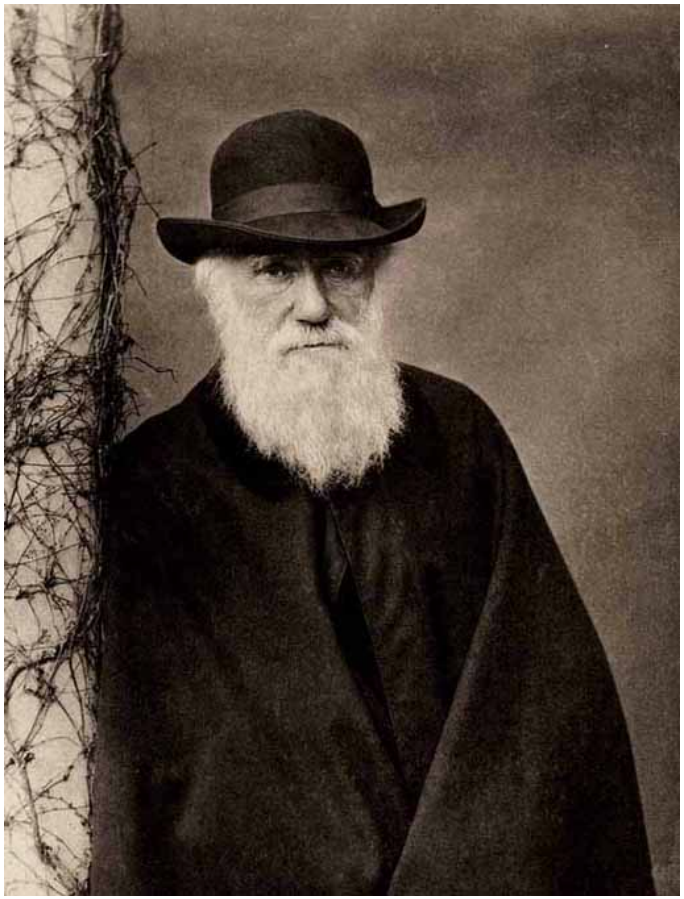
Theodosius Dobzhansky (1900-1975)

- Evolution gives us a mechanistic understanding of life
- It changed biology from a descriptive to predictive science
- Phylogenetic trees provide the easiest way to depict evolutionary relationships between organisms
- Phylogenies allow us to test different evolutionary hypotheses

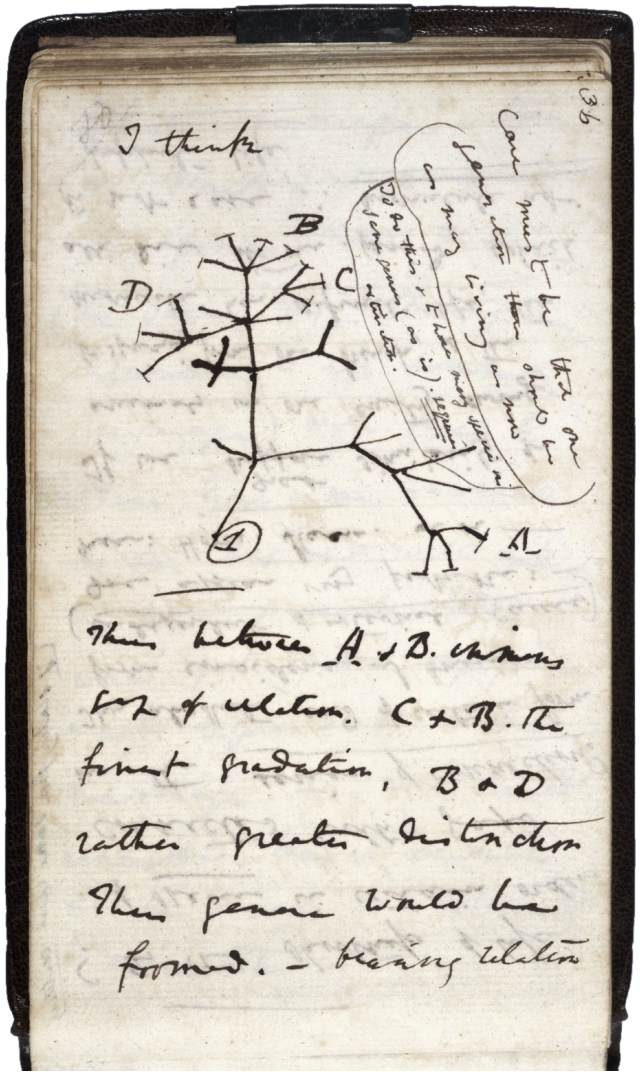
Phylogenetic Tree Terminology



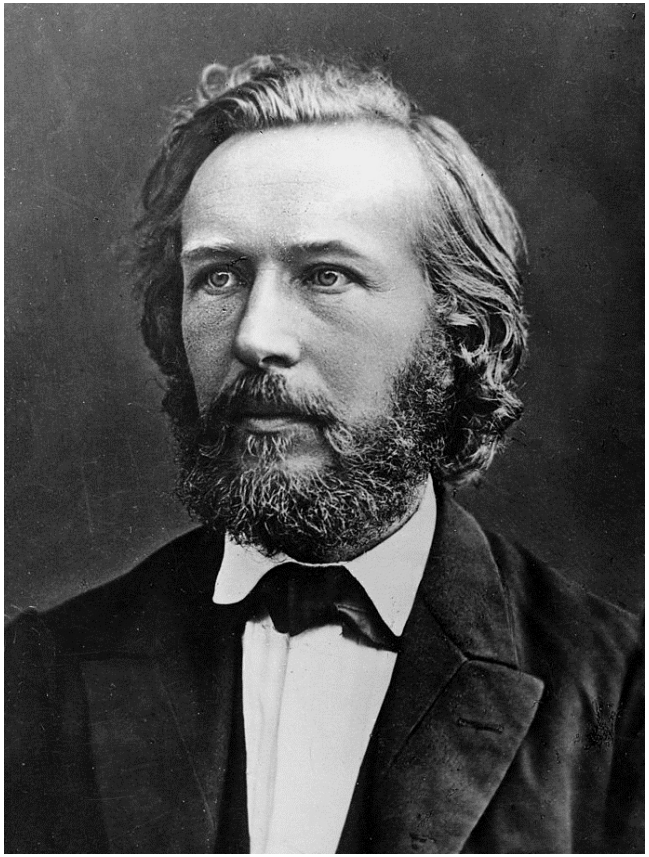
First phylogenetic trees



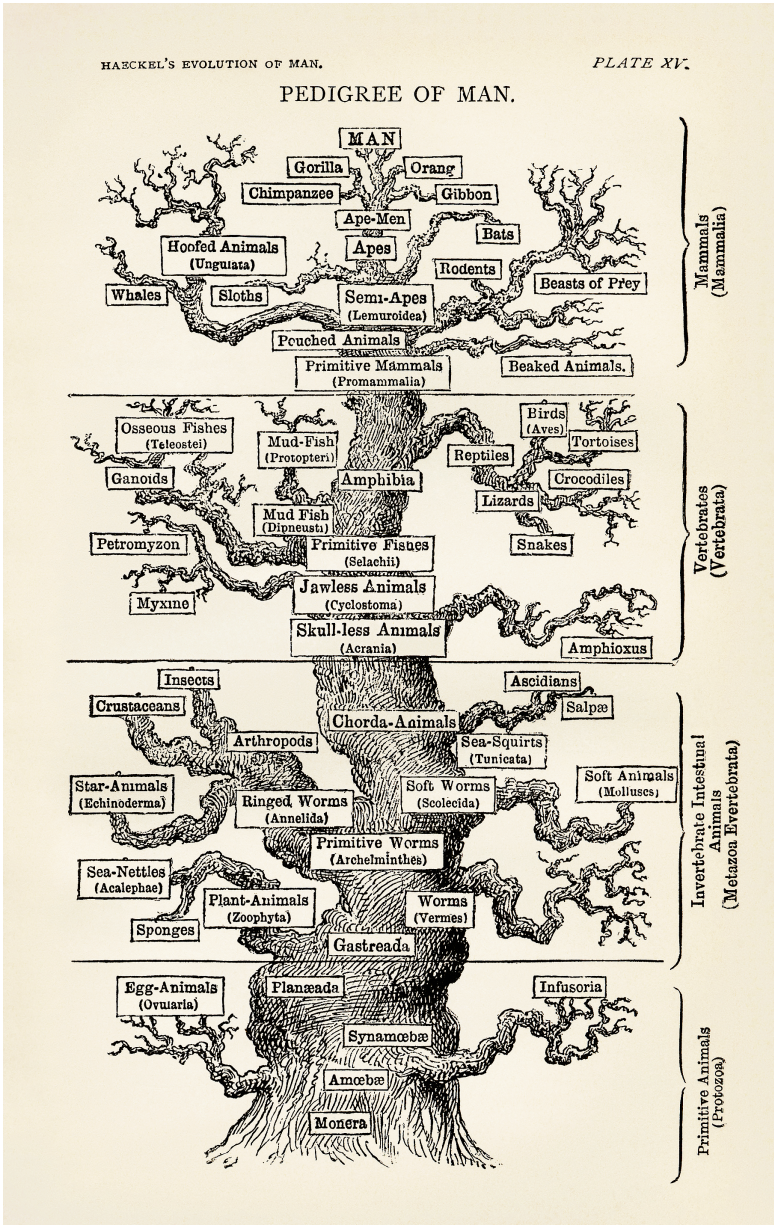
Charles Darwin (1809-1882)



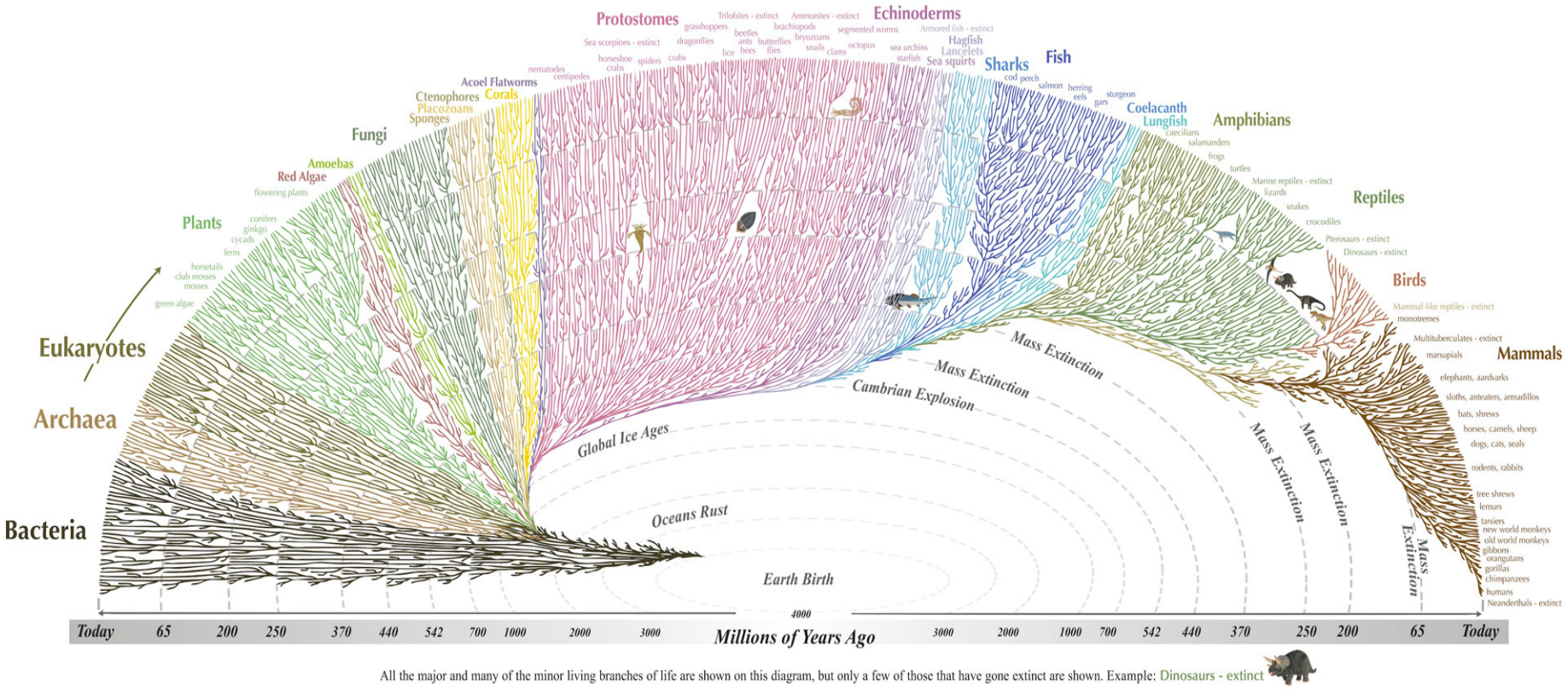
First phylogenetic trees



Ernst Haeckel (1834-1919)



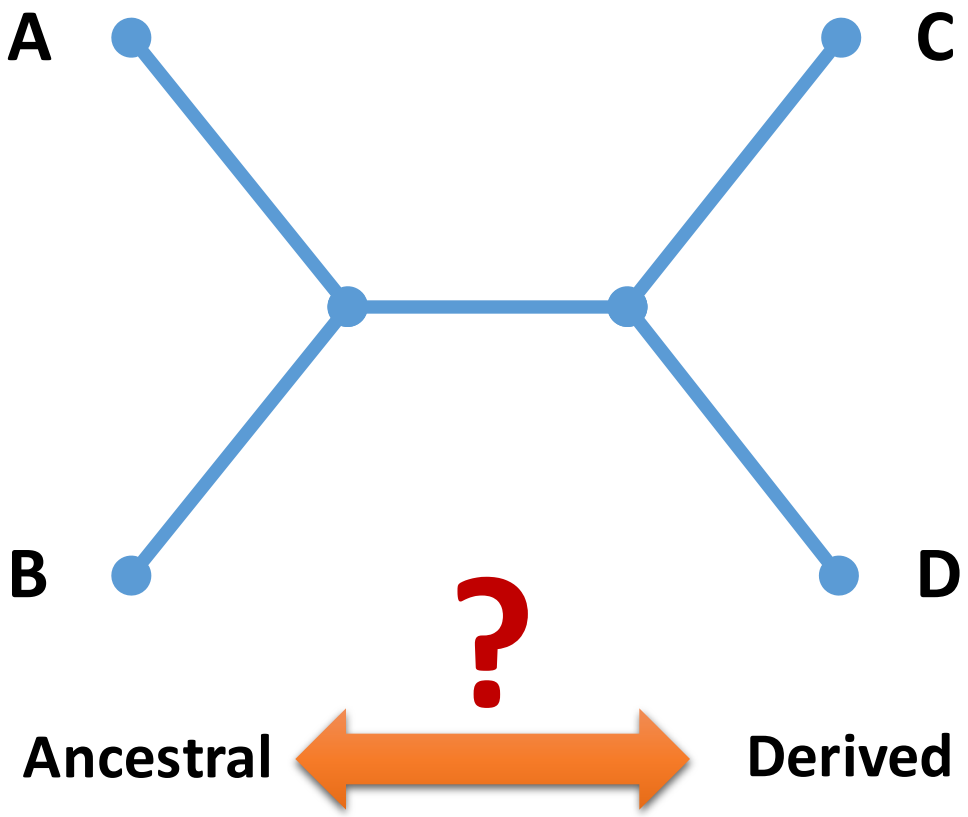
Modern phylogenetic trees: the Tree of Life



All the major and many of the minor living branches of life are shown on this diagram, but only a few of those that have gone extinct are shown. Example: Dinosaurs - extinct

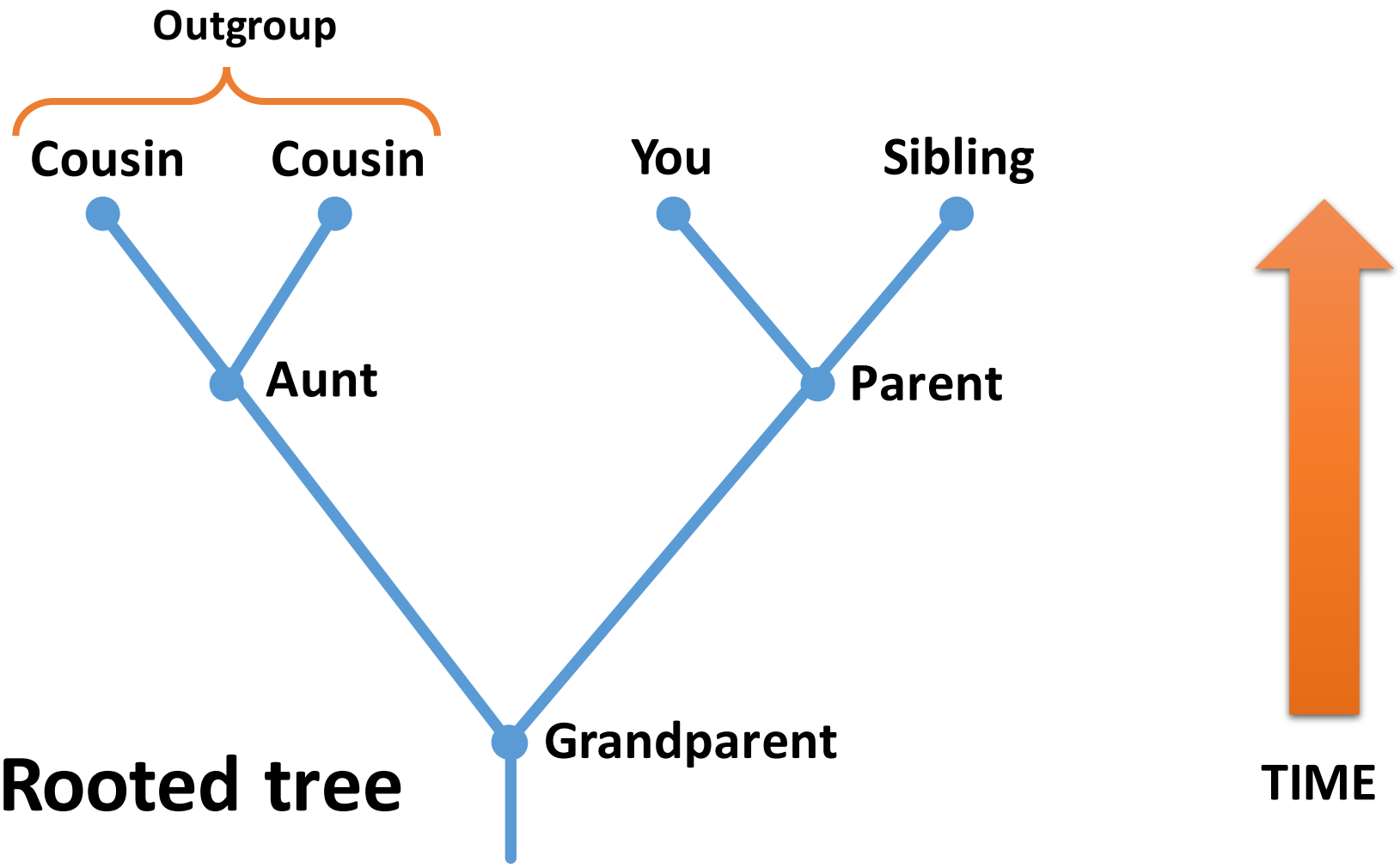
Image source: https://www.evogeneao.com

Understanding Tree Topology



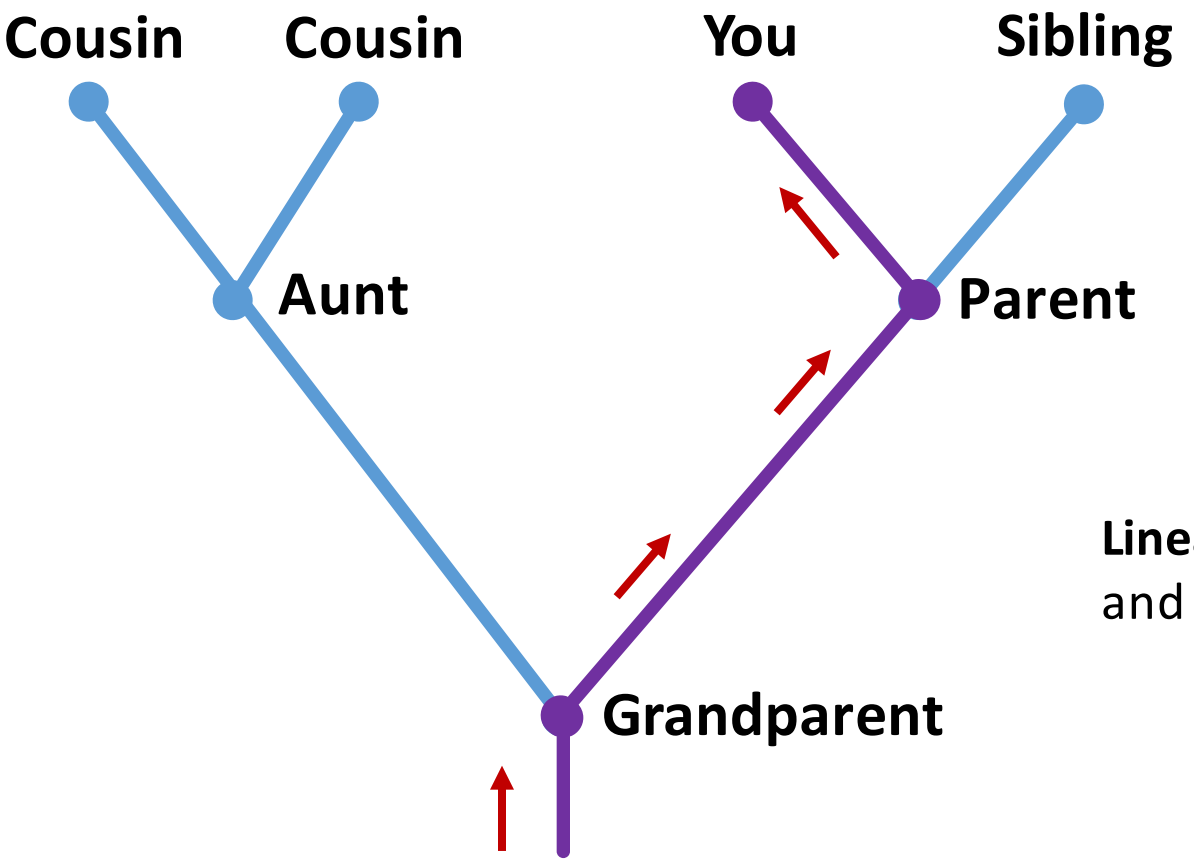
Unrooted tree

Understanding Tree Topology



Rooted tree

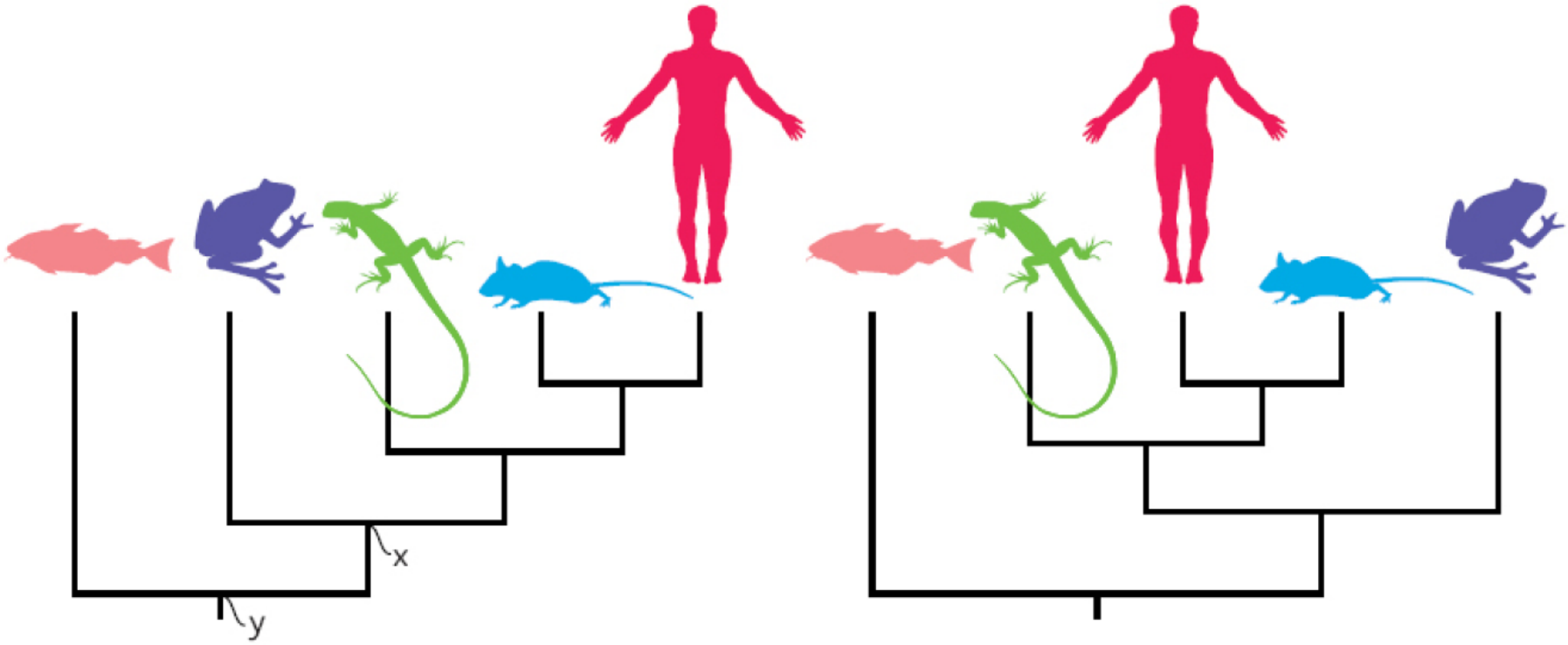
Understanding Tree Topology



Lineage – a chain of ancestors and their descendants

The tree thinking challenge

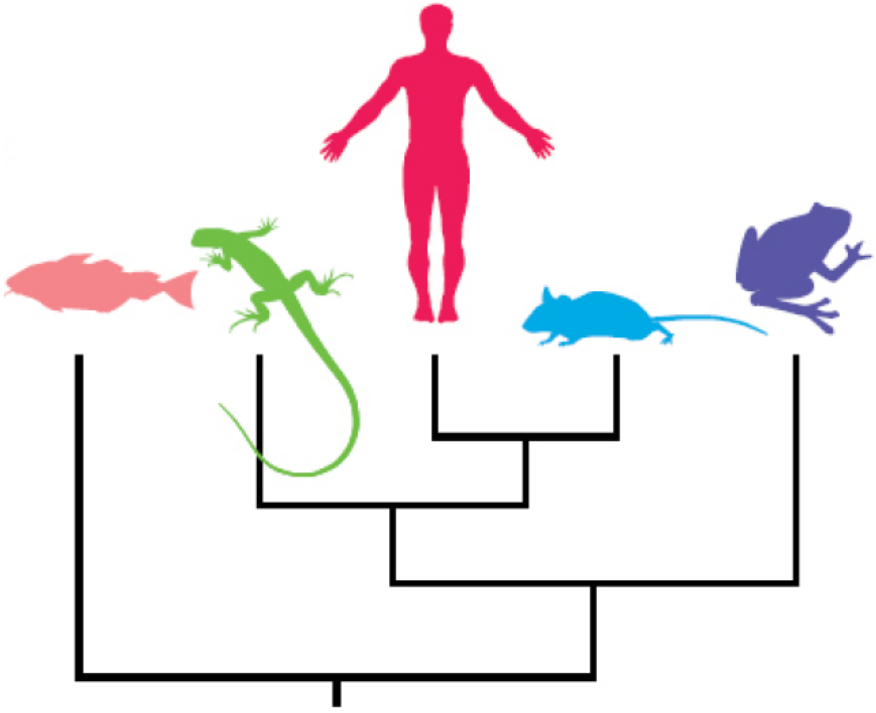
Which phylogenetic tree is more accurate?



Baum et al., 2005

The tree thinking challenge

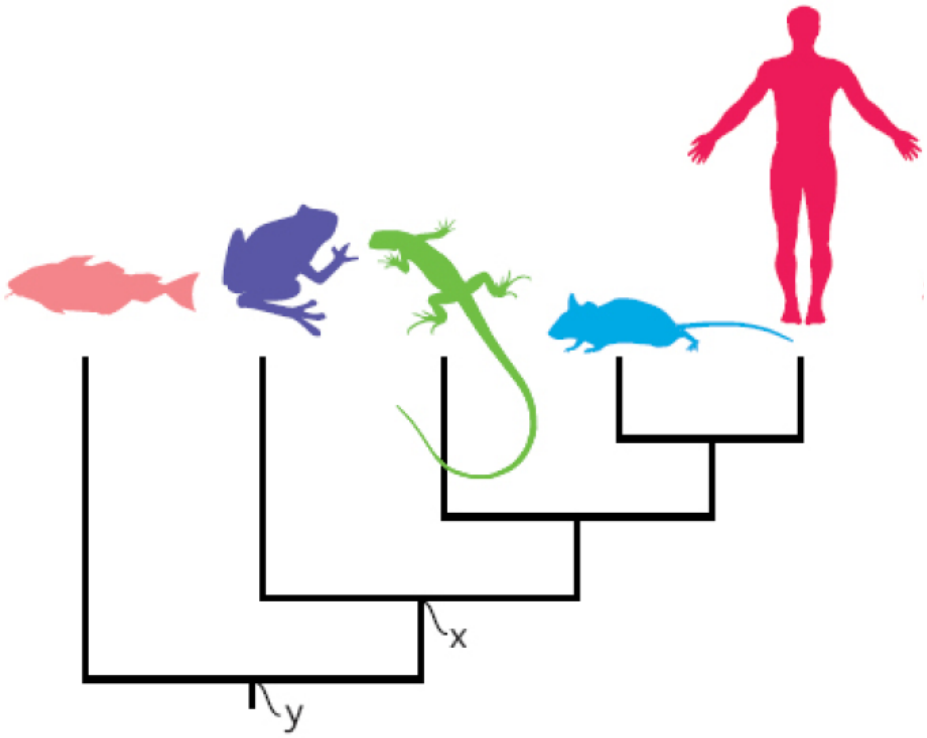
Is the frog more closely related to the fish or to the human?



Baum et al., 2005

The tree thinking challenge

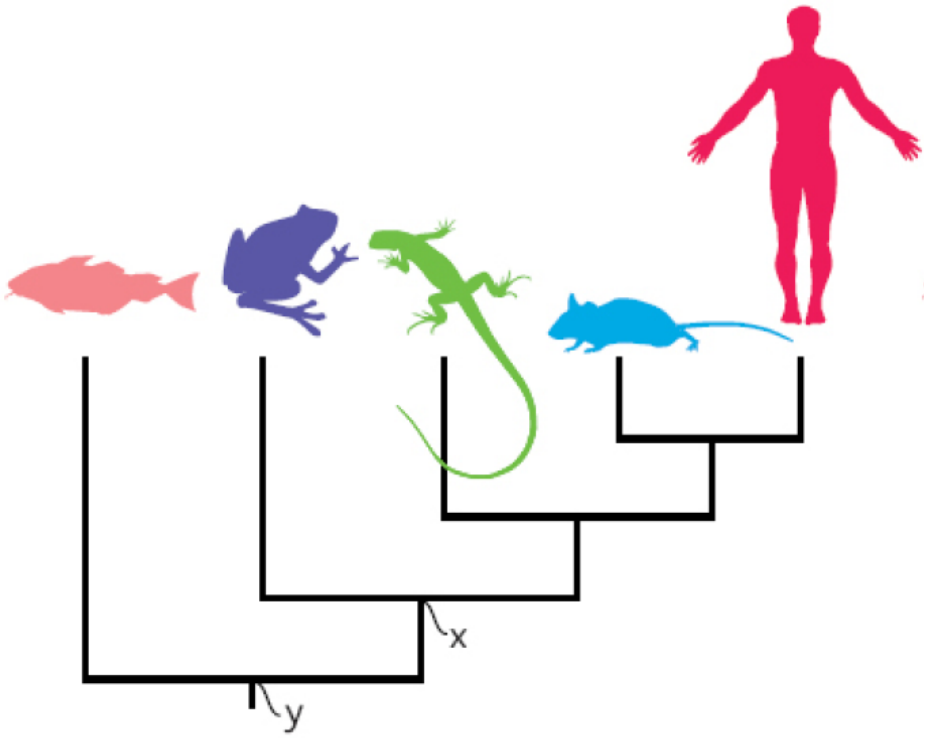
Is the frog more closely related to the fish or to the human?



Baum et al., 2005

The tree thinking challenge

Is the frog more closely related to the fish or to the human?



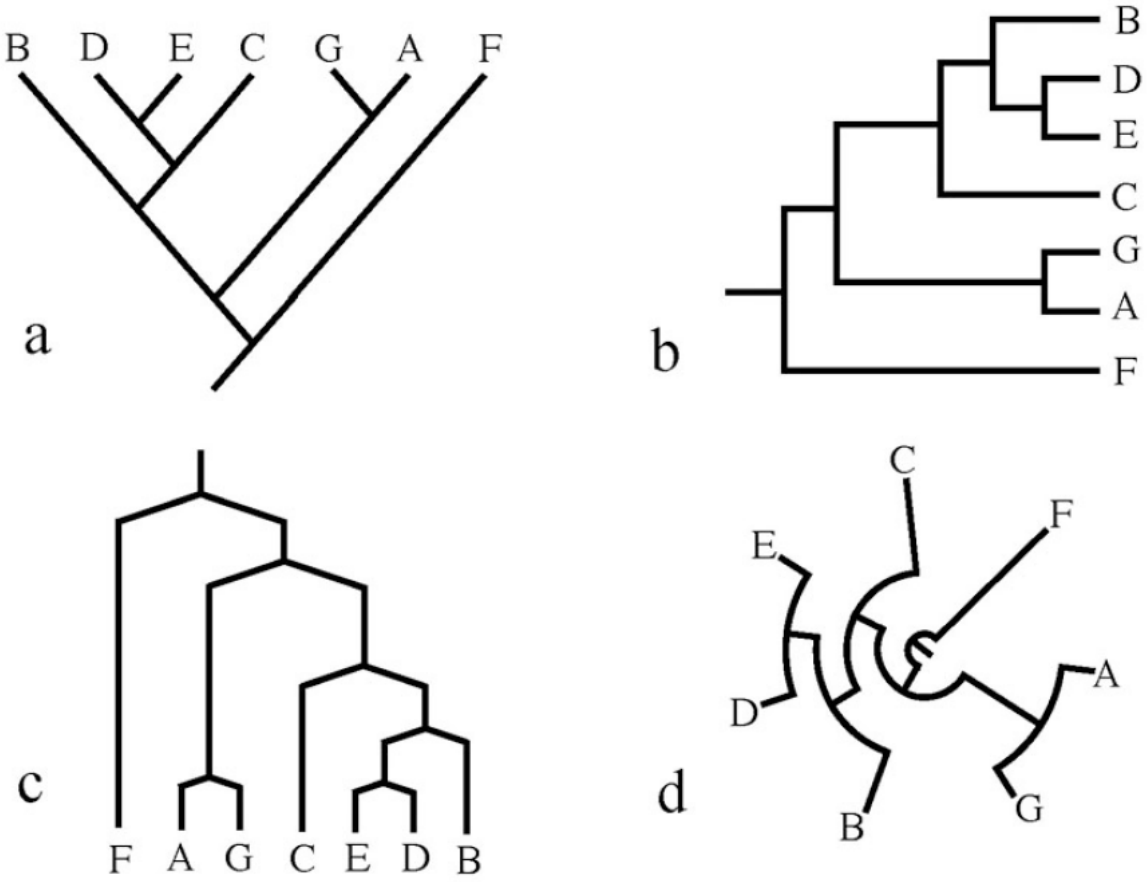
The more recently species share a common ancestor, the more closely related they are

X = the Most Recent Common Ancestor (MRCA)

Baum et al., 2005

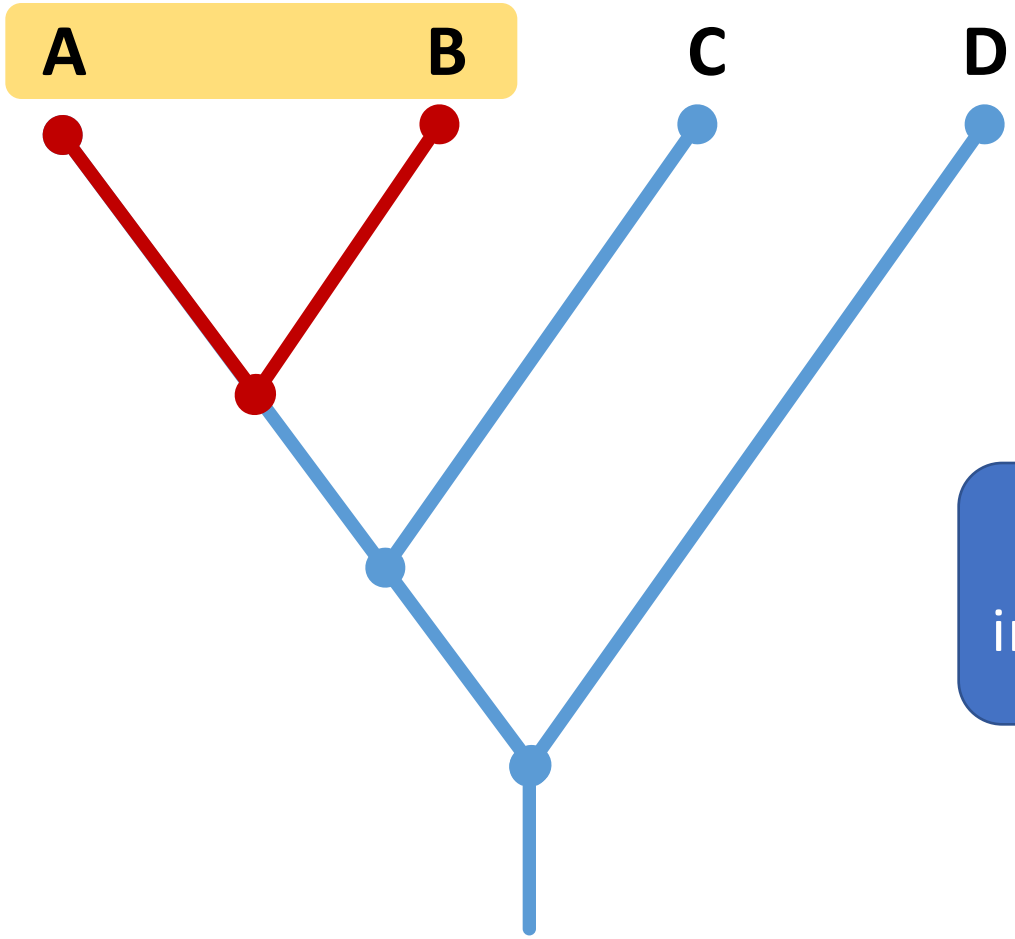
The tree thinking challenge

Which of the trees depicts a pattern different from the others?



Baum et al., 2005

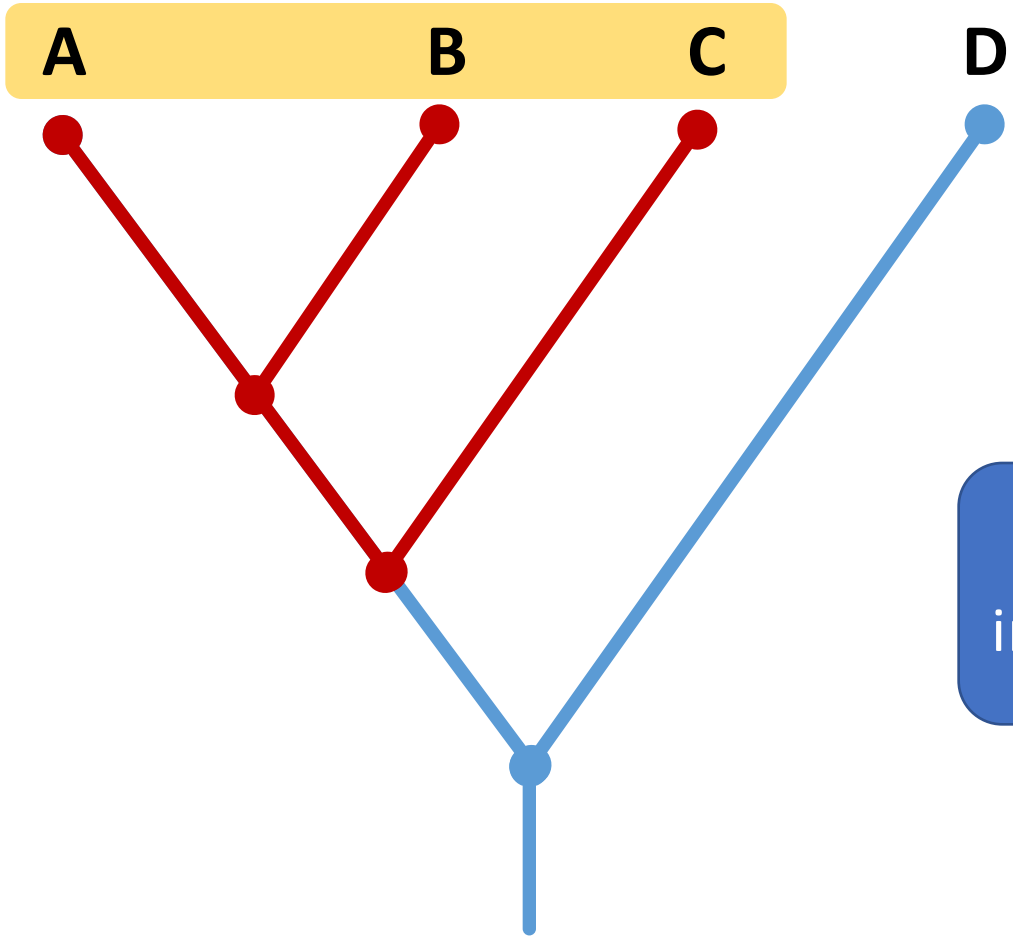
Understanding Tree Topology



Monophyletic group

OTUs share MRCA, which includes all of its descendants

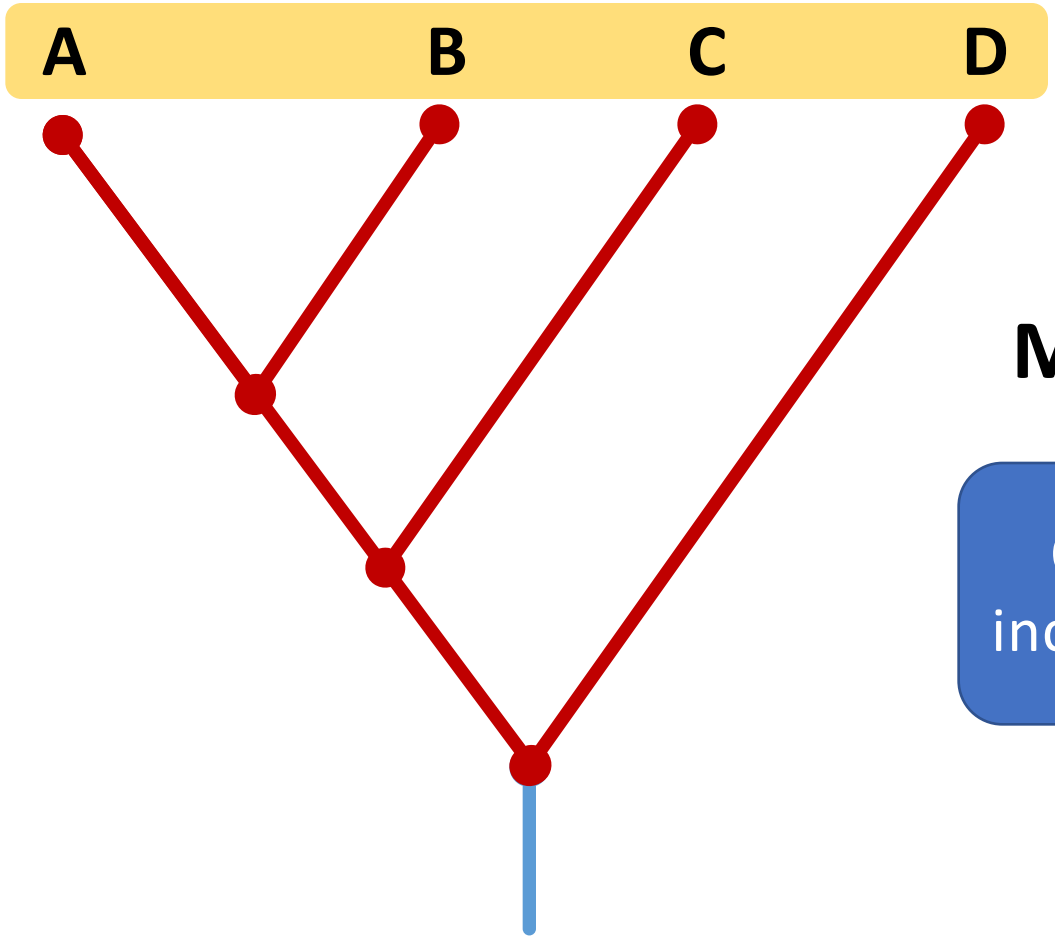
Understanding Tree Topology



Monophyletic group

OTUs share MRCA, which includes all of its descendants

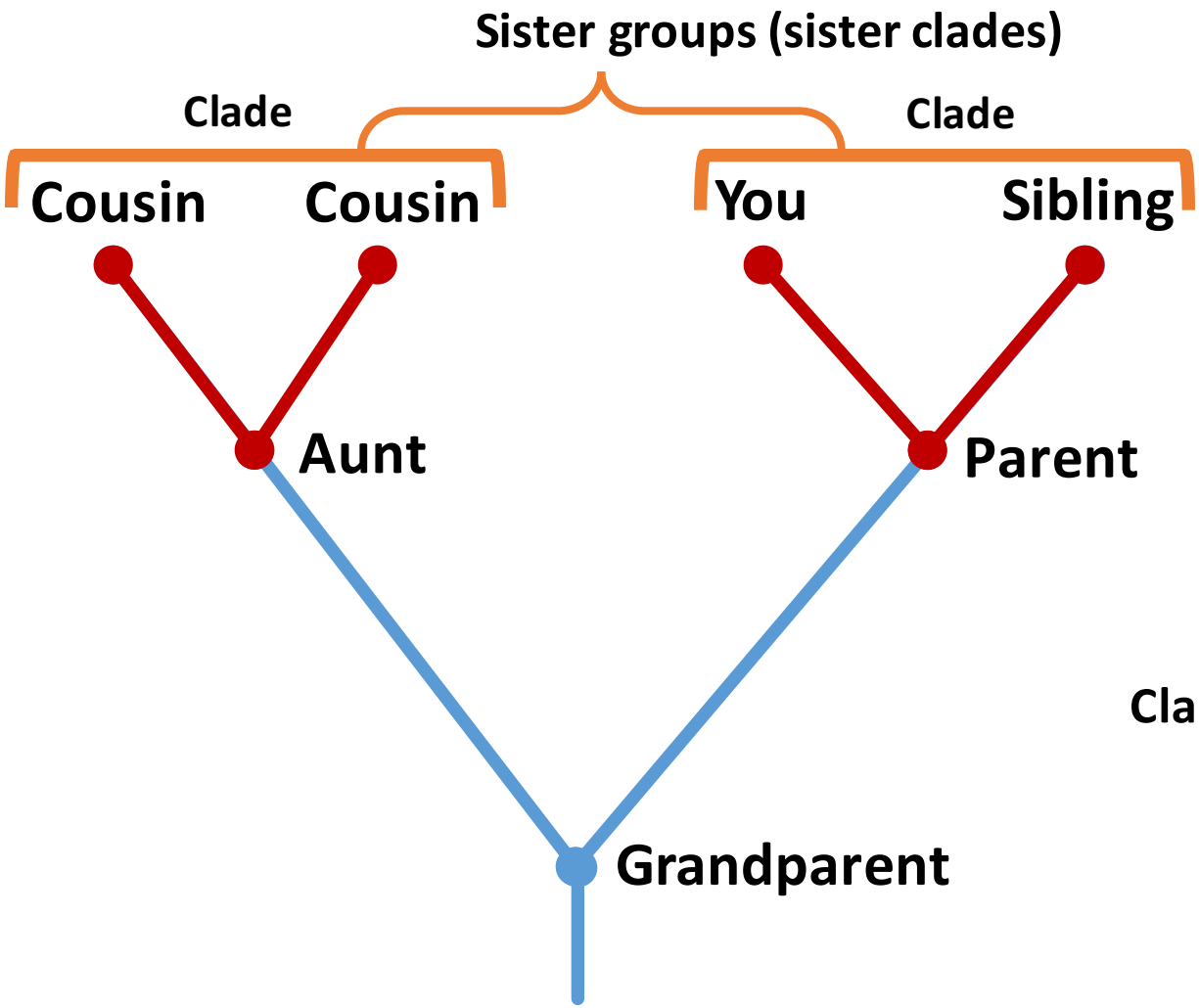
Understanding Tree Topology



Monophyletic group

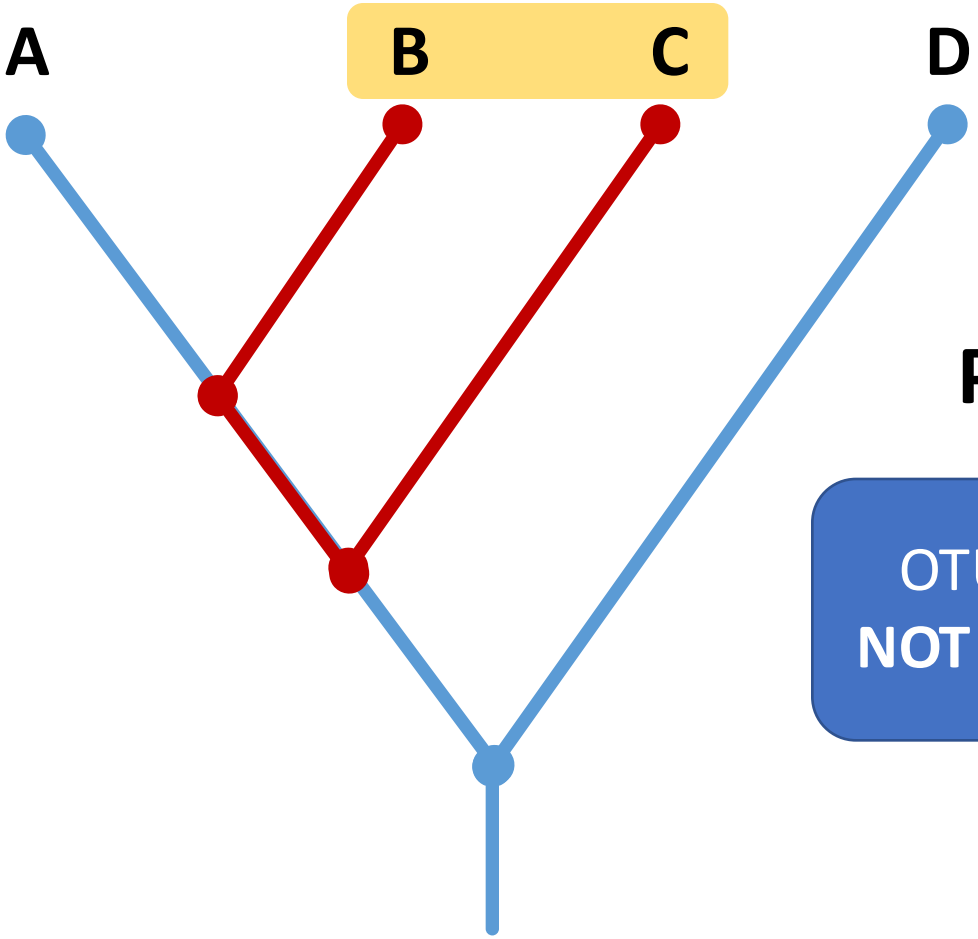
OTUs share MRCA, which includes all of its descendants

Understanding Tree Topology



Clade = monophyletic group

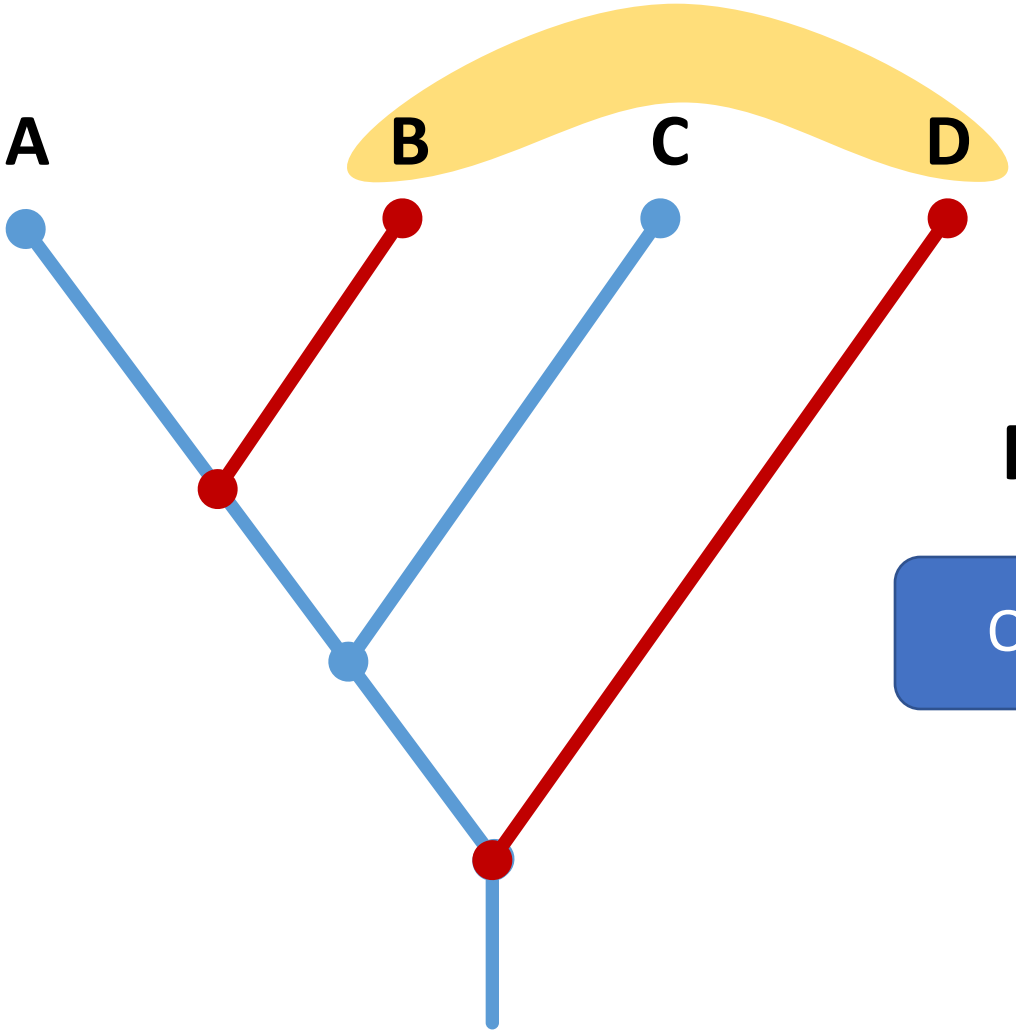
Understanding Tree Topology



Paraphyletic group

OTUs share MRCA, which does **NOT** include all of its descendants

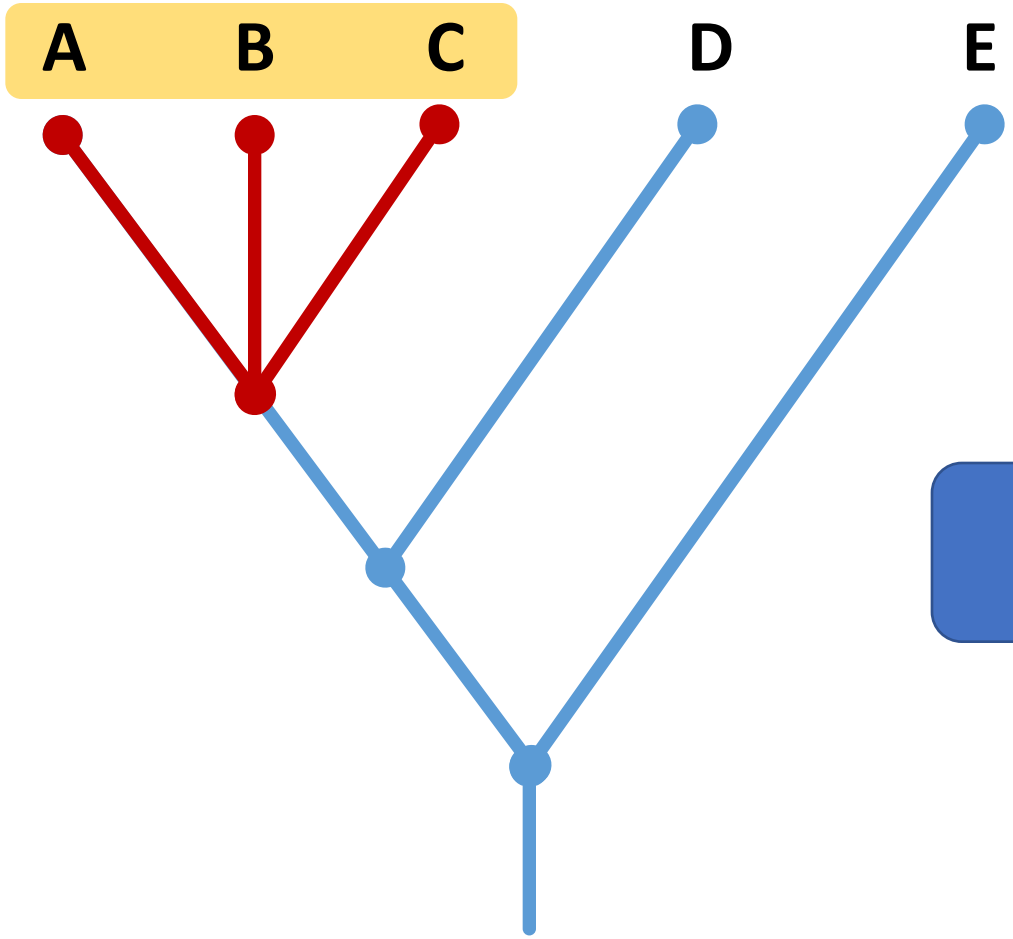
Understanding Tree Topology



Polyphyletic group

OTUs do NOT share MRCA

Understanding Tree Topology

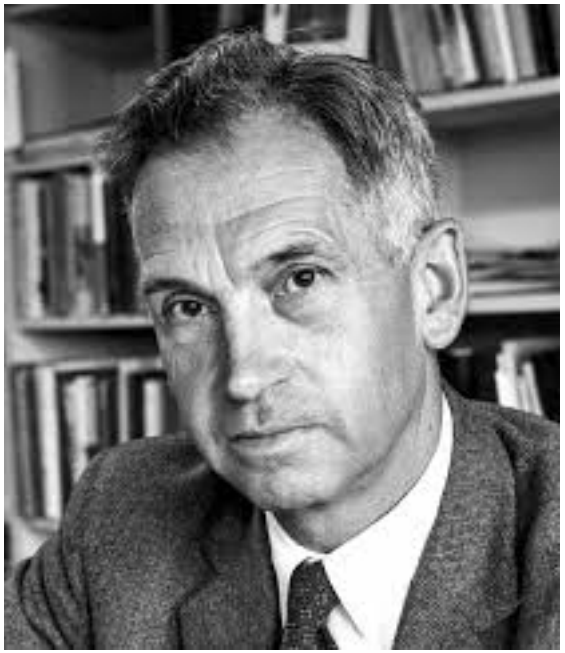


Polytomy

shows ambiguities in relationships

Rise of Evolutionary Systematics (subjective approach)

George Simpson (1902-1984)

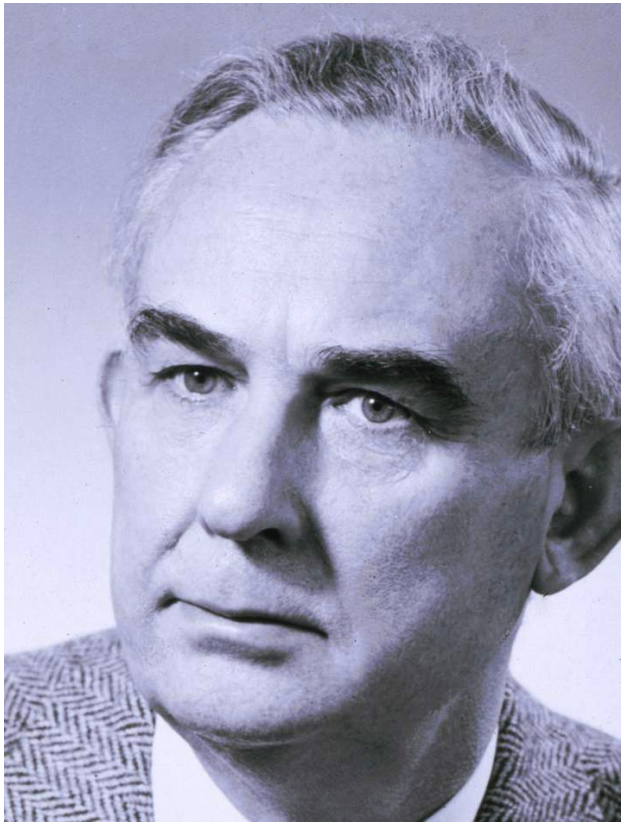


Ernst Mayr (1904-2005)

Organisms must be classified based on their evolutionary relationships rather than their morphological similarity… But how?

Phylogenetic Systematics (objective approach)

Hennig proposed the idea that evolutionary relationships between organisms must be revealed by using some objective analytical approach. He introduced the concepts of common ancestry and shared derived characters (synapomorphies) that were then implemented in cladistics analysis.



Willi Hennig (1913-1976)

"Apomorphic" – derived or specialized character

"Plesiomorphic" – ancestral character

Phylogenetic Systematics → Cladistics

Cladistics is an approach in biological classification in which organisms are grouped in taxa based on their **synapomorphies**; those are shared derived characters within a monophyletic group

THE BIGGEST CHALLENGE:

How do we know which characters to compare for shared ancestry?

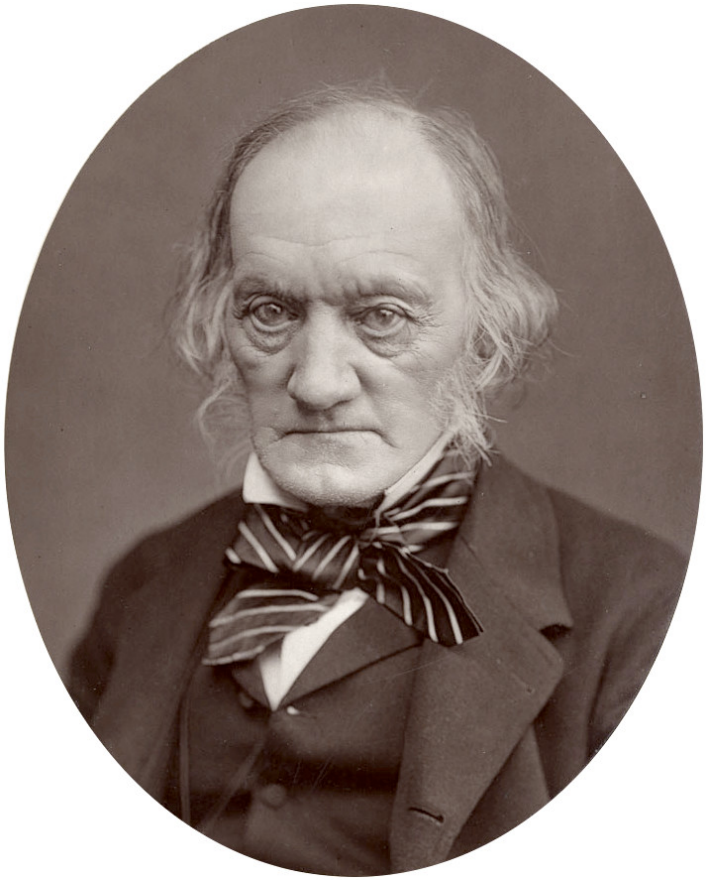
= for reconstructing phylogenetic trees?

Homology

"the same organ in different animals under every variety of form and function"

Analogous characters or genes have similar structure or function but do NOT share a common ancestor.

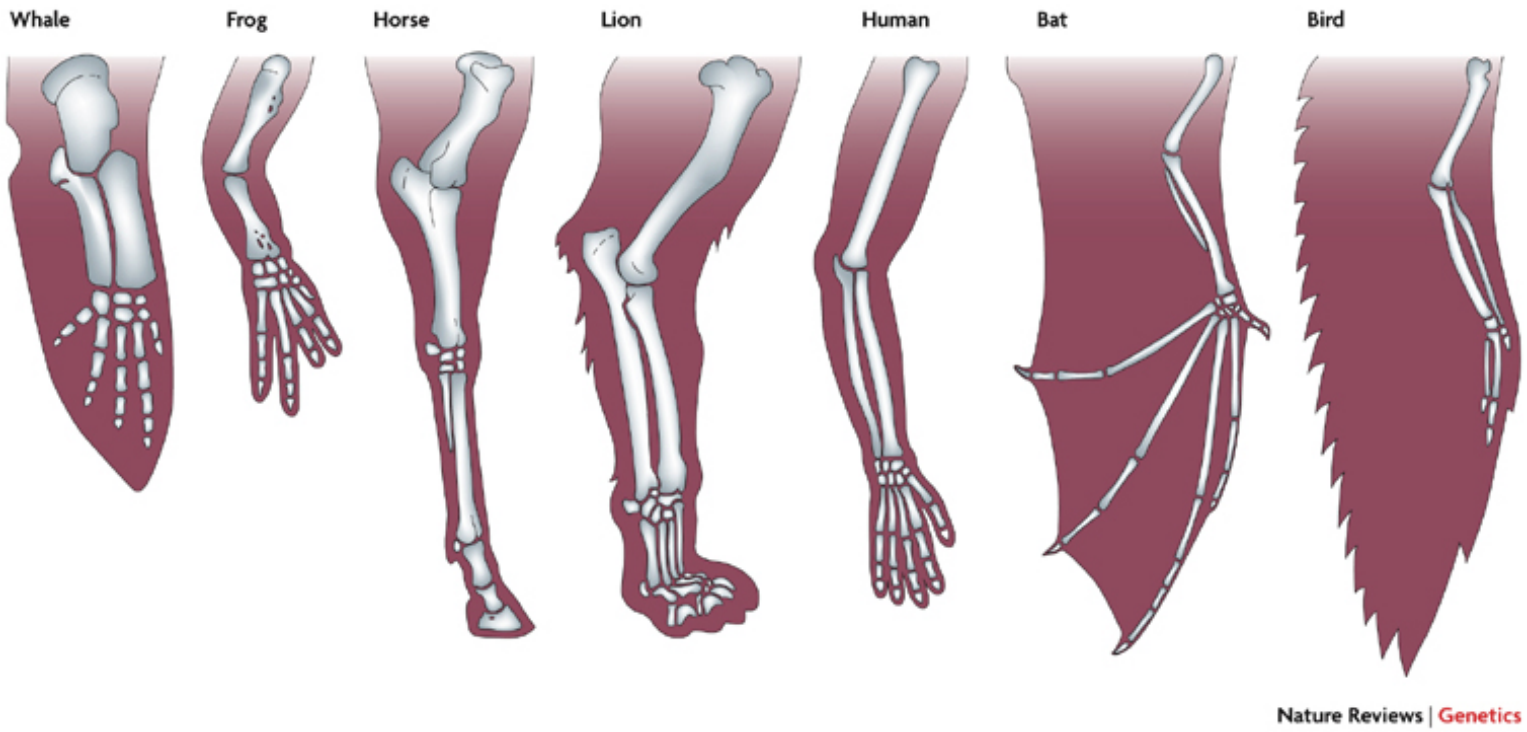
Homologous characters or genes are descending from a common ancestor.



Richard Owen (1804-1892)

Homology of forelimbs

Homologous characters can have different shapes and functions

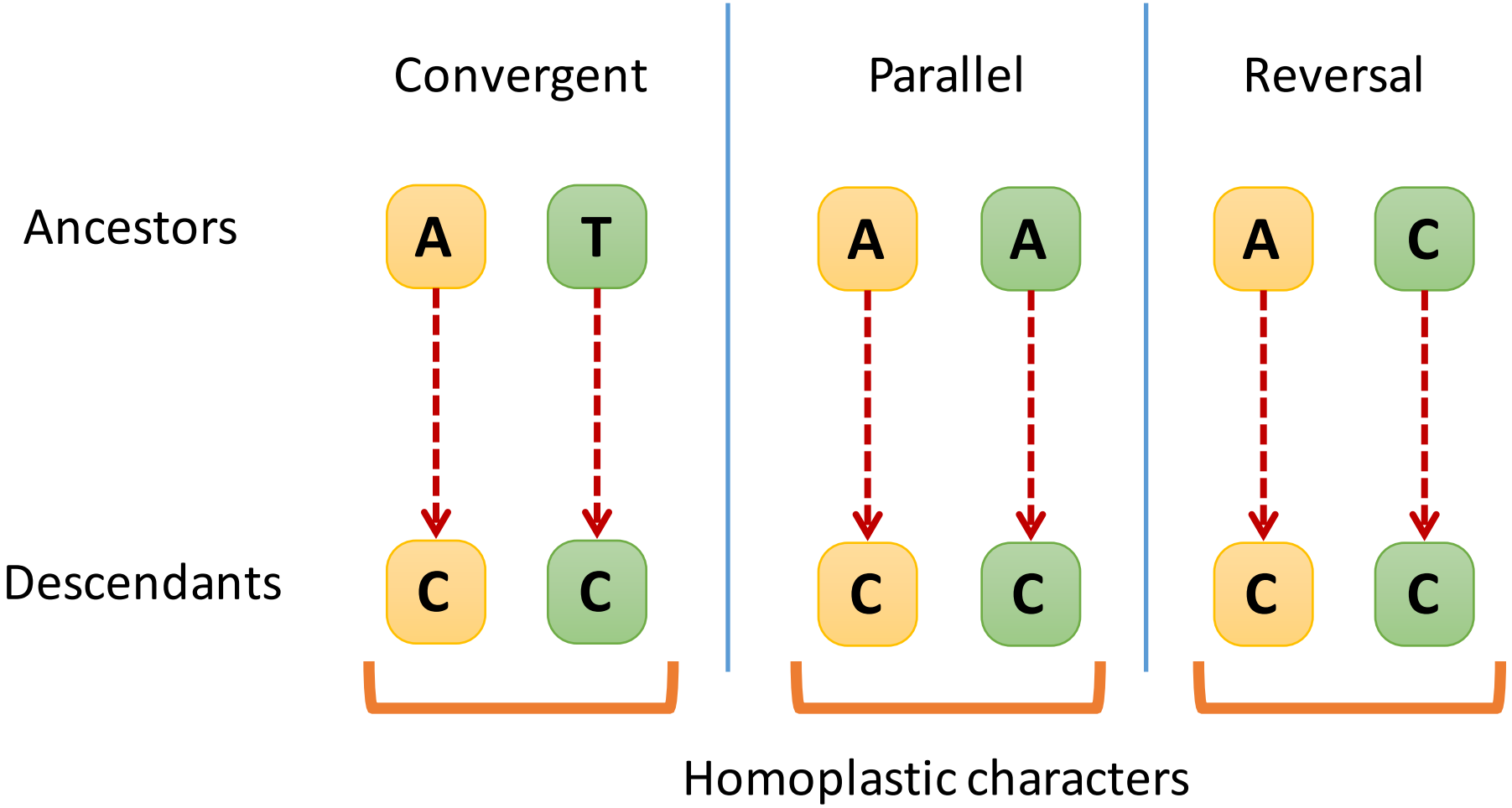


Wagner, 2007

Analogous characters look or perform very similar but derived independently (do not share MRCA)

Convergent evolution – independent evolution of similar characters in taxonomic groups from different lineages

Homoplasy (= analogy) in nucleotide substitutions

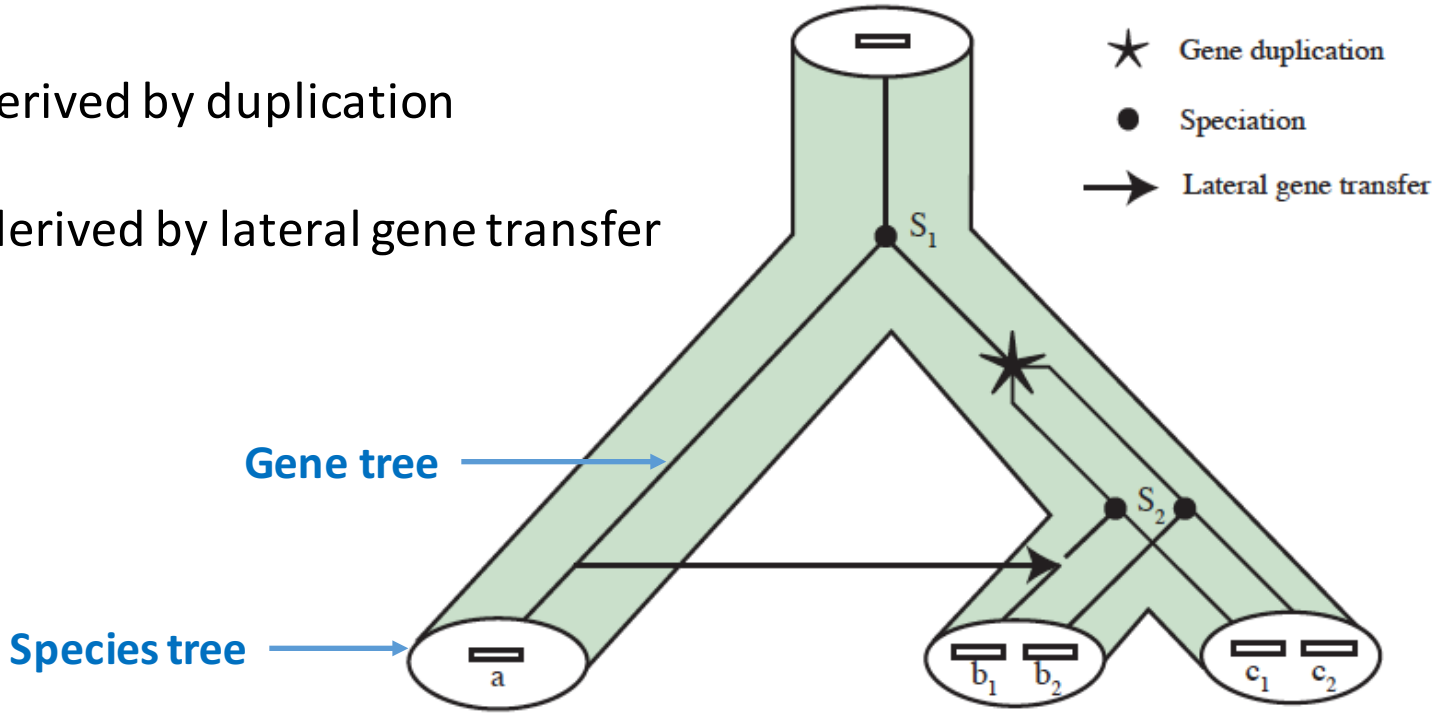


Homologous genes can be classified into:

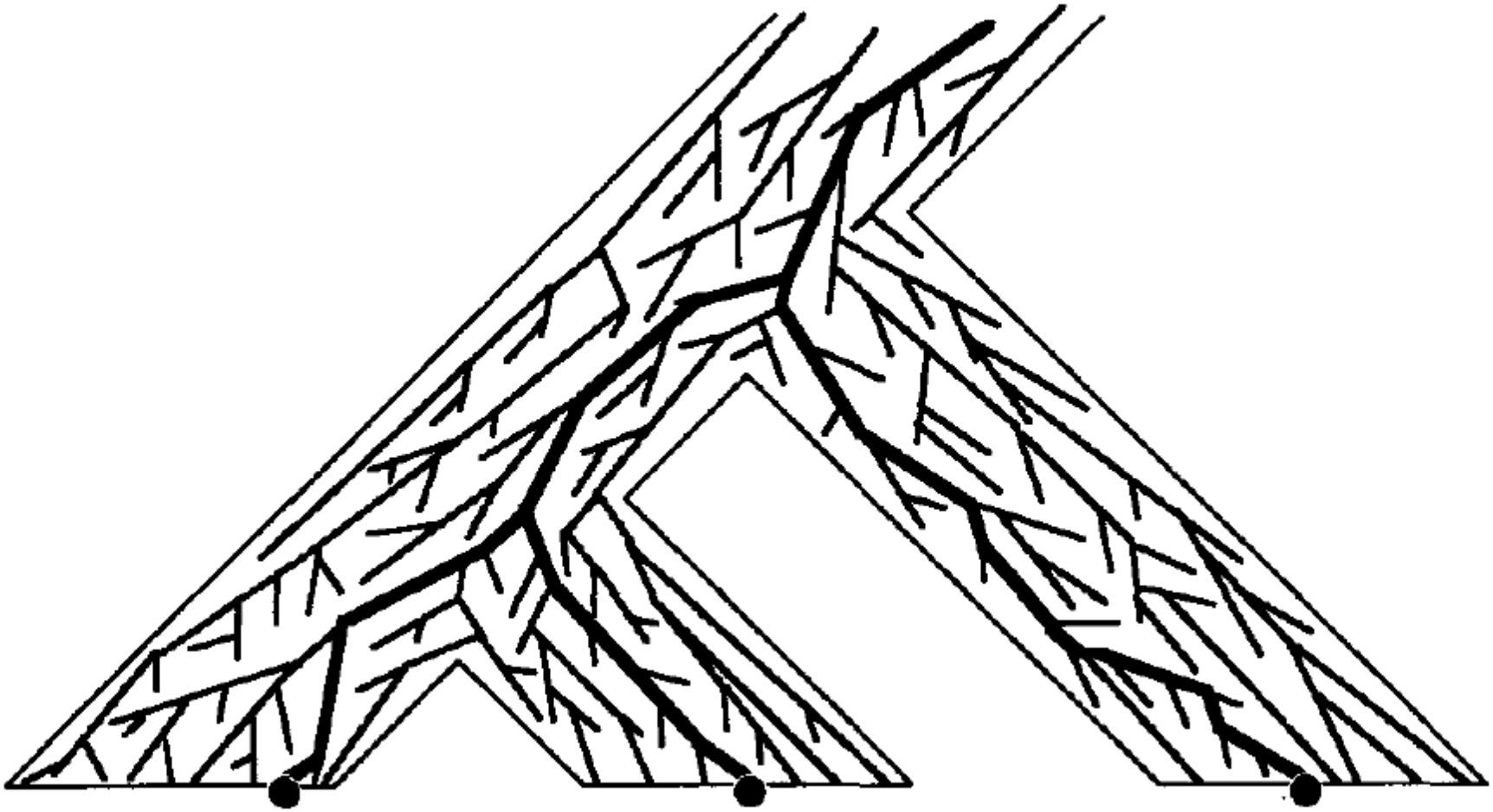
Orthologs – derived by speciation

Paralogs – derived by duplication

Xenologs – derived by lateral gene transfer



Genes trees contained within a species tree



Maddison, 1997

Molecular markers for phylogenetic inference

- Must represent homologous sequences
- Orthologs can be inferred by sequence similarity (alignment score or a graph distance) and by gene vs. species tree reconciliation
- Paralogs are good for reconstructing gene trees but bad for reconstructing species trees, and therefore, must be avoided

THE BIGGEST CHALLENGE:

How do we know which ~~characters to compare for shared ancestry~~? genes are orthologous?

How similar must two sequences be in order to be considered homologous?

PROTEINS: > 25 % of the amino acids

DNA & RNA: > 70 % of the nucleotides

MUST BE SIMILAR*

* Assumption usually applies when sequence length is more than 100 amino acids or nucleotides

Large scale orthology inference software

Phylogeny based methods:

- NOTUNG (Chen et al., 2000)
- Orthostrapper (Storm & Sonnhammer, 2002)
- LOFT (Heijden et al., 2007)
- Ensemble (Flicek et al., 2008)

Pairwise based methods:

- OrthoMCL (Li et al., 2003)
- Roundup (Deluca et al., 2006)
- MultiParanoid (Alexeyenko et al., 2006)
- OMA (Roth et al., 2008)
- InParanoid 7 (Ostlund et al., 2010)
- Ortholog-Finder (Horiike et al., 2016)

Graph based methods:

- OrthoMCL-DB (Chen et al., 2006)
- COG (Tatusov et al., 2000)
- eggNOG (Jensen et al., 2008)

Online databases for identifying orthologous genes

OrthologID

Orthologous Group Search

Query Orthology Classification

About OrthologID

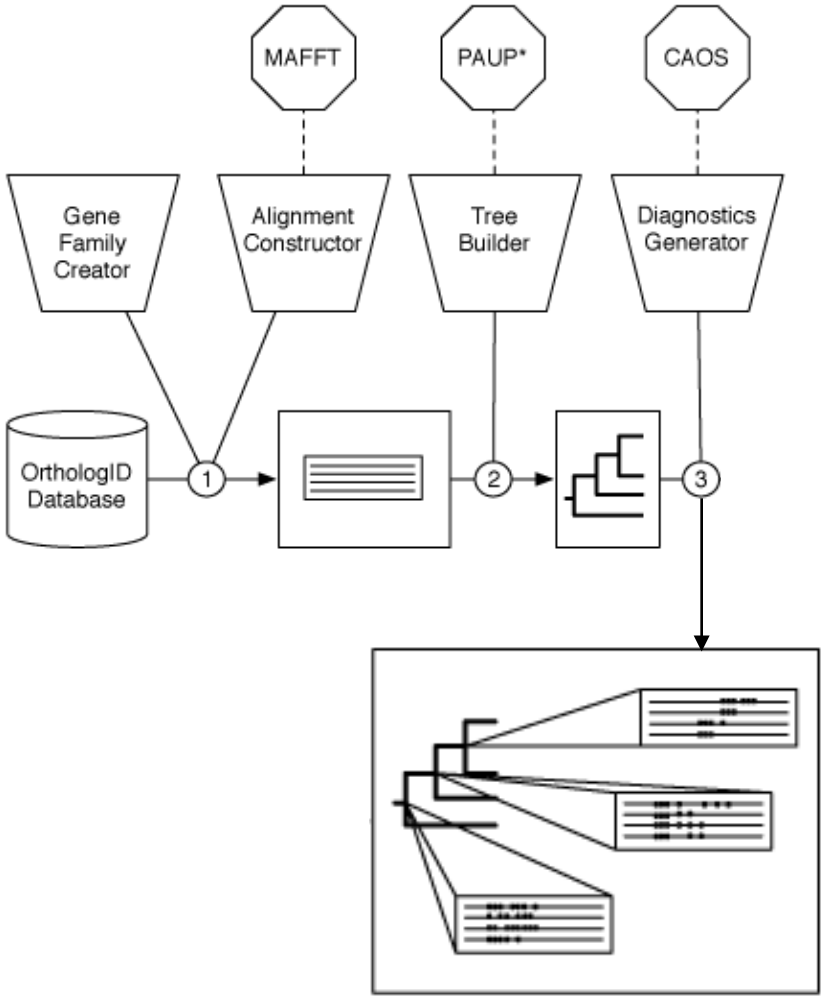
Welcome to OrthologID Online

OrthologID automates gene orthology determination within a character-based phylogenetic framework.

OrthologID Online identifies orthologous groups for complete genomes compiled in our database (*Orthologous Group Search*), and classifies user-input query sequences into orthologous groups generated from complete genomes (*Query Orthology Classification*). It identifies diagnostic characters that define each orthologous group, as well as diagnostic characters responsible for classifying query sequences. The output is presented in phylogenetic tree format.

Joanna C. Chiu, Ernest K. Lee, Mary G. Egan, Indra Neil Sarkar, Gloria M. Coruzzi, and Rob DeSalle. OrthologID: automation of genome-scale ortholog identification within a parsimony framework *Bioinformatics* (2006) 22(6): 699-707 first published online January 12, 2006 doi:10.1093/bioinformatics/btk040

This site uses an early version of the OrthologID Plant database and pipeline. To install the latest version of the OrthologID pipeline on your own cluster, follow the instructions here to download the source code from our Mercurial repository.

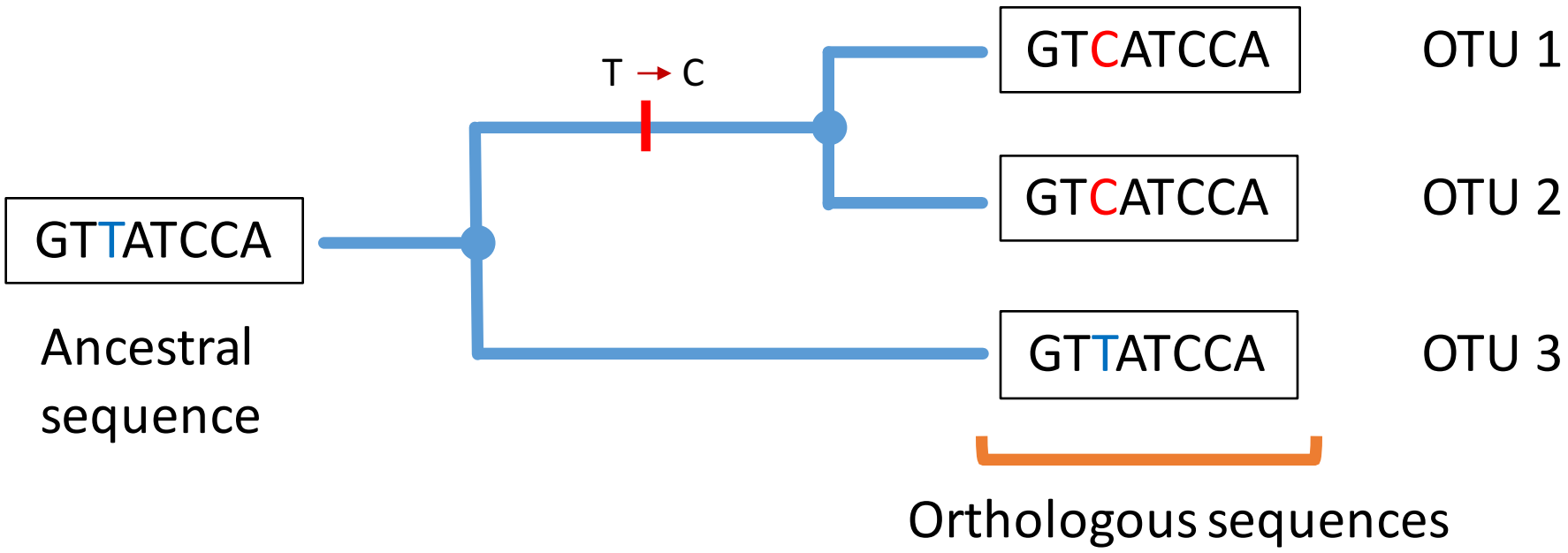


http://nypg.bio.nyu.edu/orthologid

http://www.orthodb.org

http://pantherdb.org/genes/

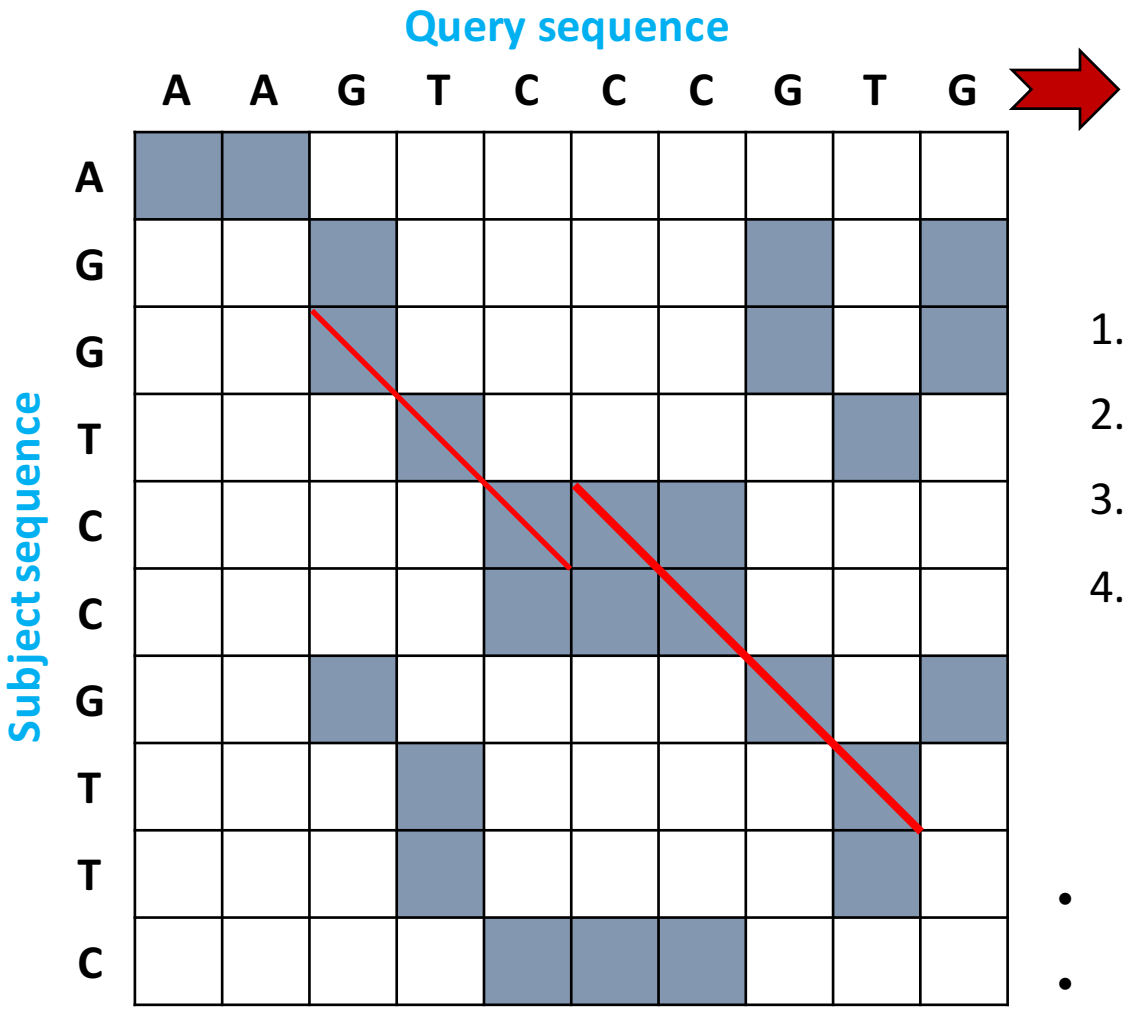
Sequence alignment is an important step in phylogenetic analysis



Important applications of sequence alignment:

- Prediction of function (e.g., protein annotation)
- Database searching (e.g., barcode gene id)
- Gene finding (e.g., finding gene within a genome)
- Sequence divergence (e.g., homology test)
- Sequence assembly (e.g., Illumina reads)

Dot-plot alignment



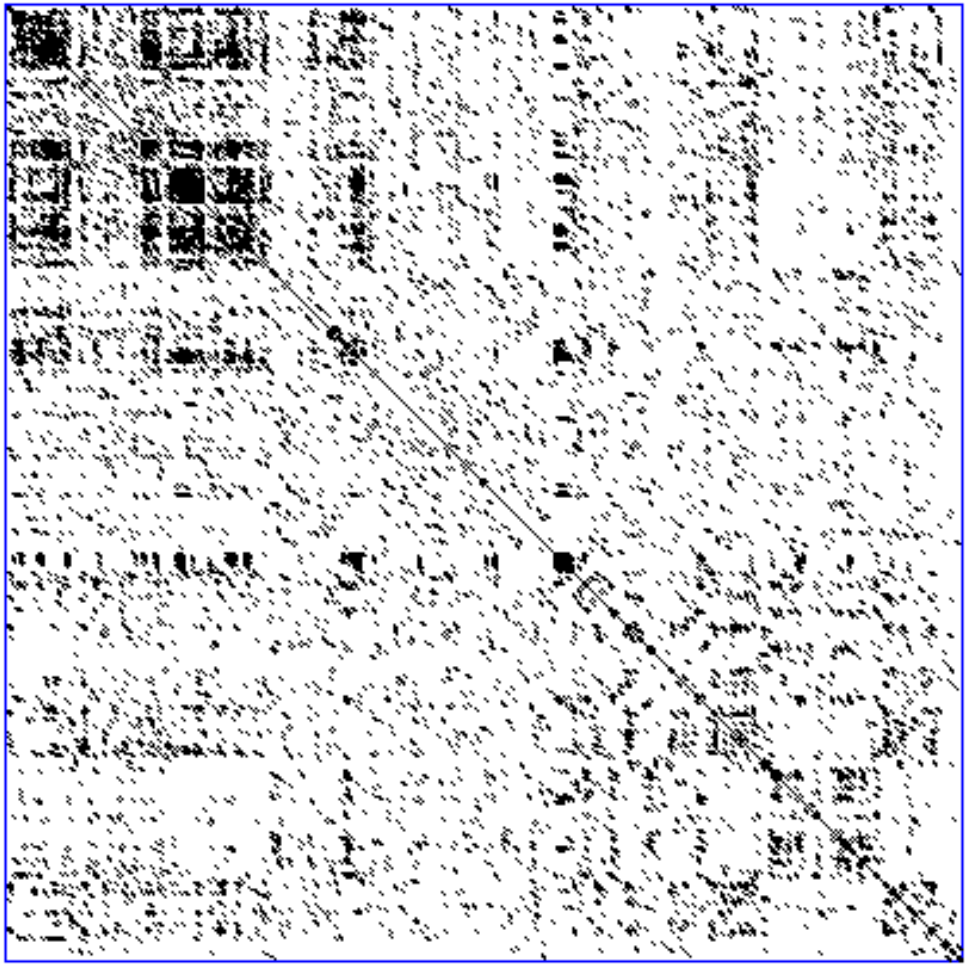
Algorithm

1. Query sequence moves right
2. Mark down all residual matches
3. Count matches on a diagonal
4. Choose the alignment with the largest score

Parameters

- Minimal diagonal length (window)
- Minimal # of matches (stringency)

Dot-plot alignment



Self-similarity of a human zinc finger transcription factor

Dot-plots are **NOT** predictive of homology. They are solely used to highlight similarity between a pair of sequences.

Dynamic alignment methods: Global vs. Local

Global (Needleman-Wunsch algorithm):

sequences are aligned over their entire lengths

```
--T--CC-C-AGT--TATGT-CAGGGGACACG-A-GCATGCAGA-GAC
  |   ||  |   ||   |  |  |  |||      ||  |  |  |   |  ||||    |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG-T-CAGAT--C
```

Local (Smith-Waterman algorithm):

only the most similar portions are aligned

```
                          TCCCAGTTATGTCAGGGGACACGAGCATGCAGAGAC
                             |||||||||||||
AATTGCCGCCGTCGTTTTCAGCAGTTATGTCAGATC
```

When to use pairwise alignments?

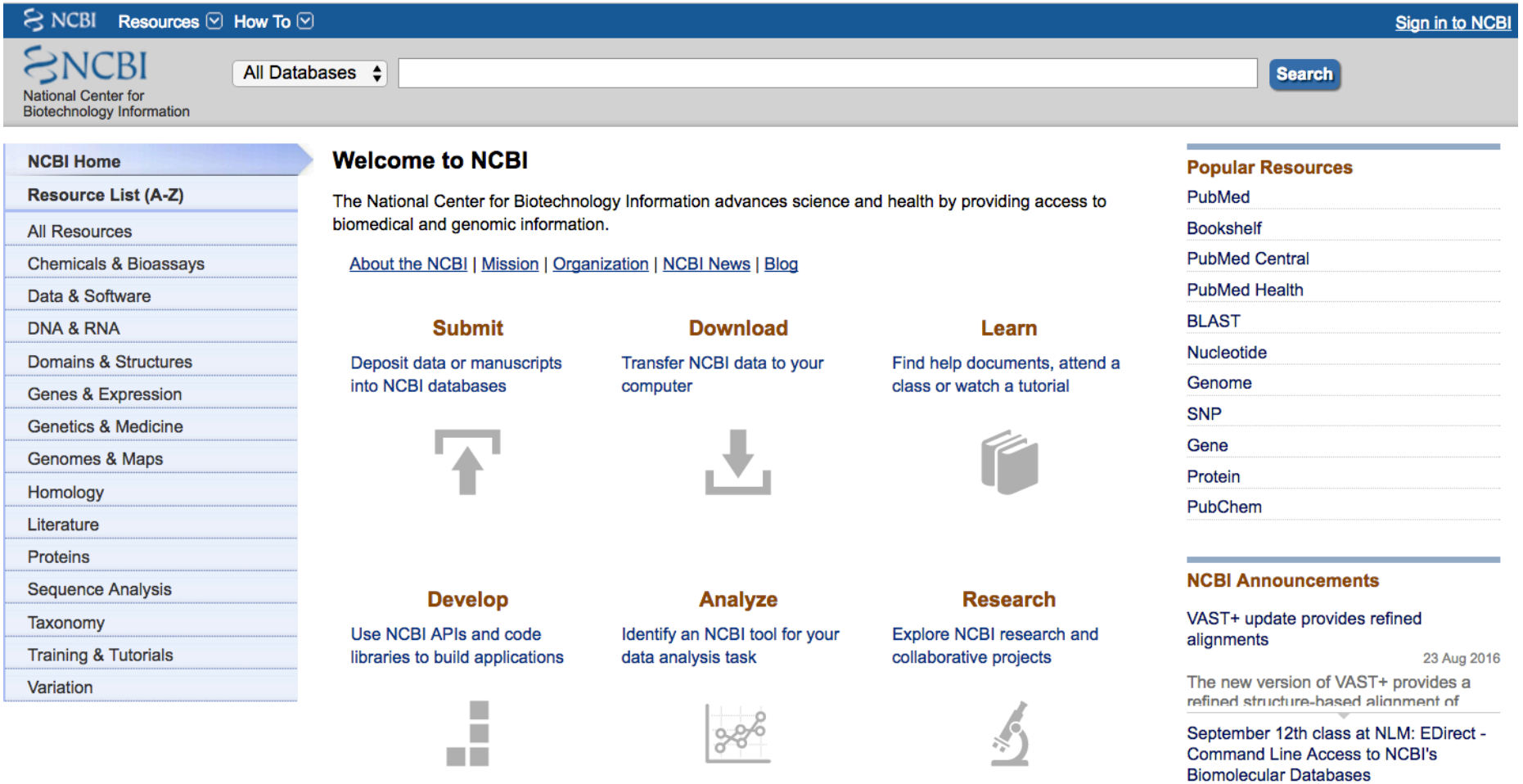
Method	Usage
Dot Plot	• General sequence exploration • Search for repeats • Finding indels
Local	• Comparing sequences with partial homology • Making high-quality alignments • Residue-per-residue analysis
Global	• Comparing sequences over their entire length • Identifying long indels and introns • Commonly used in phylogenetics/phylogenomics

Some popular sequence alignment software

Program	Alignment	Algorithm	Accuracy	Speed	Link
BLAST	local	heuristic	medium	slow	blast.ncbi.nlm.nih.gov
BLAT	local	heuristic	medium	medium	www.kentinformatics.com
Needle	global	heuristic	medium	fast	emboss.toulouse.inra.fr
FastA	local	heuristic	high	slow	fasta.bioch.virginia.edu
HMMER	both	Consistency-based	high	slow	hmmer.org
MAFFT	global	Iterative	medium	fast	mafft.cbrc.jp
MUSCLE	global	Iterative	medium	fast	www.drive5.com/muscle
SSEARCH	local	heuristic	high	very slow	www.biology.wustl.edu/gcg/ssearch.html
T-coffee	both	Structure-based Consistency-based	high	slow	www.tcoffee.org
USEARCH	global	heuristic	low	very fast	www.drive5.com/usearch

Online DATABASES for sequence data:

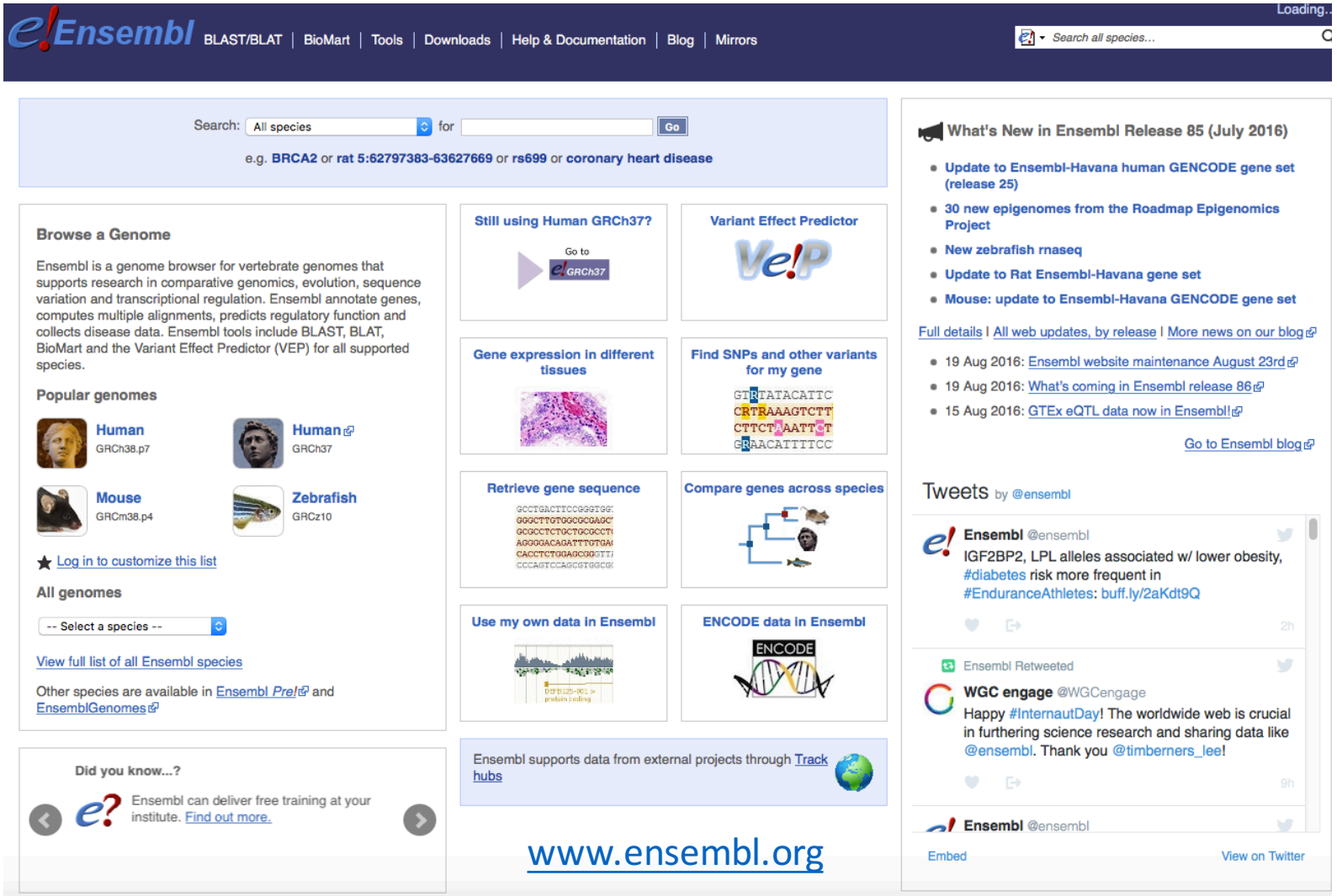
NCBI: National Center for Biotechnology Information (aka 'GeneBank')



www.ncbi.nlm.nih.gov

Online DATABASES for sequence data:

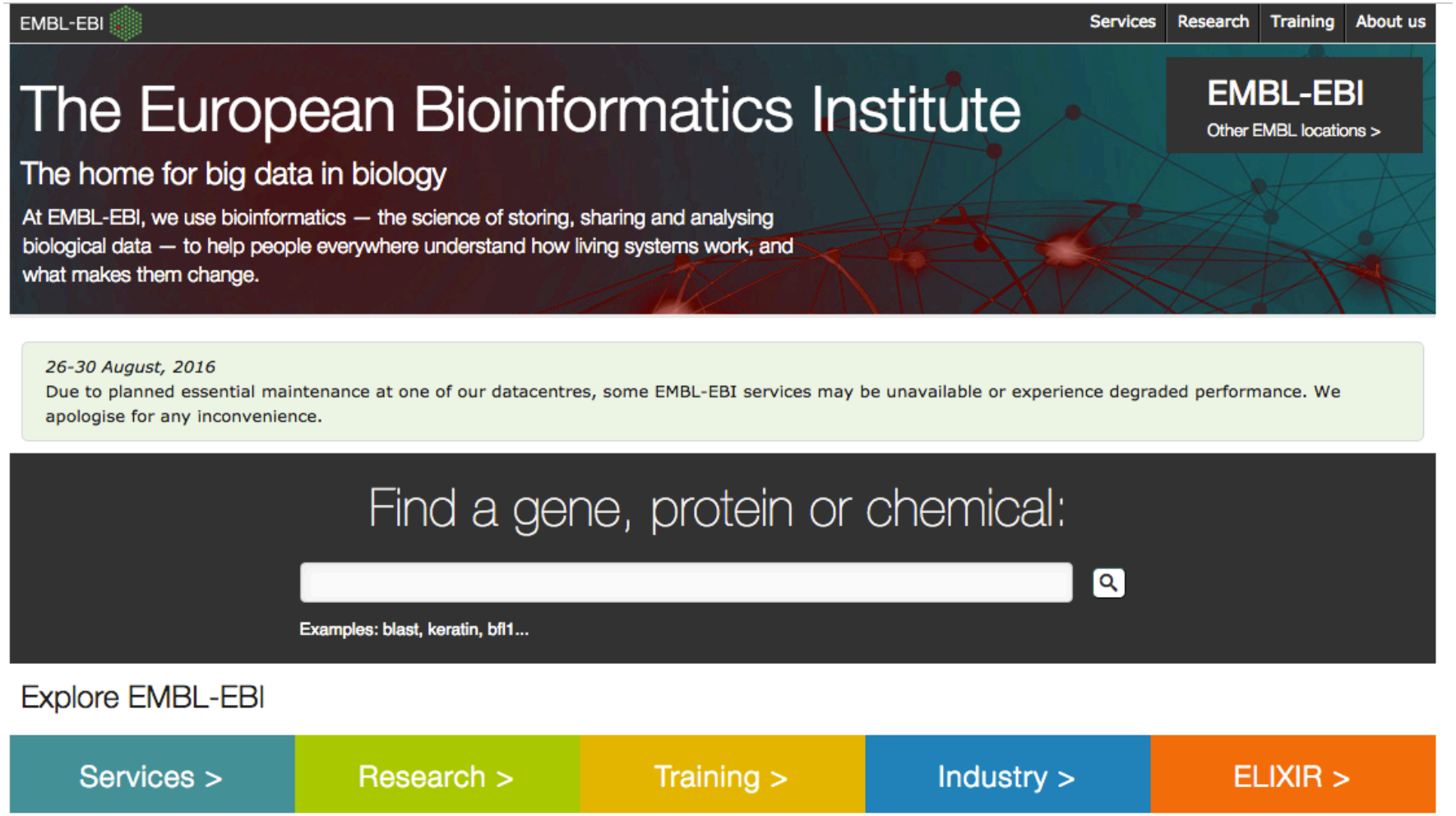
ENSEMBL: vertebrate genome browser



www.ensembl.org

Online DATABASES for sequence data:

EMBL-EBI: The European Bioinformatics Institute



www.ebi.ac.uk

Online DATABASES for sequence data:

UniProt (Swiss-Prot & TrEMBL): protein functional information

Online DATABASES for sequence data:

ExPaSy: Bioinformatics Resource Portal

www.expasy.org

Phylogenetics vs Phylogenomics: what's the difference?

	Phylogenetics	Phylogenomics
DATA	• One or a few genes with known function • Long sequences	• Genomic data which often includes non-coding regions • Short sequences
Alignment	• Accurate and relatively fast • Easy to visualize and analyze	• Not accurate and slow • Hard to visualize and analyze
Phylogenetic inference	• Partition test • Model test • Model based methods	• No partition test • No model test • Distance based methods • Simple model based with uniform prior for all genes
Tree visualization	• Conventional	• Super-trees • Phylogenetic networks • Cloudogram