Sistemática Teórica

Hernán Dopazo

Biomedical Genomics and Evolution Lab

Lesson 01

Introduction

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Argentina

2013

About

- This short, but intensive course, has the purpose to introduce students to **the main concepts of molecular evolution and phylogenetic analysis**:
  - Historical Background
  - Applications of Phylogenies
  - Cladograms & Phylograms
  - Homology & Similarity
  - Molecular Evolution
  - Models of Sequence Evolution
  - Phylogenetic Methods:
    - Distances,
    - Statistical Approaches
      - Maximum Likelihood
      - Bayesian Inference

- The course consists of **a series of lectures, and PC. Lab. sessions that will familiarise the student with the statistical problem of phylogenetic reconstruction** and its multiple uses in biology.

Tree Basic questions

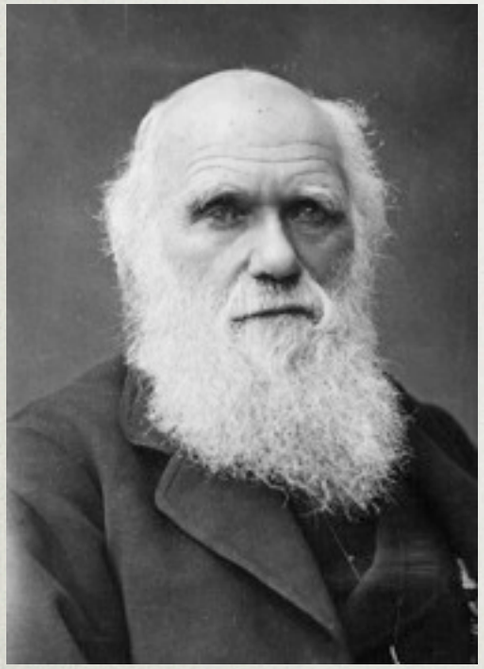
Tree basic questions

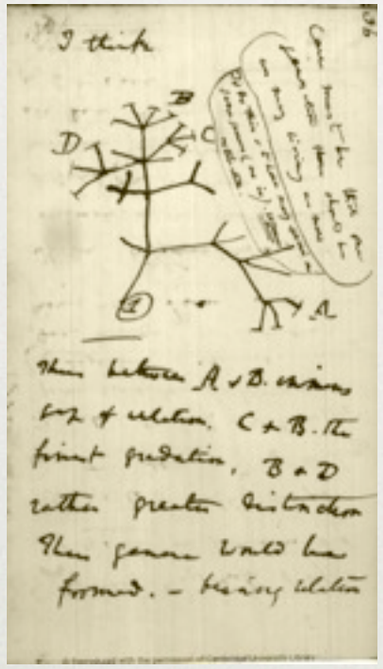
- Why use phylogenies?
  - Like astronomy, biology is an **historical science** !
  - The knowledge of the past is important to solve many questions related to biological patterns and processes.
- Can we know the past?
  - We can postulate alternative evolutionary scenarios (**hypothesis**)
  - Obtain the proper dataset and get statistical confidence
- What means to know "...the phylogeny"?
  - The ancestral-descendant relationships (**tree topology**)
  - The distances between them (**branch lengths**)

Phylogenies are working hypotheses !!!

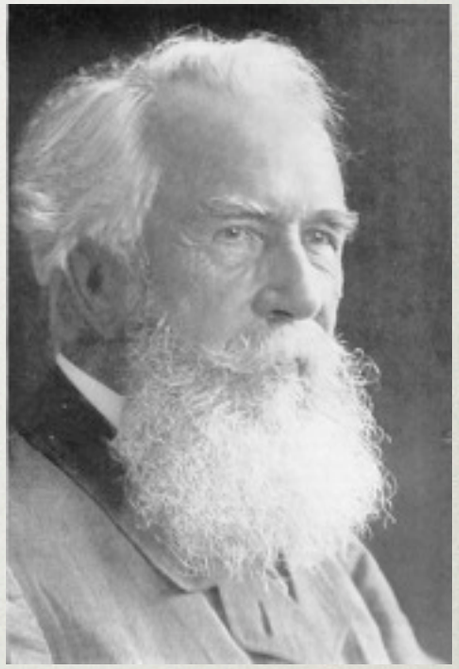
What are the roots of modern phylogenetics?

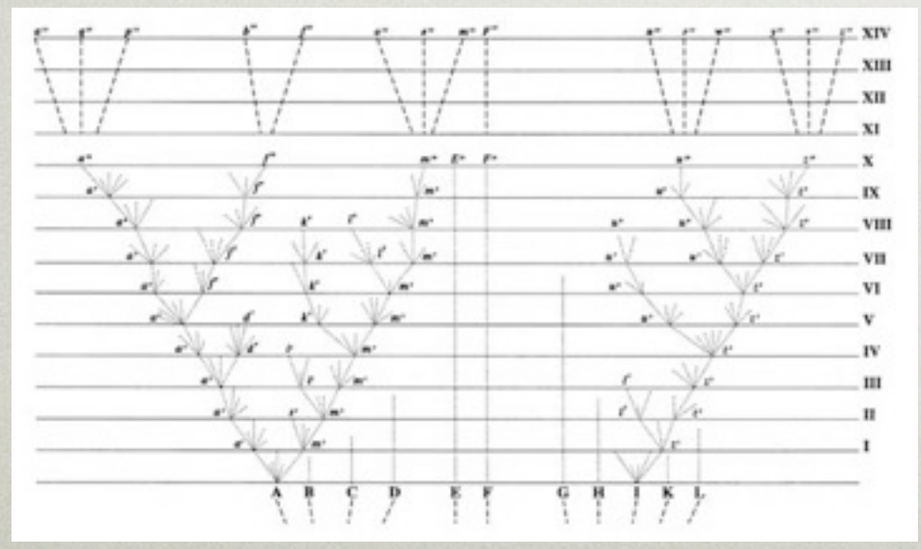
- Phylogenies have been inferred by systematics ever since they were discussed by **Charles Darwin** *(1837)*, and **Ernst Haeckel** *(1866)*. In mid-July 1837 Darwin started his "B" notebook on Transmutation of Species, and on page 36 wrote "I think" above his first evolutionary tree.



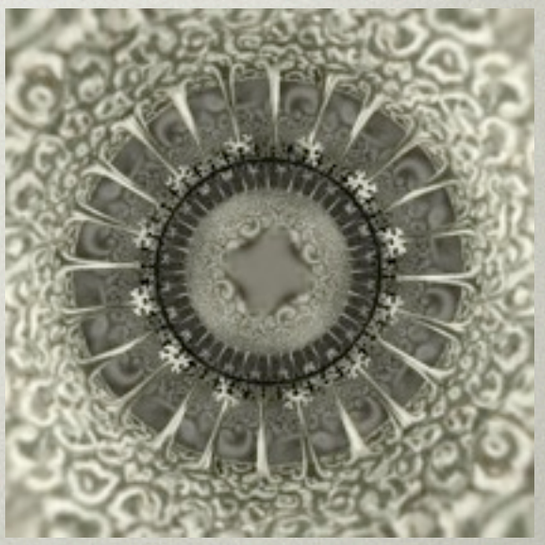








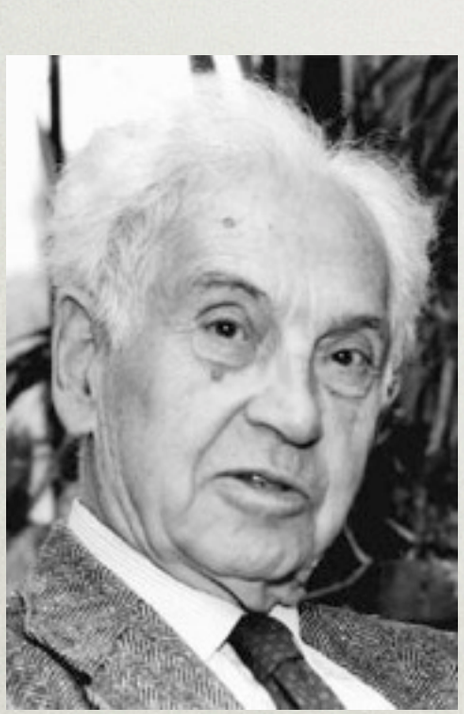


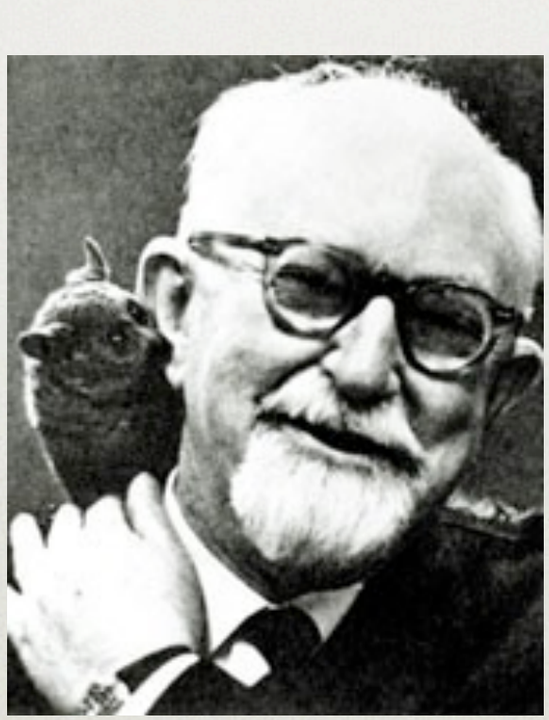


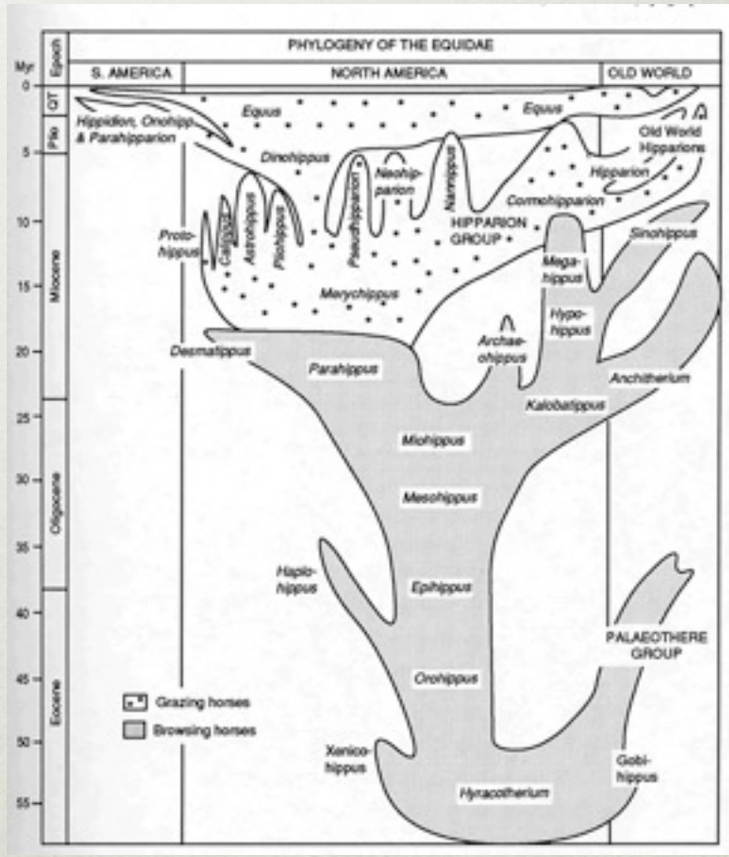
However, since 1950s-60s classifications began to be **more numerical, algorithmic and statistical**. Principally due to progress in molecular biology, protein sequence data and computer development (initially, using punched card machines).

Roughly, systematists divided in two:

1. Proponents of the "Evolutionary Systematic School" classify organisms using different historical, ecological, numerical, and evolutionary arguments. It attempts to represent, not only the branching of phyletic lines (**cladogenesis**) but also its subsequent divergence (**anagenesis**) leading the invasion of a new adaptive zone by a particular class of organisms (a grade). Its represents were **Ernst Mayr** *(1942)*, and **George G. Simpson** *((1944),* among others.







2. Proponents who rejected the notion of theory-free method of classification, introduced **objectivity** by using explicit numerical approaches.

(i) **Numerical Taxonomy's school** (**Phenetics**) originated by in Kansas:

Charles Michener. *A quantitative approach to a problem of classification. Evolution, 11: 490–499, 1957,* **Peter Sneath**. *The application of computers to taxonomy. Journal of general microbiology, 17: 201–226, 1957,* and **Robert Sokal.** *1963. Principles of Numerical Taxonomy, 1963.*

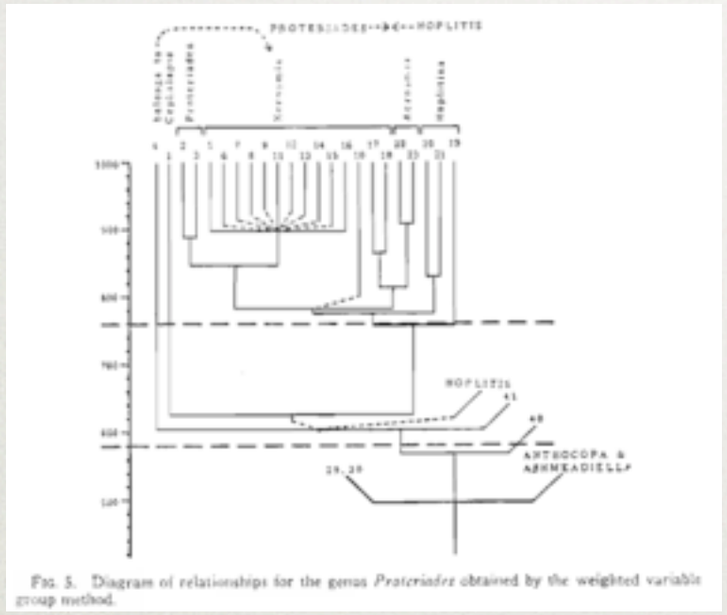
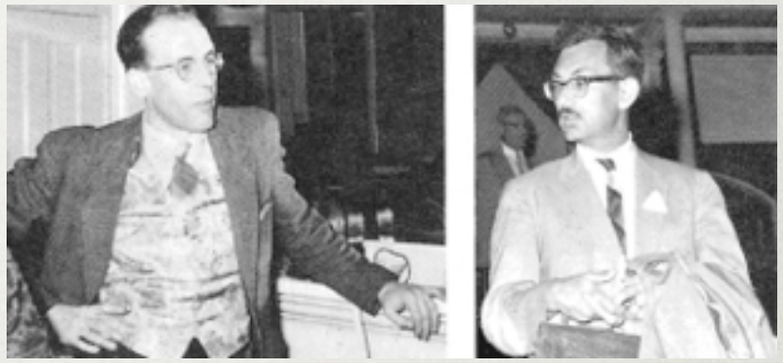
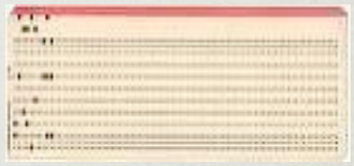


FIG. 5. Diagram of relationships for the genus *Prateriades* obtained by the weighted variable group method.



Punch Cards



A QUANTITATIVE APPROACH TO A PROBLEM IN CLASSIFICATION [1]

CHARLES D. MICHENER AND ROBERT R. SOKAL [2]

Department of Entomology, University of Kansas, Lawrence

Received October 10, 1956

SNEATH, P. H. A. (1957). *J. gen. Microbiol.* 17, 201–226

The Application of Computers to Taxonomy

BY P. H. A. SNEATH

National Institute for Medical Research, Mill Hill, London, N.W. 7

Main idea

To score pairwise differences between OTU's (**Operational Taxonomic Units**) using **as many characters as possible.** Cluster by similarity using an algorithm that produces a single dendogram (**Phenogram**). The objective is **classification not phylogeny**

(ii) **Statistical Approaches** developed around human blood groups data sets.

- **Edwards** and **Cavalli-Sforza** worked on the spatial representation of human gene frequencies differences, developed the **Minimum Evolution** and the **Least Square distance methods**, respectively. In order to reconcile results, they worked out an impractical **Maximum Likelihood method** and found that it was not equivalent to either of their two methods! Indeed, they discussed similarities between a **Maximum Parsimony method** and likelihood



Ronald Fisher

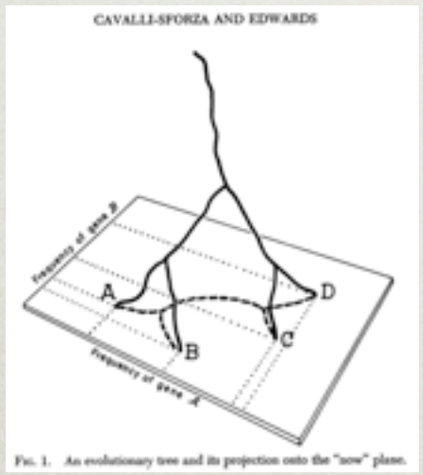


FIG. 1. An evolutionary tree and its projection onto the "now" plane.



Luigi L. Cavalli-Sforza

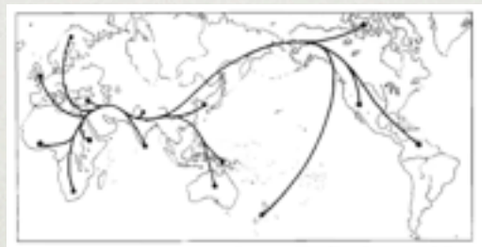


FIG. 1. Topology of the minimum-evolution tree uniting fifteen human populations; constructed on the basis of the frequency of blood-group alleles.

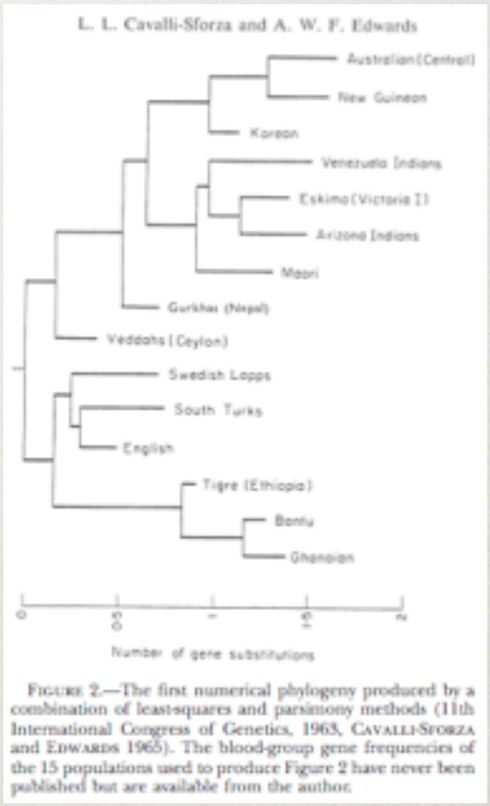
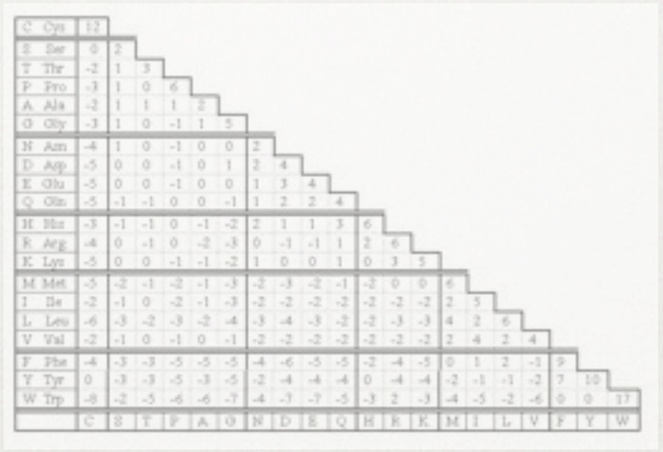


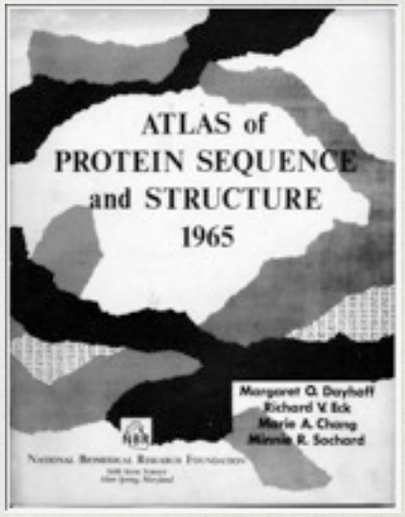
FIGURE 2.—The first numerical phylogeny produced by a combination of least-squares and parsimony methods (11th International Congress of Genetics, 1963, CAVALLI-SFORZA and EDWARDS 1965). The blood-group gene frequencies of the 15 populations used to produce Figure 2 have never been published but are available from the author.

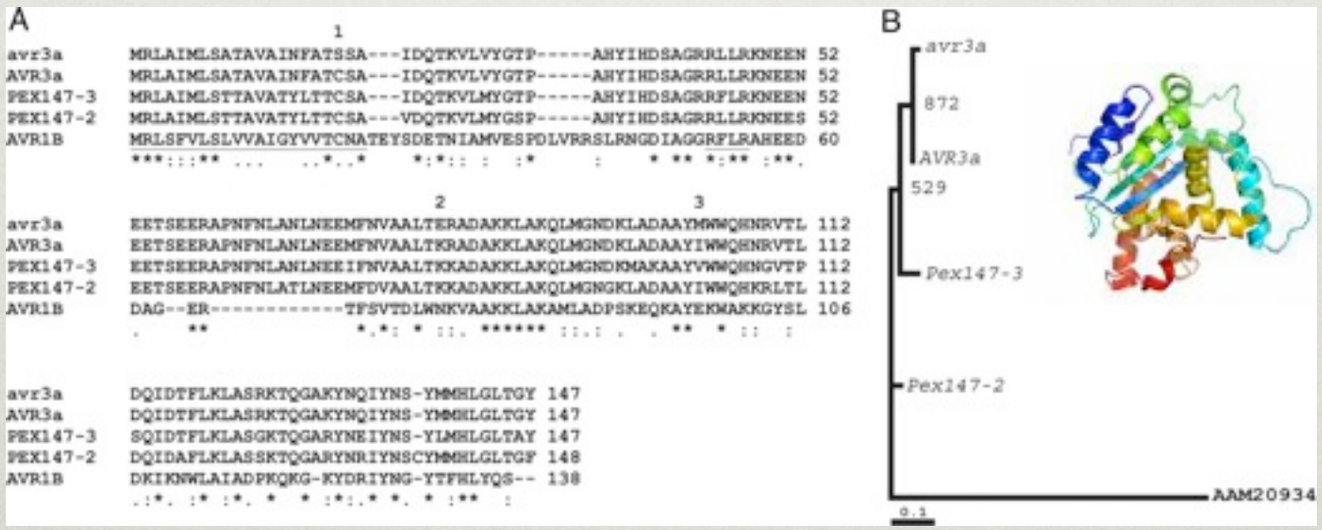
L. L. Cavalli-Sforza and A. W. F. Edwards. Analysis of human evolution. In Genetics Today. Proceeding of the XI International Congress of Genetics, The Hague, The Netherlands., volume 3, pages 923–933. Pergamon Press, Oxford, 1965.

- In the 1960s the molecular sequence data was mostly proteins. **Margaret Dayhoff** began to accumulate in the first molecular database! produced in a printed form. In the second edition of the "Atlas..." **they describe the first molecular parsimony method**, based on a model in which each of the 20 amino acids was allowed to change to any of the 19 others in a single step (**unordered method**).









- She originated one of the first substitution matrices, Point accepted mutations or (PAM). The one-letter code used for amino acids was developed by her, reflecting an attempt to reduce the size of the data files used to describe amino acid sequences in an era of punch-card computing.

- Although distance methods were first described by Edwards and Cavalli-Sforza, **Walter Fitch** and **E. Margoliash** *(Construction of phylogenetic trees: a method based on mutation distances as estimated from cytochrome c sequences is of general applicability. Science, 155:279–284, 1967)* **popularised distance matrix methods based on least squares**. The distances were fractions of amino acids differences between a particular pair of sequences. **The least squares was weighted with a greater observed distance given less weight**. This introduces the concept that large distances would be more prone to random error owing to the stochasticity of evolution.



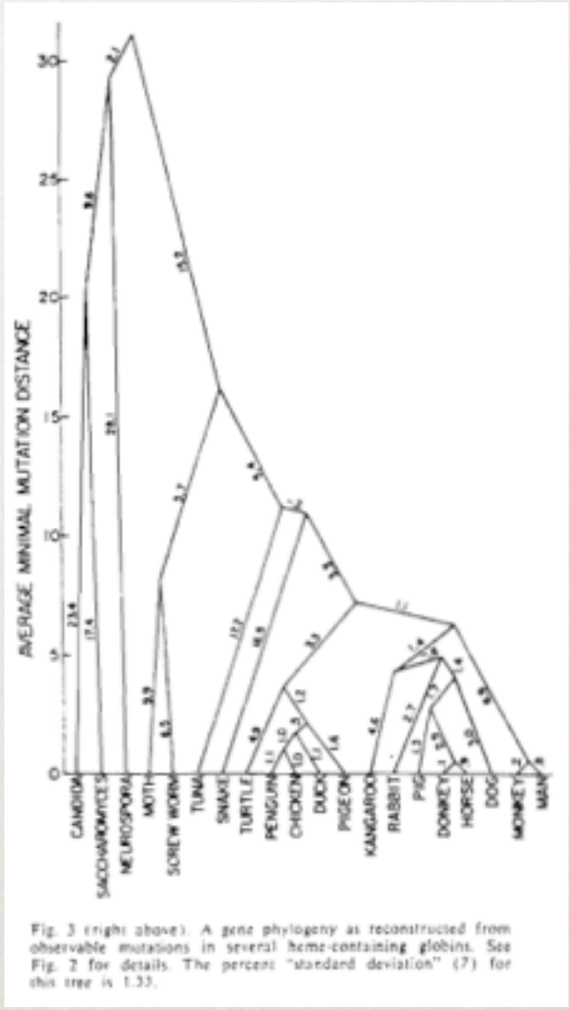


Fig. 3 (right above). A gene phylogeny as reconstructed from observable mutations in several heme-containing globins. See Fig. 2 for details. The percent "standard deviation" (7) for this tree is 1.33.

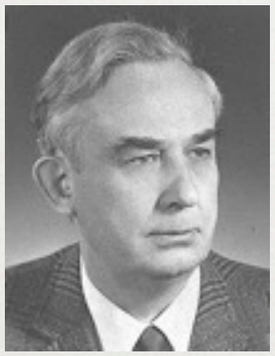
- Explicit models of sequence evolution correcting the effects of multiple replacement was first implemented by **Jukes & Cantor** in 1969

(iii) **Phylogenetic Systematic school** (**Cladistics**) originated by:

- **Willi Hennig** in Germany. *(Grundzu ¨geeinertheoriederphylogenetischensystematik. Deutscher Zentralverlag, Berlin, 1950. W. Hennig. Phylogenetic systematics. University of Illinois Press, Urbana, 1966);* and followed **in USA** by **W. H. Wagner Jr.**, **A. G. Kluge** and **J. S. Farris**. No numerical methods postulated from them. Camin, Edwards, and Dayhoff elaborated numerical methods for parsimony..
- **Osvaldo Reig** translates the book to Spanish (*Elementos de una Sistemática Filogenética. Ed Eudeba, Buenos Aires, 1966*). Many spanish scientists learnt cladism from this book



Joseph Camin



Willi Hennig



Warren Wagner



James Farris

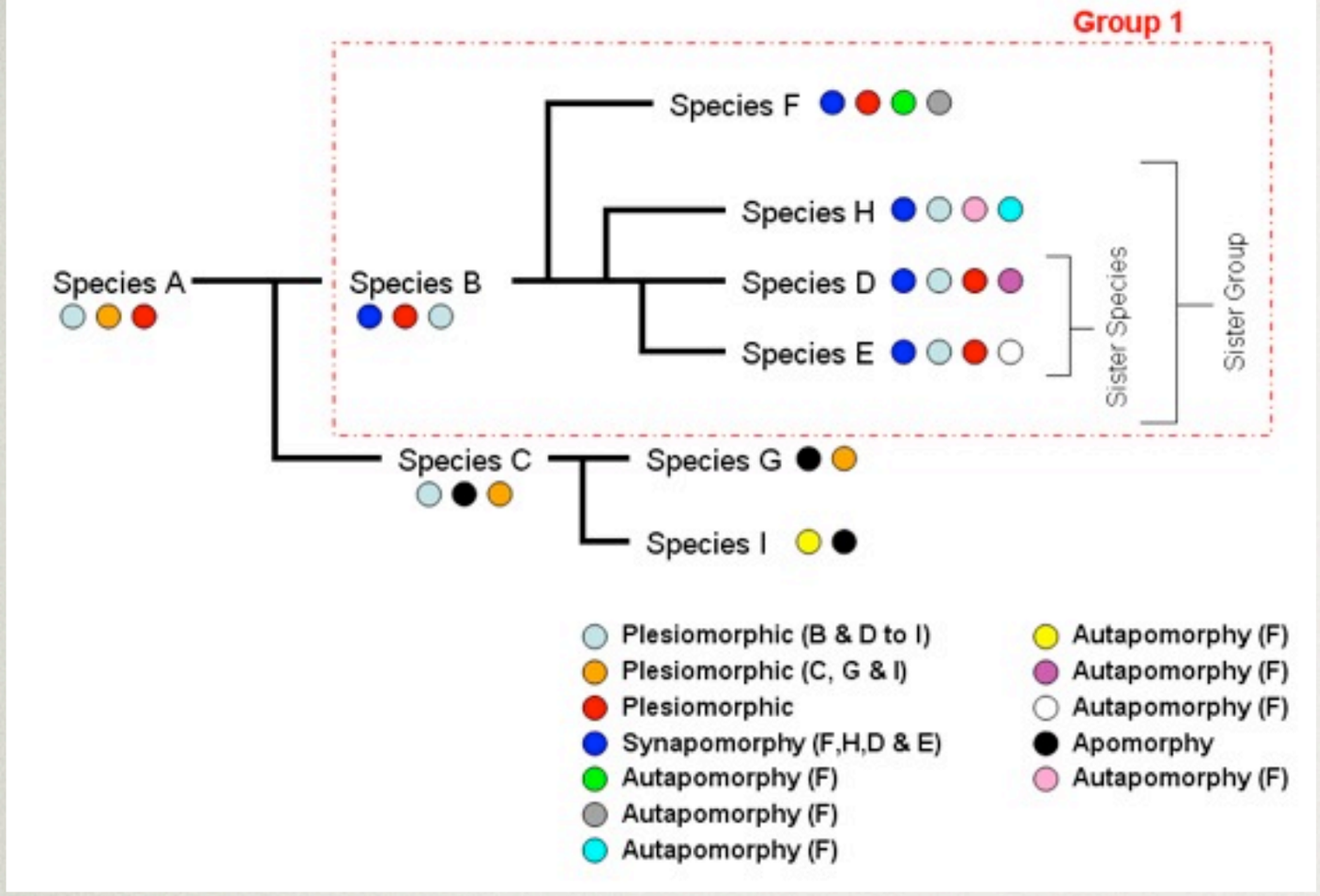


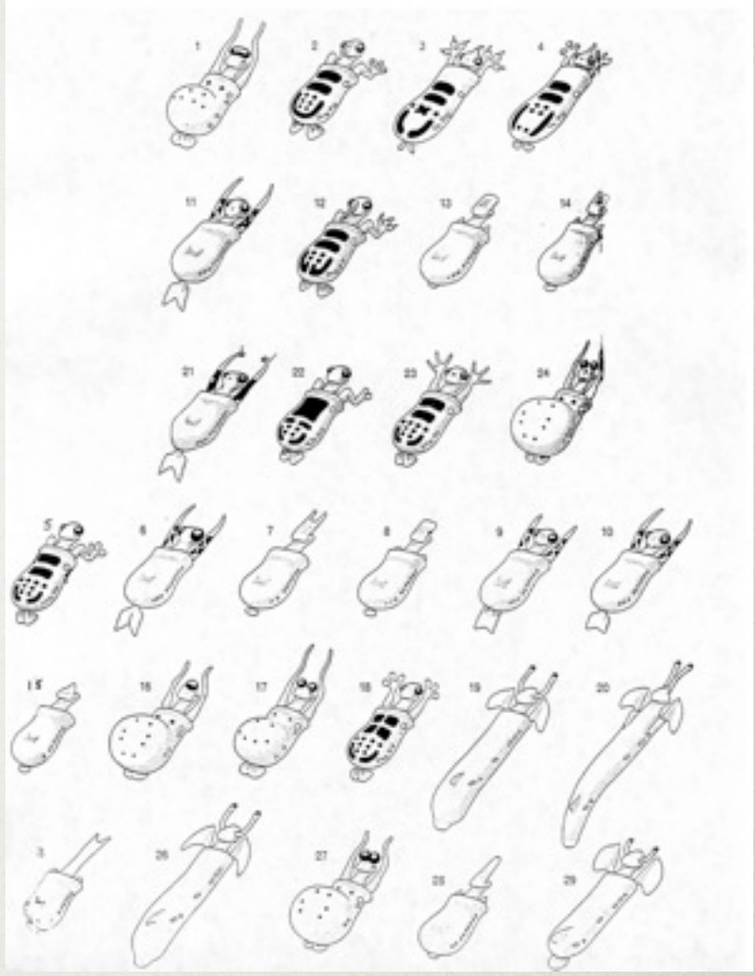
Osvaldo Reig

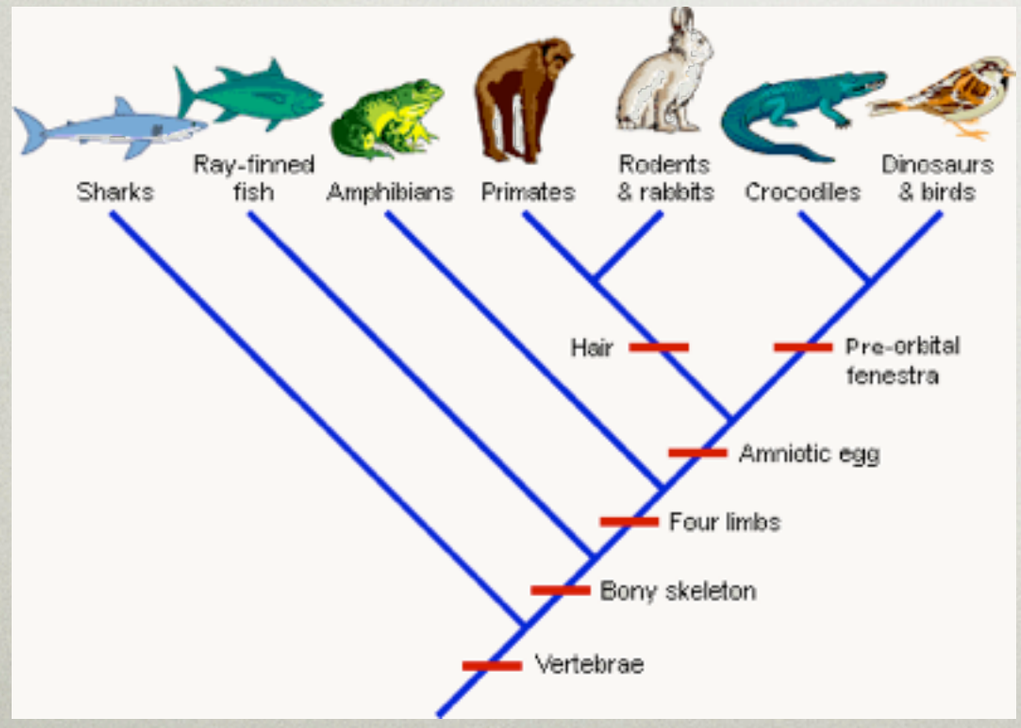
Main idea

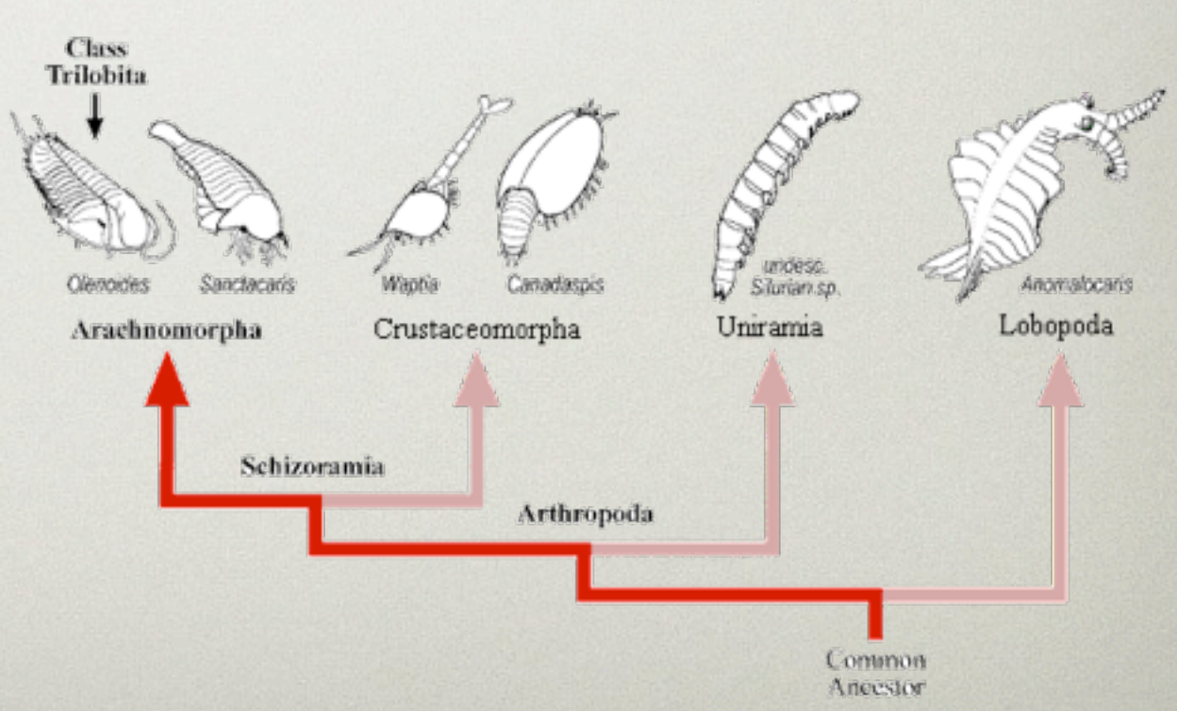
To use recency of common ancestry to construct hierarchies of relationship, NOT similarity. Relationships depicted by phylogenetic tree show sequence of speciation events (**Cladogram**). **Hennig's auxiliary principle** states that homology should be presumed in the absence of evidence to the contrary. Homoplasy ought not be postulated beyond necessity, that is to say parsimony.

- Felsenstein (2004) asserts that although Edwards and Cavalli-Sforza introduced parsimony, modern work on it springs from the paper of Camin and Sokal (1965)


Group 1
Species A
Species B
Species C
Species F
Species H
Species D
Species E
Species G
Species I
Sister Species
Sister Group
Plesiomorphic (B & D to I)
Plesiomorphic (C, G & I)
Plesiomorphic
Synapomorphy (F,H,D & E)
Autapomorphy (F)
Autapomorphy (F)
Autapomorphy (F)
Autapomorphy (F)
Autapomorphy (F)
Apomorphy
Autapomorphy (F)




Sharks
Ray-finned fish
Amphibians
Primates
Rodents & rabbits
Crocodiles
Dinosaurs & birds
Hair
Pre-orbital fenestra
Amniotic egg
Four limbs
Bony skeleton
Vertebrae


Class Trilobita
Olenoides
Sanctacaris
Arachnomorpha
Waptia
Canadaspis
Crustaceomorpha
undesc. Silurian sp.
Uniramia
Anomalocaris
Lobopoda
Schizoramia
Arthropoda
Common Ancestor

Applications of phylogenetics

Phylogenetic information is used in different areas of biology. From population genetics to macroevolutionary studies, from epidemiology to animal behaviour, from forensic practice to conservation ecology. In spite of this broad range of applications, phylogenies are used by making inferences from:

1. **Tree topology and branch lengths:**
   - Applications in evolutionary genetics deducing partial internal duplication of genes, recombination, reassortment, gene conversion, translocations or, xenology.
   - Applications in population genetics[1] in order to quantify parameters and processes like gene flow, mutation rate, population size, natural selection and speciation
   - Applications by estimating rates and dates in order to check clock- like behaviour of genes, to date events in epidemiological studies , or macroevolutionary events.
   - Applications by testing hypotheses of coevolution, co-speciation , biogeography, molecular adaptation, neutrality, convergence, tissue tropisms (HIV clones), the origin of genetic code, stress effects in bacteria, etc.
   - Applications in conservation biology, forensic or legal cases, and the list is far less than exhaustive!!!

2. **Mapping character states on to a known tree:**
   - Applications in comparative biology, in areas of animal behaviour, development, speciation and adaptation

---

[1] Coalescence theory brings many possibilities to test for different hypotheses associated to population genetic parameters.
See: Hein, J; Schierup, M. H., and Wiuf, C. Gene Genealogies, Variation and Evolution – A Primer in Coalescent Theory. Oxford University Press, 2005.

Developmental biology

Hox gene variation and evolution

Axel Meyer

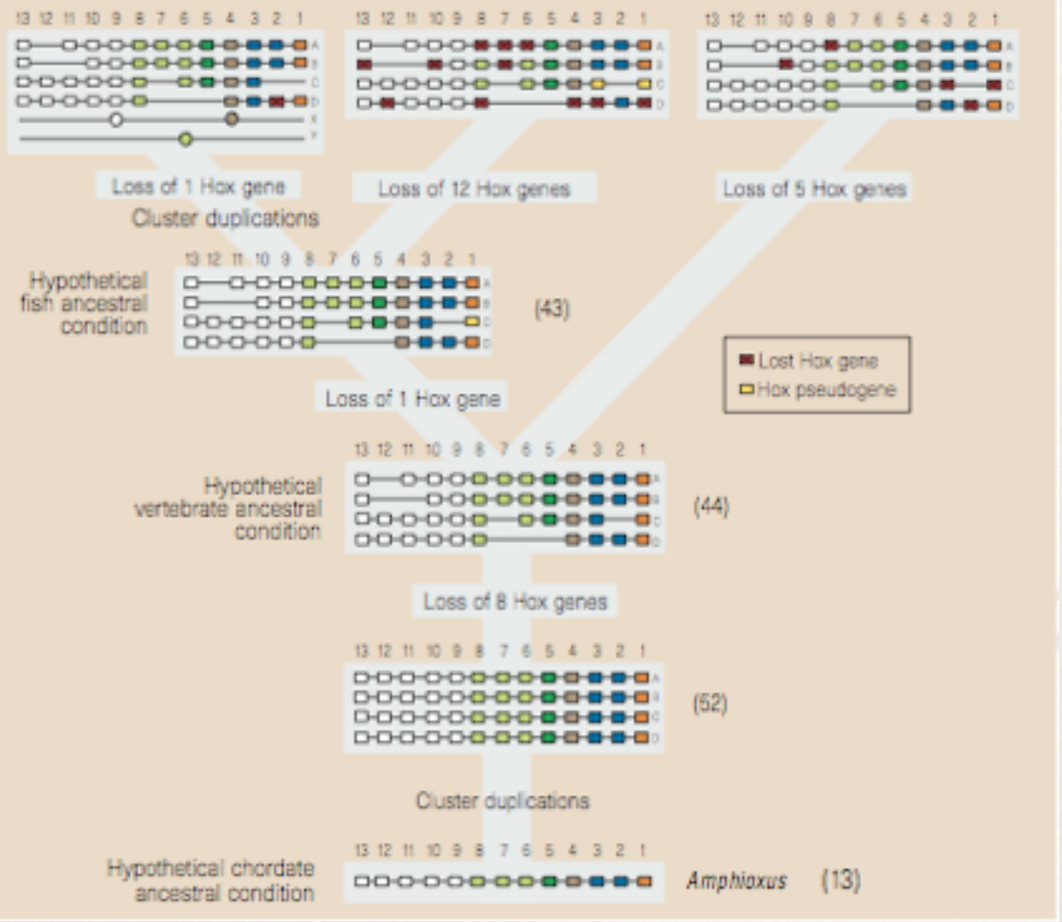


Figure 1 Evolutionary conservation of the Hox gene clusters in zebrafish, pufferfish, mouse and *Amphioxus*. The zebrafish Hox gene clusters are shown with the 34 genes confirmed by Prince *et al.*[1], and an additional eight, for the total of 42 that the zebrafish is likely to have. Ten gene losses differentiate the mouse from the pufferfish, but there are only three losses between the mouse and zebrafish, even though fish and tetrapods shared their last common ancestor about 400 million years ago. The common ancestor of human and mouse probably lived around 60 million years ago, but the architectures of their Hox gene clusters are identical. The losses in the *AbdA (abdominal A)*-related genes (light green) in the D-cluster probably occurred early in the history of vertebrates. (Figure drawn by Heike Haunstetter.)

Convergent evolution of behavior in an adaptive radiation of Hawaiian web-building spiders

Todd A. Blackledge† and Rosemary G. Gillespie

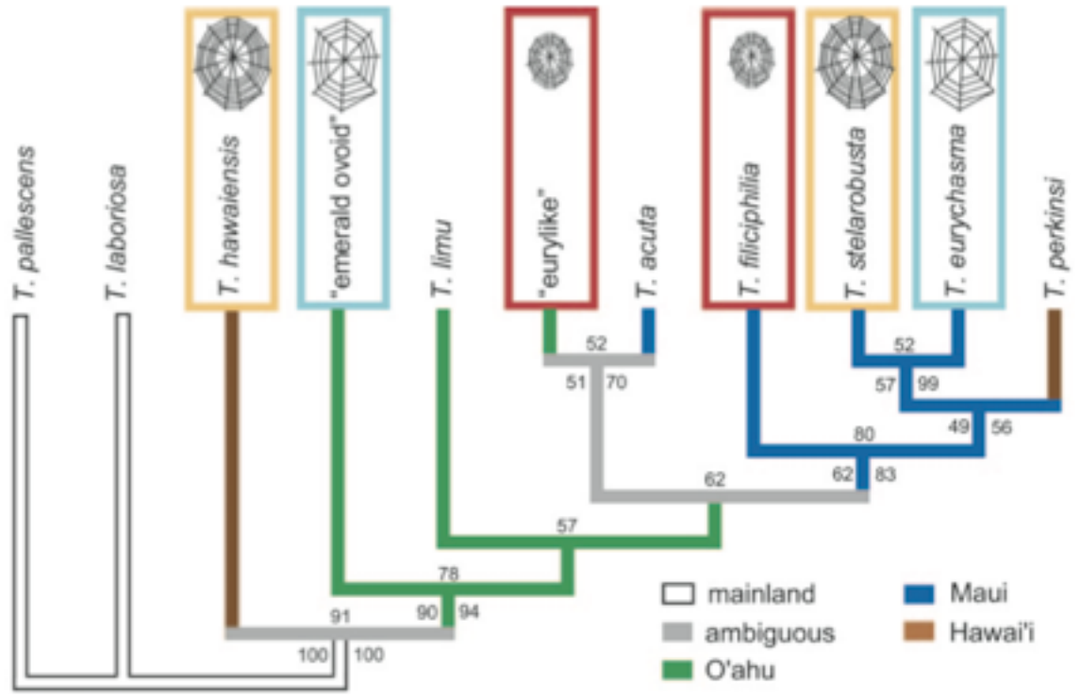


Fig. 4. Phylogenetic analysis of the historical relationships among species. Bootstrap values from the maximum parsimony analysis are presented above each node, bootstrap values from the maximum likelihood analysis are presented to the left of each node, and clade posterior probabilities from the Bayesian inference analysis are presented to the right of each internode. Colors of the branches denote different islands optimized as unordered and unpolarized character states. Web illustrations denote the three ethotypes and are identified by the same colors as in Fig. 2. Members of each ethotype arose independently of one another in at least two of the three instances.

Ancient origin of the integrin-mediated adhesion and signaling machinery

Arnau Sebé-Pedrós[a], Andrew J. Roger[b], Franz B. Lang[c], Nicole King[d], and Iñaki Ruiz-Trillo[a,e,1]



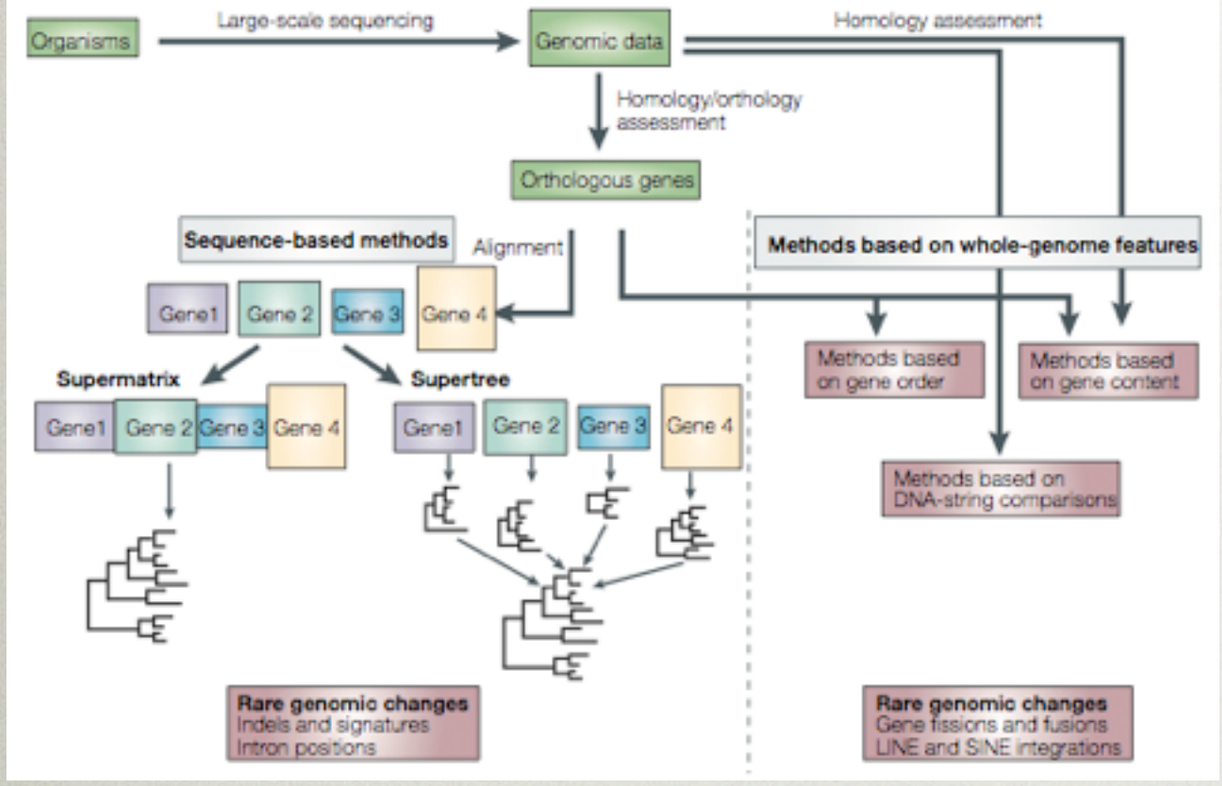
Fig. 3. Schematic representation of integrin-mediated cell-adhesion and cell-signaling evolution. *Left:* The canonical metazoan integrin adhesion complex. The colors correspond to the three main steps in the evolution of the integrin adhesion mechanism, as shown in the cladogram. Dots indicate origin, and crosses indicate losses. The branch leading to *Amastigomonas* sp. is shown dashed, because its phylogenetic position remains unresolved (see main text for discussion).

Phylogenomics

PHYLOGENOMICS AND THE RECONSTRUCTION OF THE TREE OF LIFE

Frédéric Delsuc, Henner Brinkmann and Hervé Philippe

Box 2 | **Methods of phylogenomic inference**

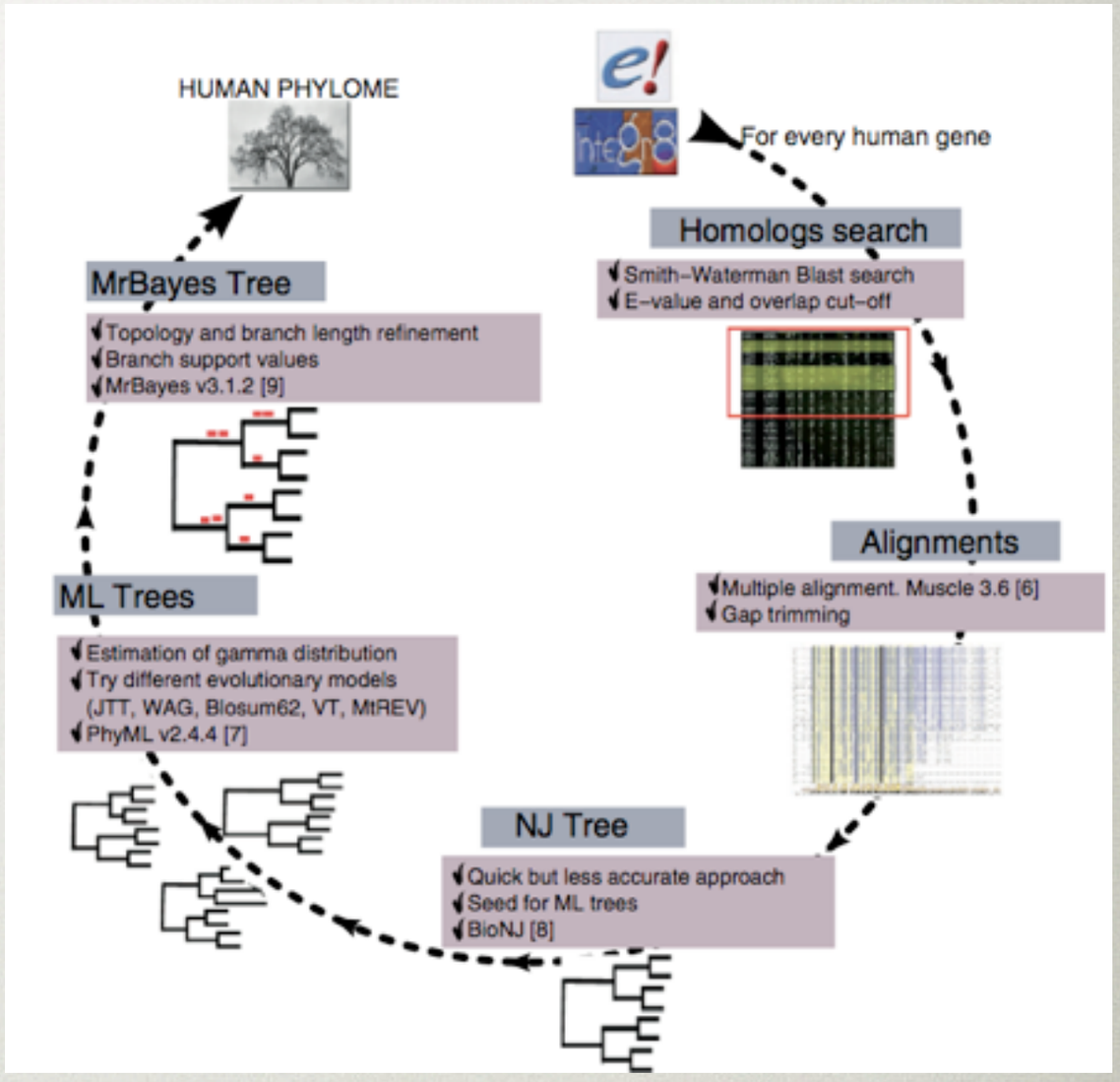


Research

Open Access

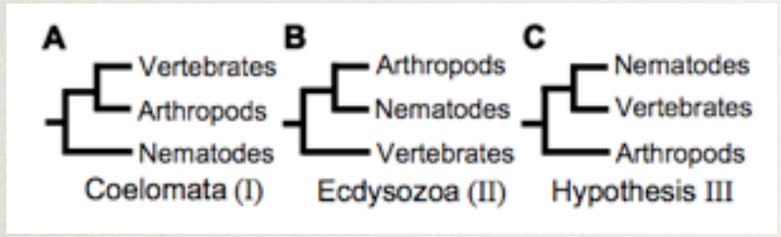
The human phylome

Jaime Huerta-Cepas, Hernán Dopazo, Joaquín Dopazo and Toni Gabaldón



Phylogenomics

- Genome scale phylogenetic analysis to be applied in :
  - ✳ ***Systematic problems***
    - Testing the new animal phylogeny, ecdysozoa (arthropods + nematodes) vs coelomata (vertebrates + arthropods). Phylogenetic relationships among *H. sapiens, D.melanogaster* and *C. elegans* were unsolved.
    - They are model species with full genome sequenced. Single gene and phylogenomics results contradicts each other.



Multigene Analyses of Bilaterian Animals Corroborate the Monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia

Hervé Philippe, Nicolas Lartillot,[1] and Henner Brinkmann

Research

Open Access

Genome-scale evidence of the nematode-arthropod clade

Hernán Dopazo* and Joaquín Dopazo†‡

Rare Coding Sequence Changes are Consistent with Ecdysozoa, not Coelomata

Manuel Irimia,† Ignacio Maeso,† David Penny,‡ Jordi Garcia-Fernàndez,† and Scott William Roy*‡§*

Ecdysozoan Clade Rejected by Genome-Wide Analysis of Rare Amino Acid Replacements

Igor B. Rogozin, Yuri I. Wolf, Liran Carmel, and Eugene V. Koonin

The Opisthokonta and the Ecdysozoa May Not Be Clades: Stronger Support for the Grouping of Plant and Animal than for Animal and Fungi and Stronger Support for the Coelomata than Ecdysozoa

Gayle K. Philip,[1] Christopher J. Creevey,[1] and James O. McInerney

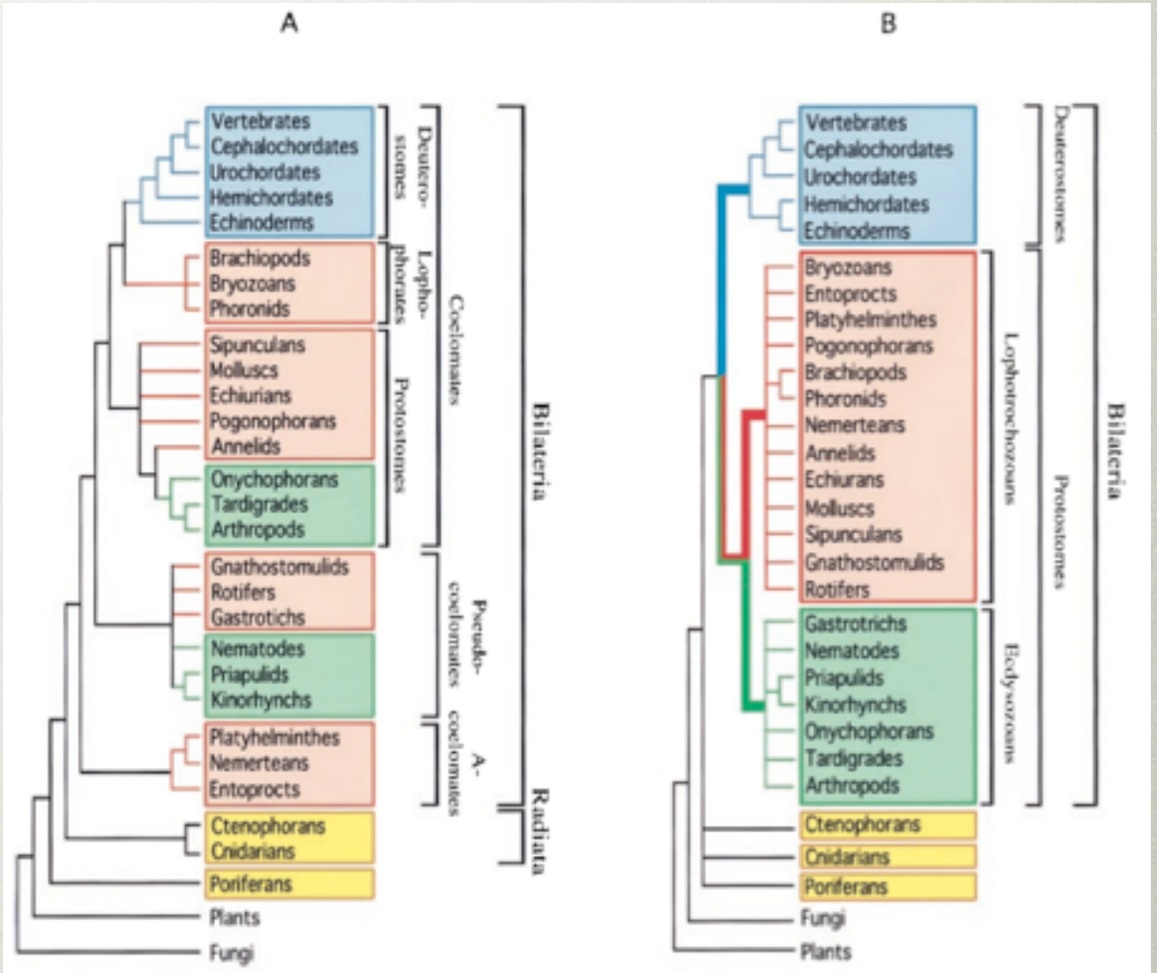
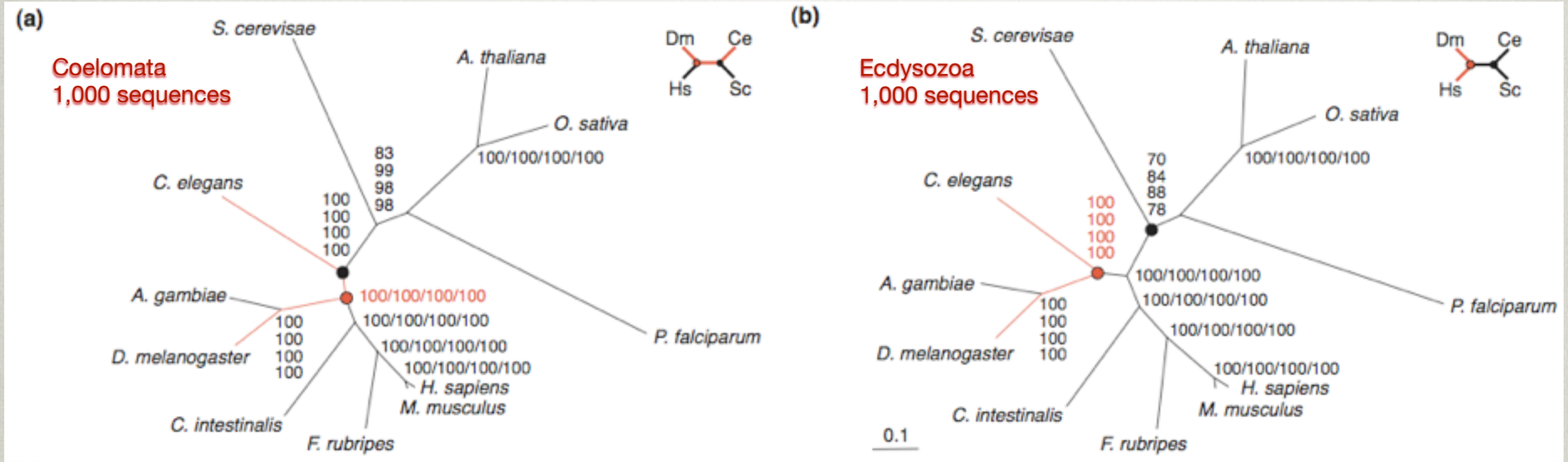


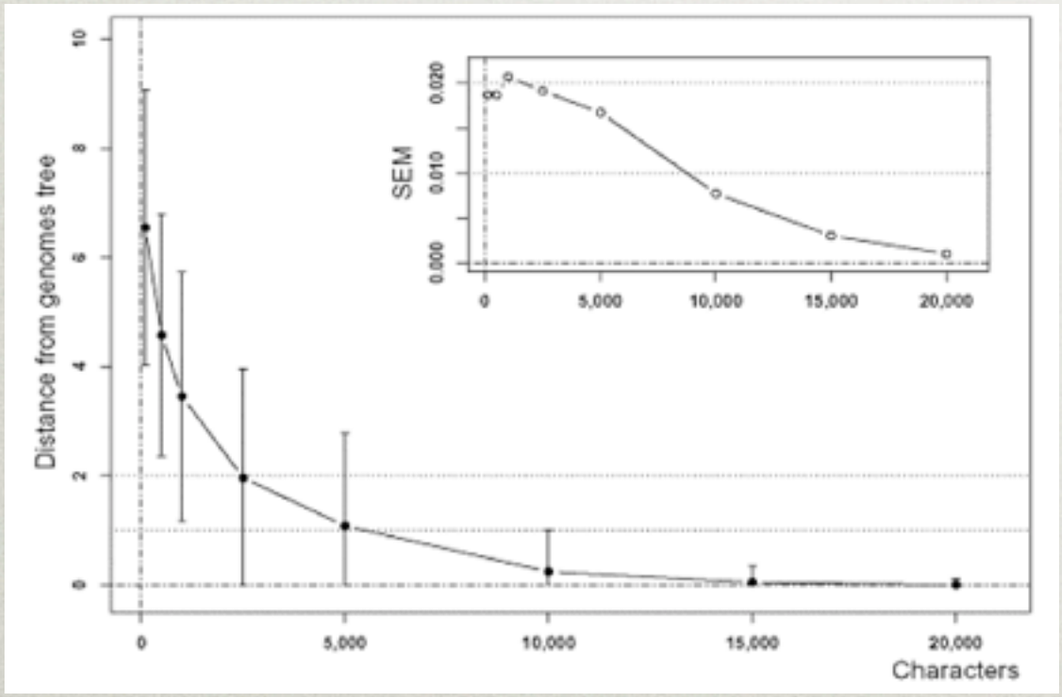
Fig. 1. Metazoan phylogenies. (*A*) The traditional phylogeny based on morphology and embryology, adapted from Hyman (11). (*B*) The new molecule-based phylogeny. A conservative approach was taken in *B*: i.e., some datasets provide resolution within some of the unresolved multifurcations displayed, but we have limited the extent of resolution displayed to that solidly provided by rRNA only.

The new animal phylogeny: Reliability and implications

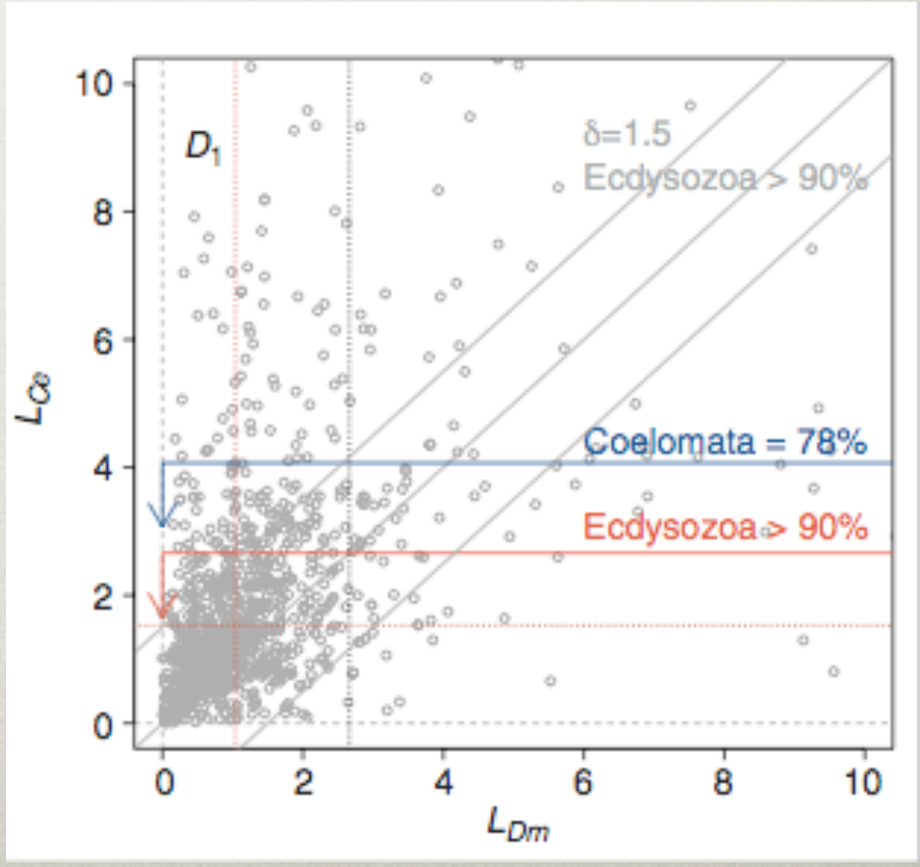
André Adoutte*, Guillaume Balavoine, Nicolas Lartillot, Olivier Lespinet, Benjamin Prud'homme, and Renaud de Rosa

Figure 3
Phylogenetic trees. Trees derived from M_1 and M_6 datasets, respectively support (a) the Coelomata and (b) the Ecdysozoa hypothesis. From left to right or top to bottom, values besides nodes show the maximum likelihood reliability values of the quartet-puzzling tree and bootstrap values using maximum likelihood, least squares, and neighbor-joining methods, respectively. Values in red show the support for (a) Coelomata and (b) Ecdysozoa nodes. Red branches display distances between *C. elegans* and *D. melanogaster*. Smaller trees are minimal representations of both hypothesis.



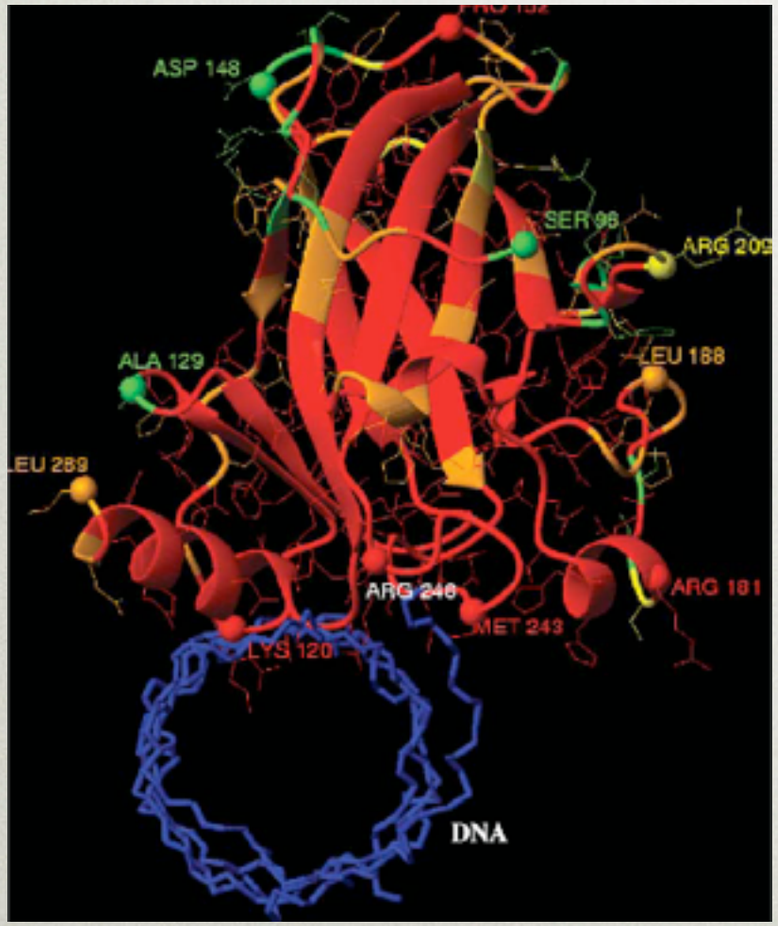


The use of a high number of characters gives strong support on trees, but sometimes it is not sufficient ... LBAE



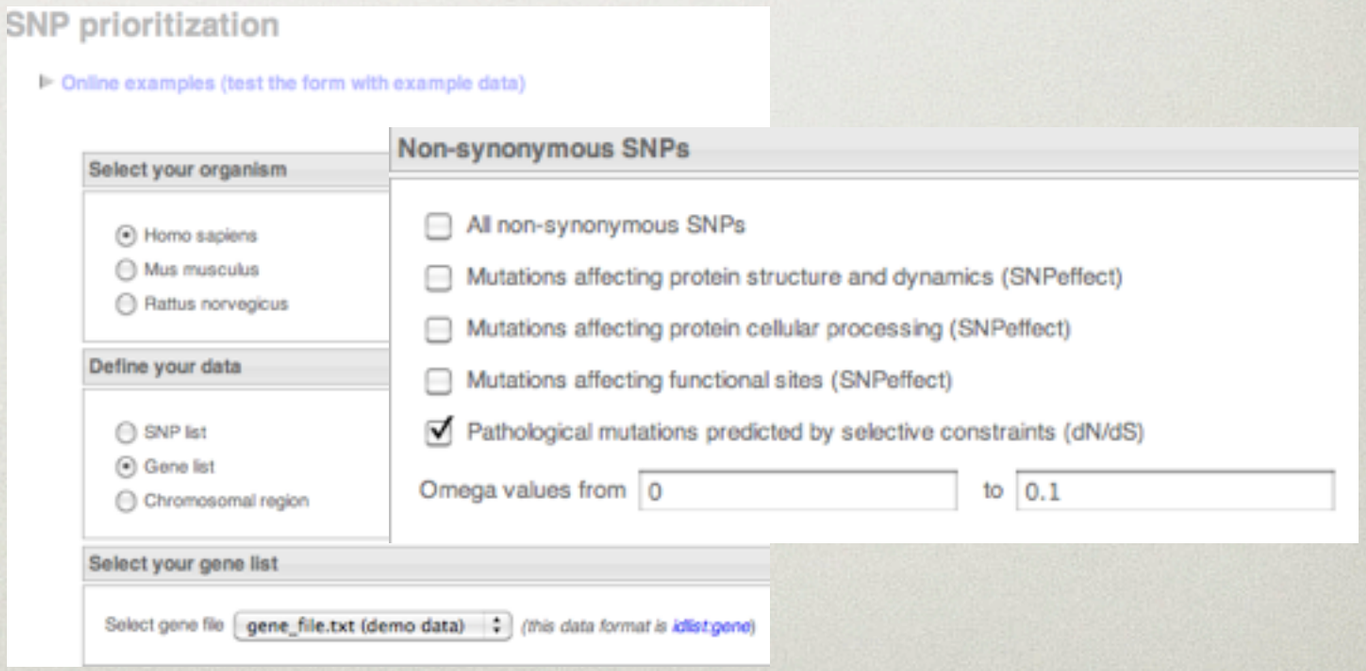
Phylogenomics

- Genome scale phylogenetic analysis to be applied in :
  - ✳ ***Medical and Bioinformatics problems***
    - Mapping selective constraints on protein codon sequences and
    - Apply the method to all the human, rat and mouse genome proteins



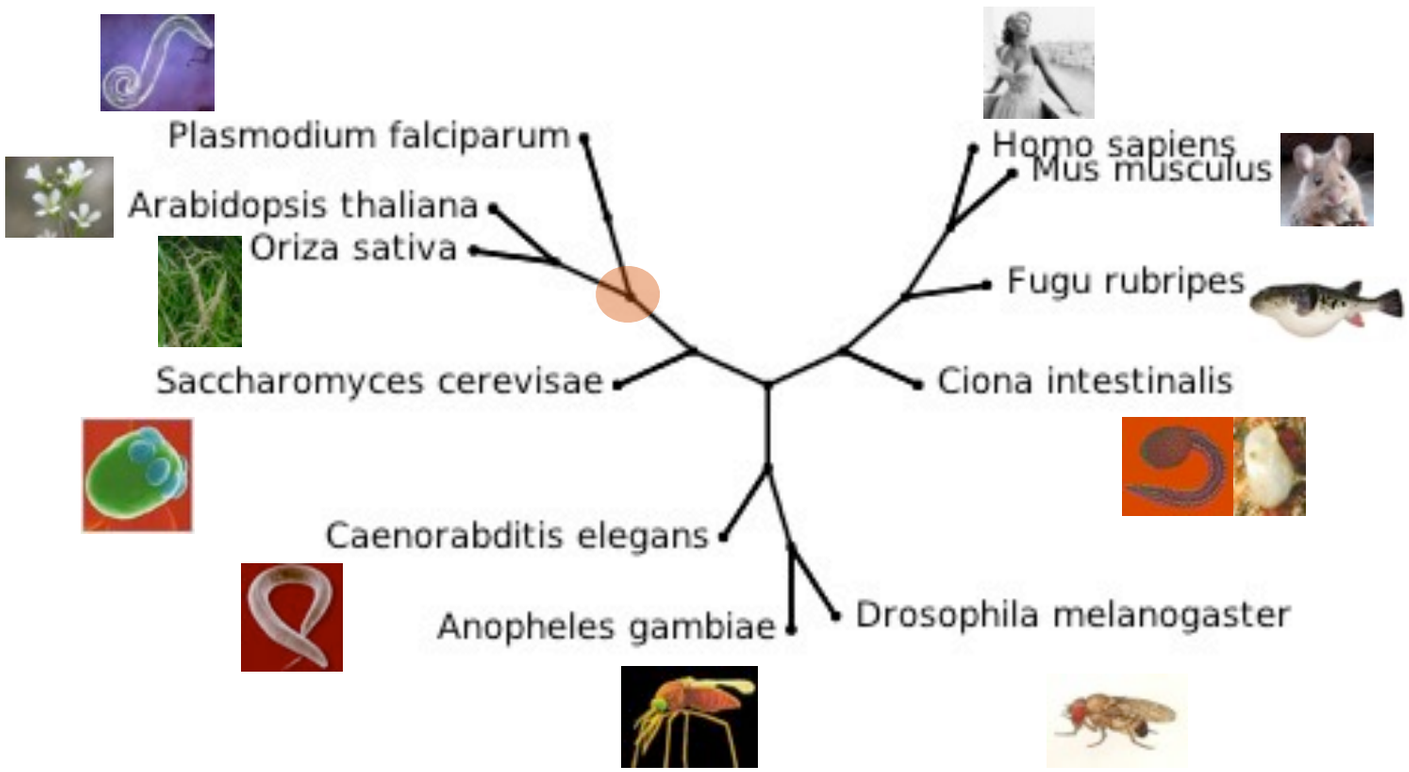
Pupasuite 3

prioritization of SNPs in genotyping experiments based on functional properties

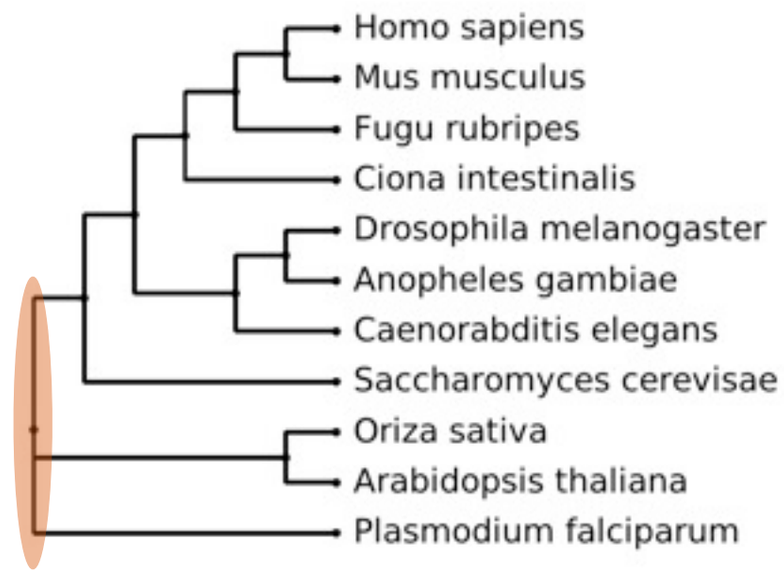


Cladograms, Phylograms & Tree Representations

Trees showing **branching order** exclusively (**cladogenesis**) are principally the interest of **systematists**[1] to make inferences on **taxonomy**[2]. Those interesting in the **full evolutionary processes** emphasise on **branch lengths** information also (**anagenesis**).



Unrooted tree topology



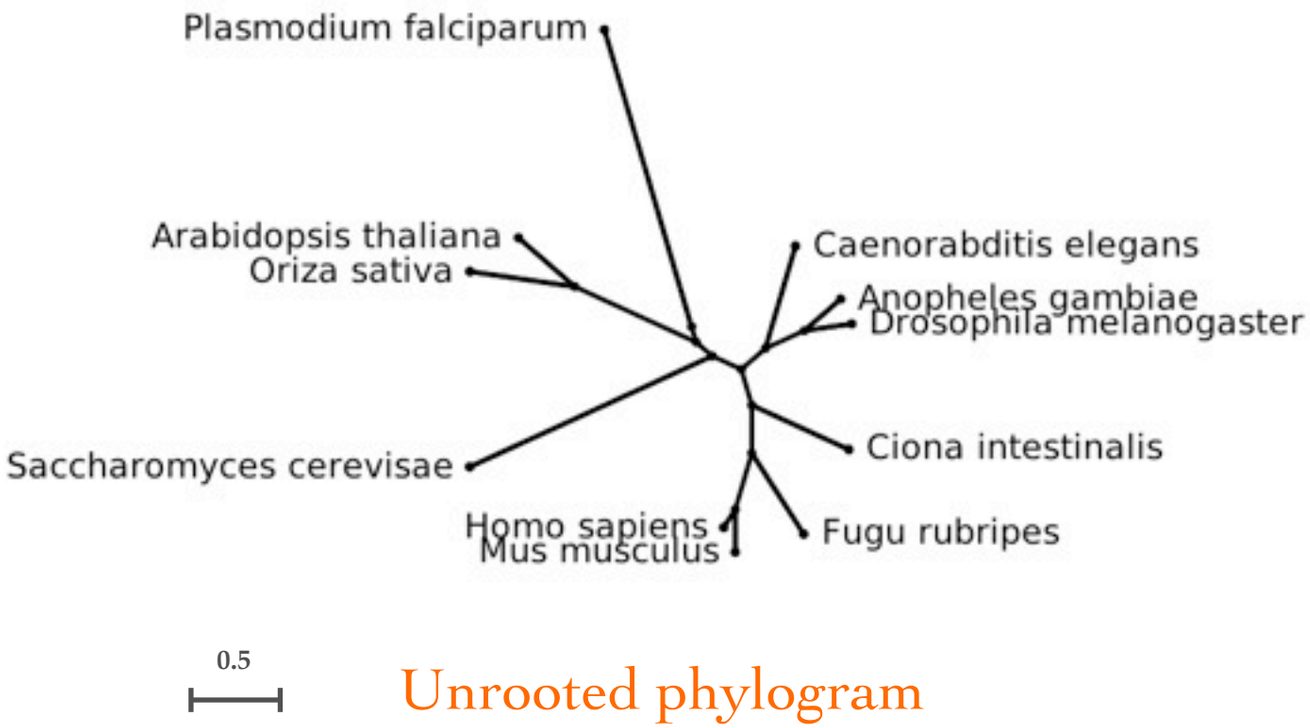
Unrooted tree topology

- **Cladogram** is a branching diagram depicting the hierarchical arrangement of taxa defined by **cladistic methods** (the distribution of shared derived characters -**synapomorphies**-)

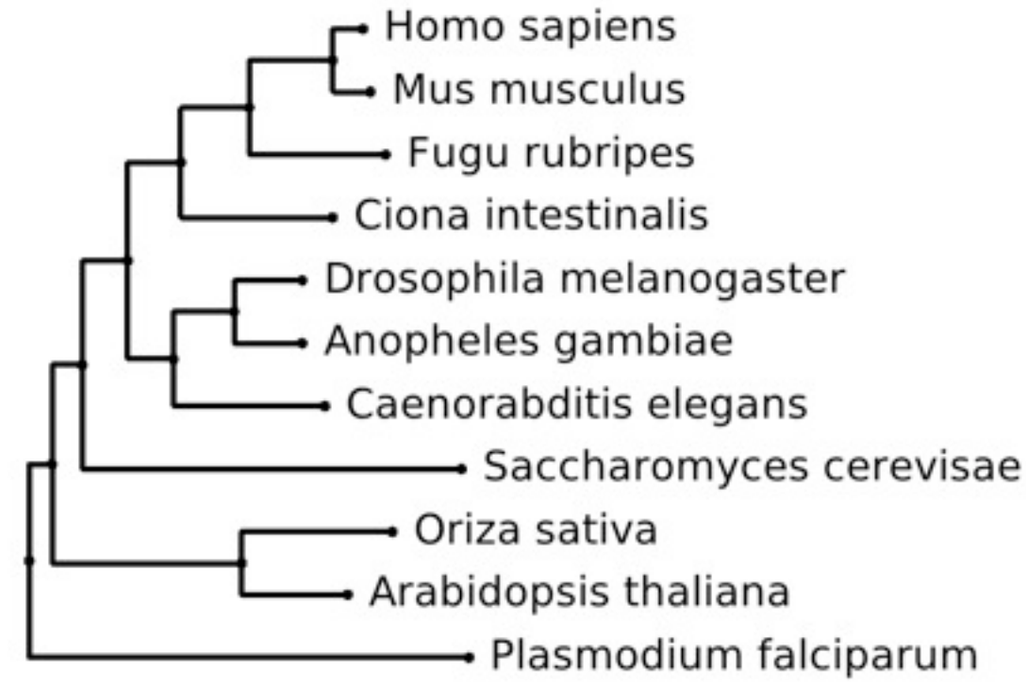
* **Phylogram** is a phylogenetic tree that indicates the relationships between the taxa **and** also conveys a sense of time or rate of evolution. The temporal aspect of a phylogram is missing in a cladogram.

---

[1] The study of biological diversity. [2] The theory and practice of describing, naming and classifying organisms

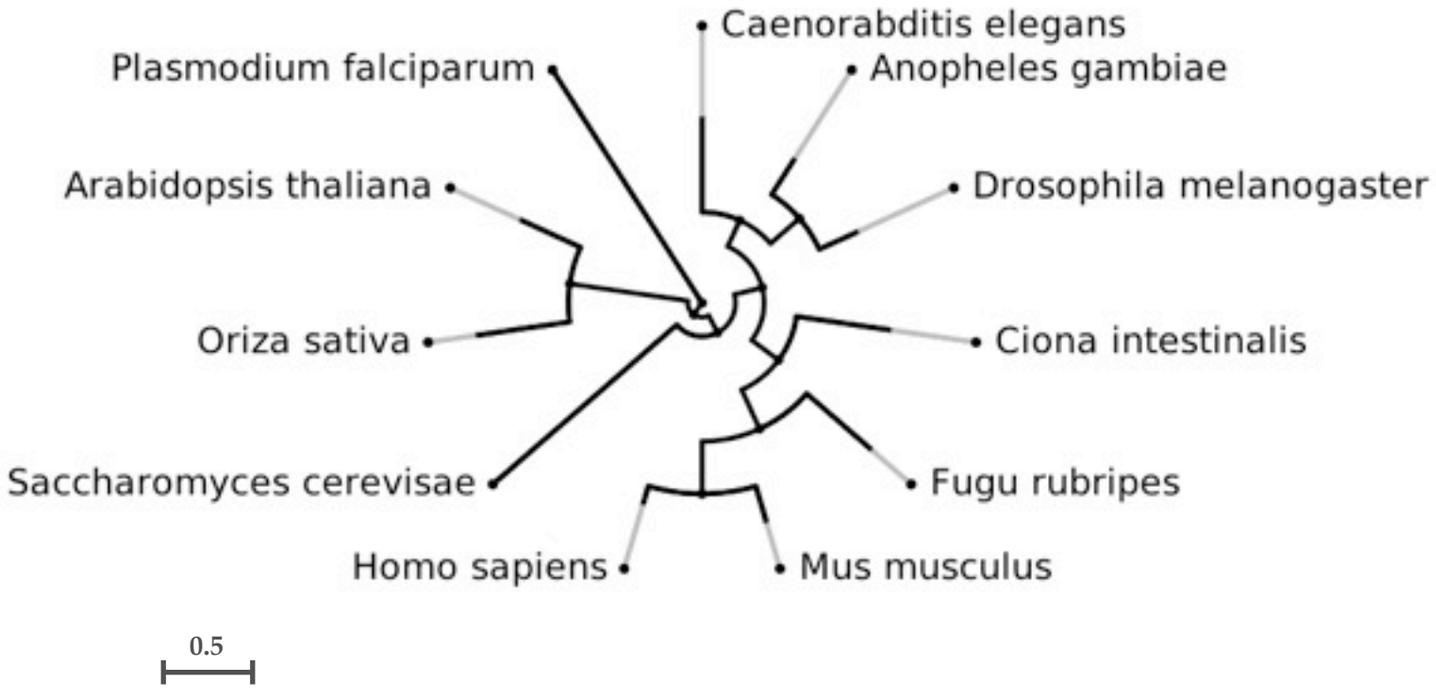


Unrooted phylogram

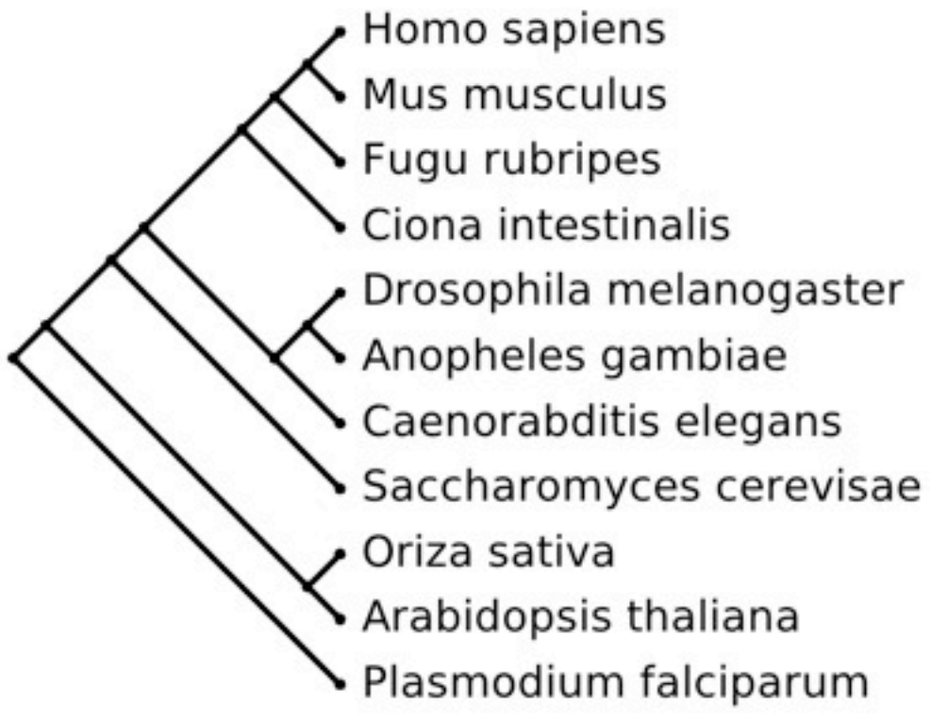


Rooted phylogram

- Distance scale represents the number of differences between sequences (e.g. 0.1 means 10 % differences)



Circular phylogram

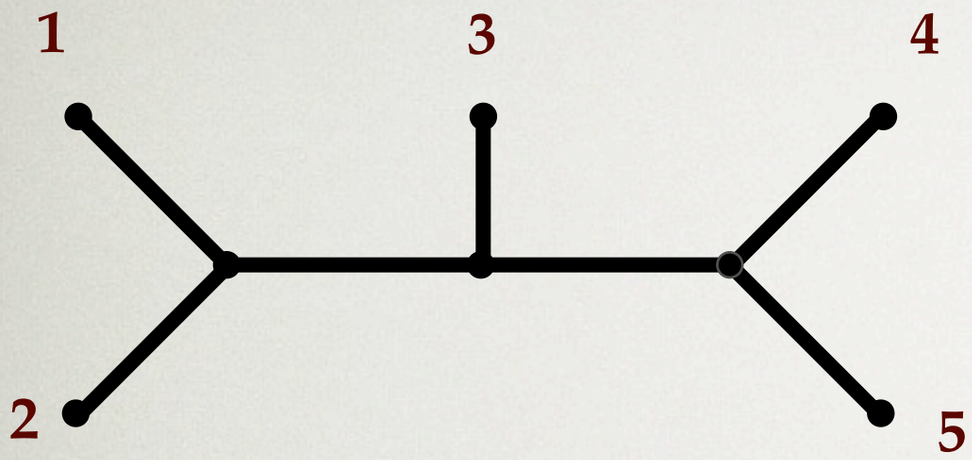


Triangular tree representation

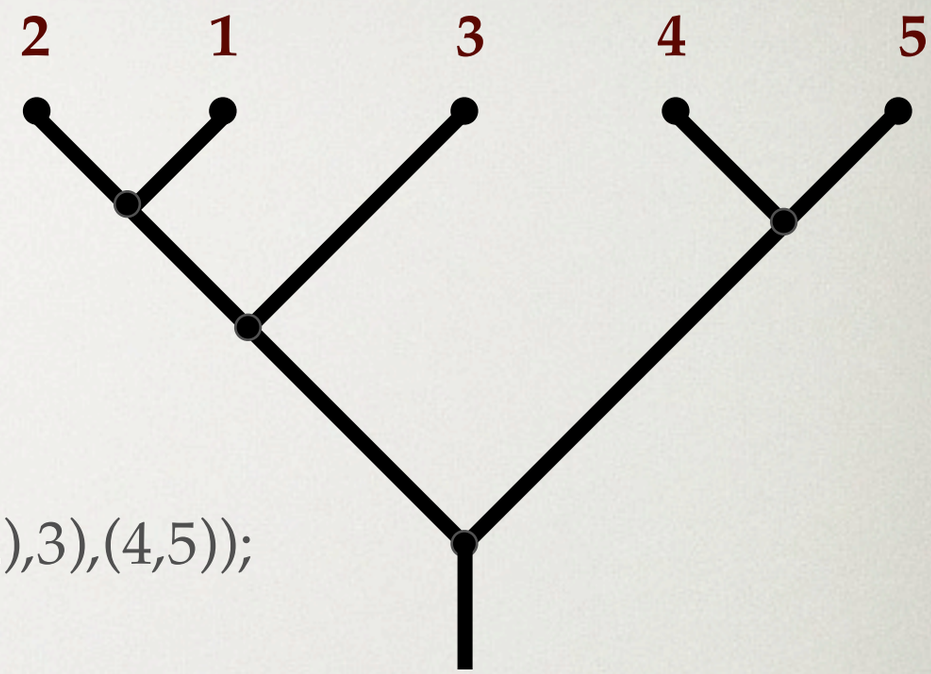
cladogram or phylogram

No distance information

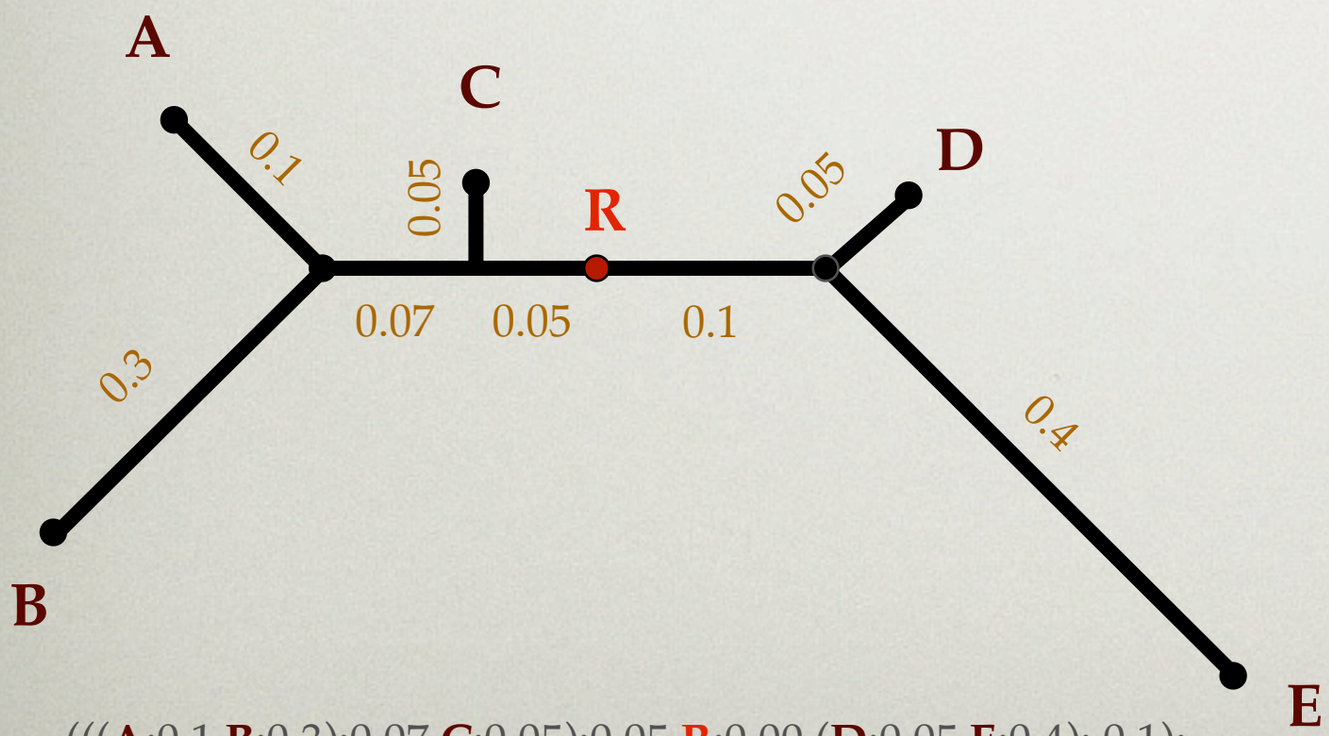
Newick Format & Visualisation Programs



((((1,2),3),4,5);



((((1,2),3),(4,5));



(((**A**:0.1,**B**:0.3):0.07,**C**:0.05):0.05,**R**:0.00,(**D**:0.05,**E**:0.4): 0.1);

E.T.E.

A python Environment for Tree Exploration

Archaeopteryx

Archaeopteryx Evolutionary Tree Visualization and Analysis Software

TreeView

FigTree

Homology





The Archetype Concept

* **Richard Owen**'s (1847) most famous contributions to theoretical comparative anatomy was to distinguish between homologous and analogous features in organisms and to present the **concept of archetype**.

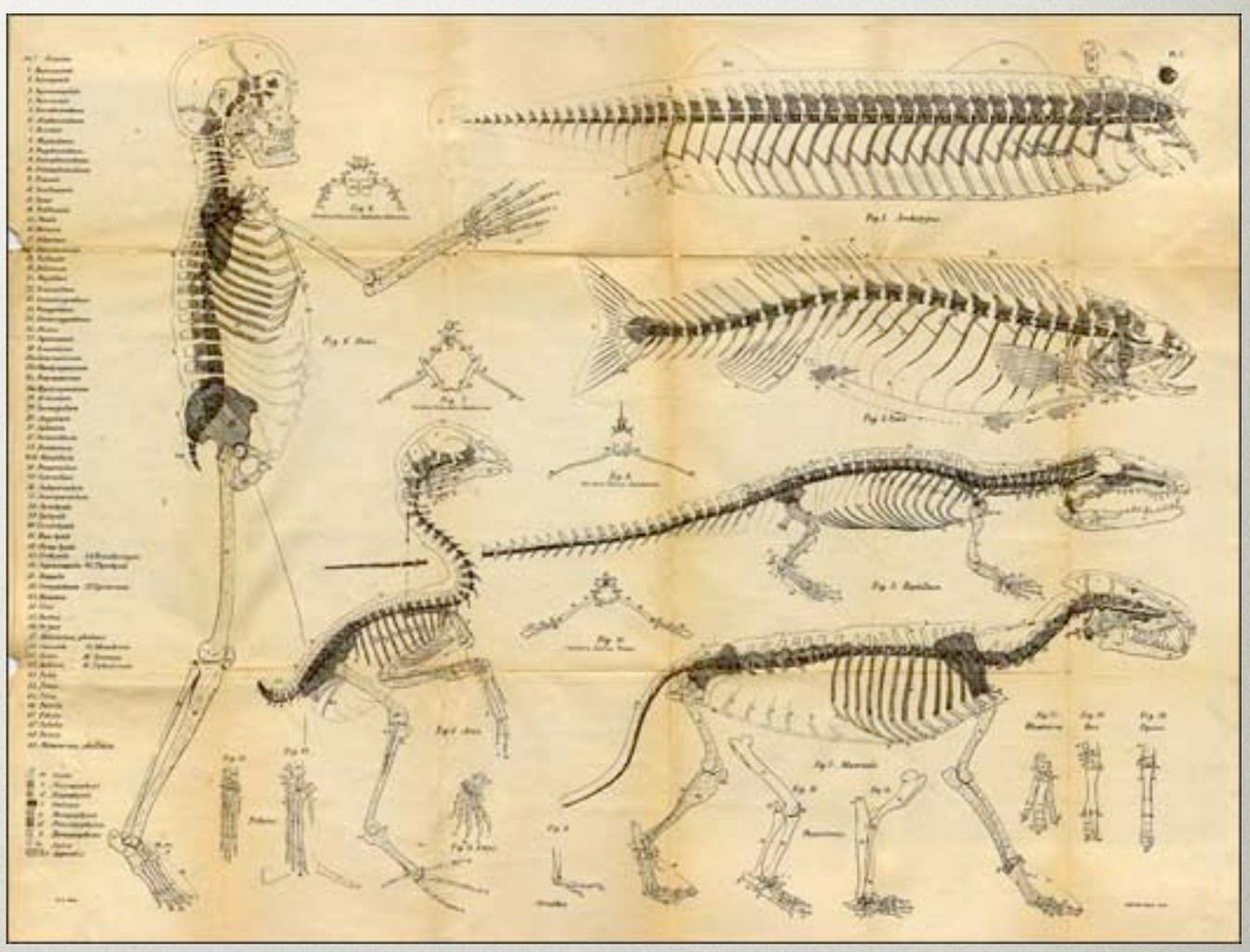
* The **vertebrate archetype** consists of a linear series of "vertebrae" and "appendages", little modified from a single basic plan.

* Each vertebra of the archetype is a **serial homologue** of every other vertebra of the archetype.

* Two corresponding vertebrae, each from different animal, are **special homologues** of one another, and **general homologues** of the corresponding vertebra of the archetype.

* **Homologue**..."*The same organ in different animals under every variety of form and function*".

* **Analogue**..."*A part or organ in one animal which has the same function as another part or organ in a different animal*".

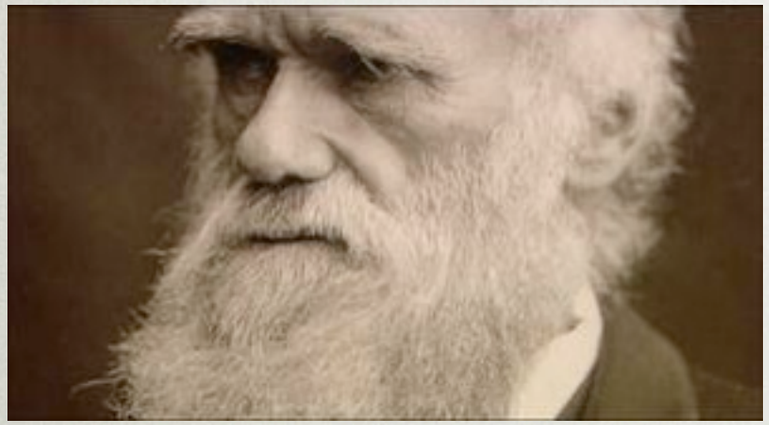


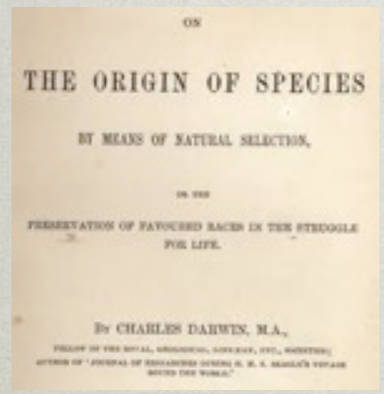
Darwin & Homology. The Origin of Species. Chapter 14.

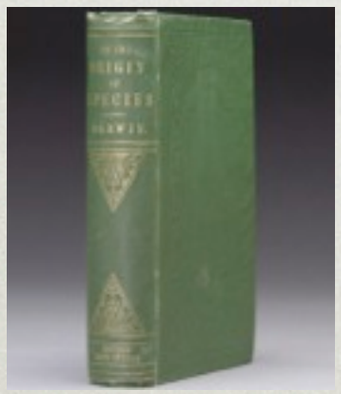
* *"What can be more curious than that the hand of a man, formed for grasping, that of a mole for digging, the leg of the horse, the paddle of the porpoise, and the wing of the bat, should all be constructed on the same pattern, and should include similar bones, in the same relative positions?"*

* *"How inexplicable are the cases of serial homologies on the ordinary view of creation!"*

* *"Why should similar bones have been created to form the wing and the leg of a bat, used as they are for such totally different purposes, namely flying and walking?"*







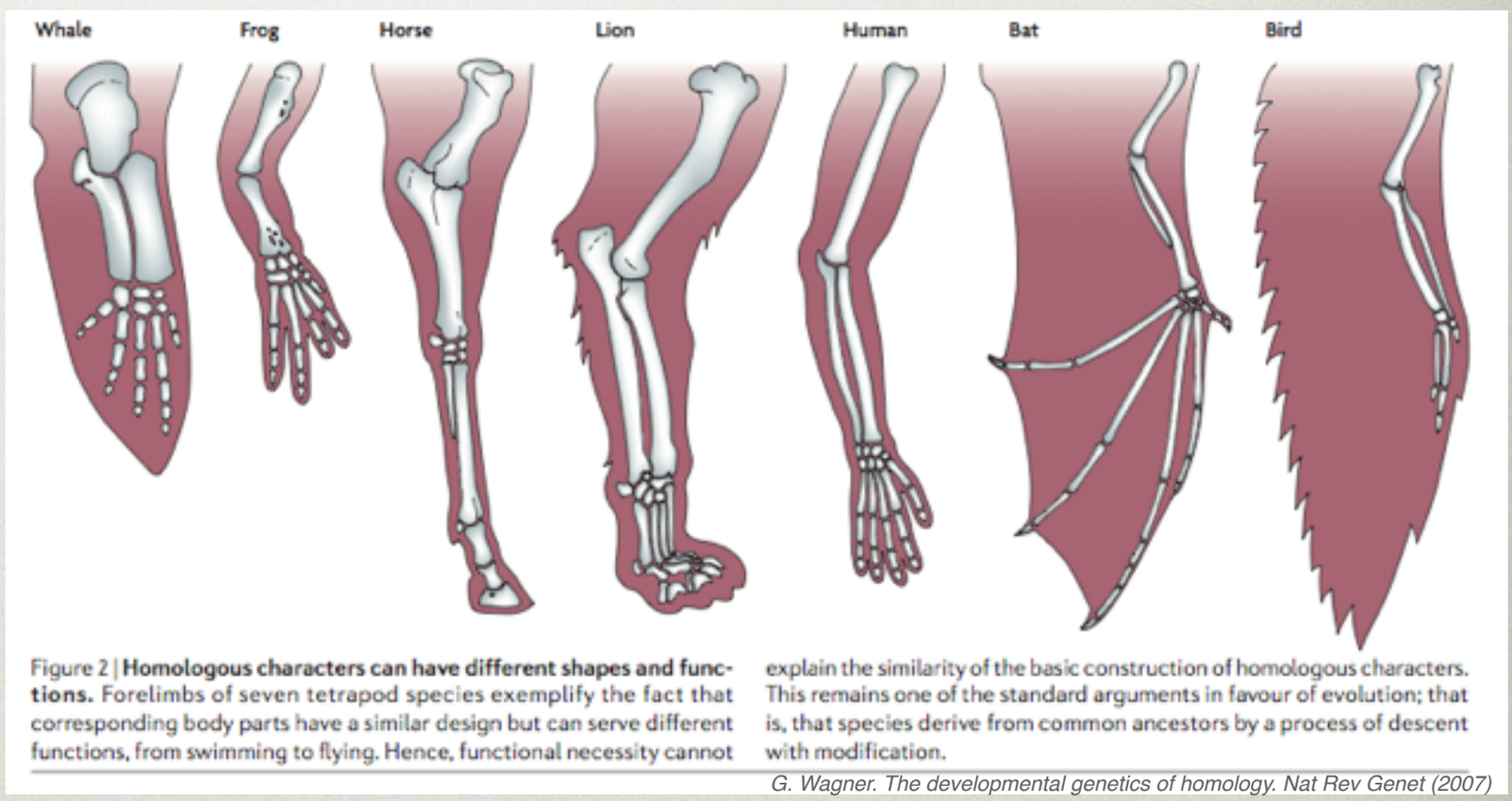


Figure 2 | **Homologous characters can have different shapes and functions.** Forelimbs of seven tetrapod species exemplify the fact that corresponding body parts have a similar design but can serve different functions, from swimming to flying. Hence, functional necessity cannot explain the similarity of the basic construction of homologous characters. This remains one of the standard arguments in favour of evolution; that is, that species derive from common ancestors by a process of descent with modification.

G. Wagner. The developmental genetics of homology. Nat Rev Genet (2007)

Homology is the result of descent with modification from a common ancestor *(Charles Darwin, 1859)*

Homology and Homoplasy

Similarity among species could represent true **homology** (just by sharing the same ancestral character state) or, **homoplastic** events like *convergence, parallelism* or *reversals*



Morphological diversity

Homology is *a posteriori* tree construction definition !!

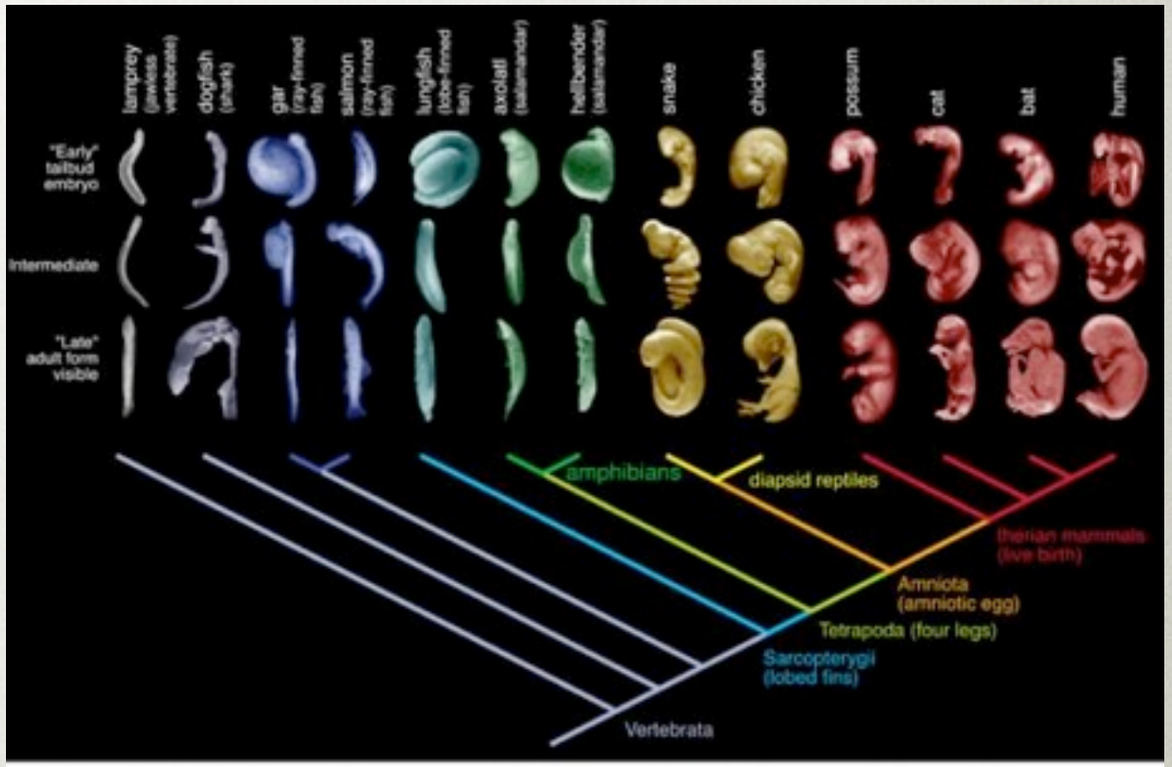
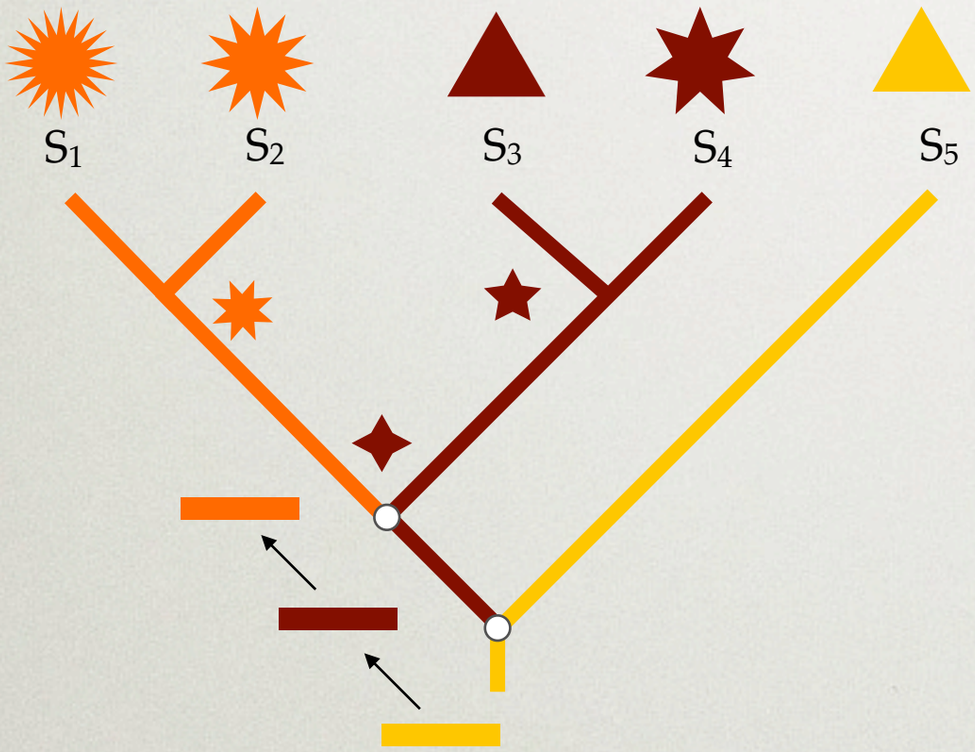


Figure 8. Developmental sequences of various vertebrates shown in phylogenetic context. Note the shared similarities of some closely related taxa, particularly the amniotes (modified from Richardson et al. 1998.)



Homologous character states

Colour conserves more phylogenetic signal than morphology

S_1	S_2	S_3	S_4	S_5
PRO	PRO	LYS	LYS	PHE
A	A	C	C	A

Homoplasy: *Convergences*

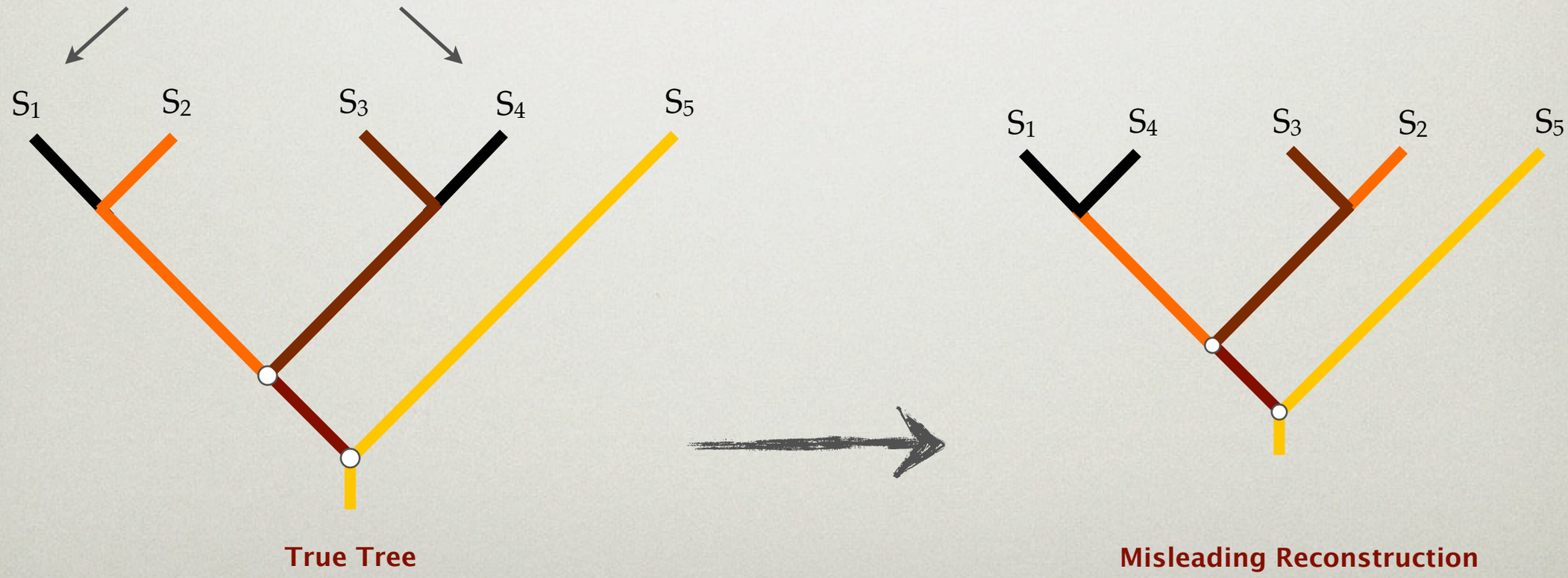
Homoplasy is the similarity that is not due to common ancestry. If such similarity is misinterpreted as homology, we probably mislead the phylogeny reconstruction.



Also, think of hedgehogs, porcupines, echidnas, cactus & sea urchin — they all have (independently-developed) spikes!

Convergence

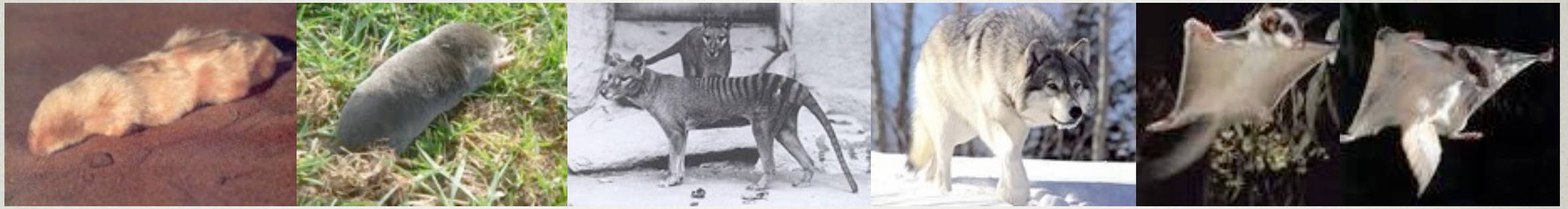
Independent evolution of the same feature from **different** ancestral condition



True Tree **Misleading Reconstruction**

Homoplasy: *Parallelism*

Homoplasy is the similarity that is not due to common ancestry. If such similarity is misinterpreted as homology, we probably mislead the phylogeny reconstruction.



Marsupial and placental faunas: moles (Notoryctes sps), tasmanyan wolf, Sugard-gliders (Petaurus breviceps), Flying Squirrel (Sciuridae family)

Thylacosmilus

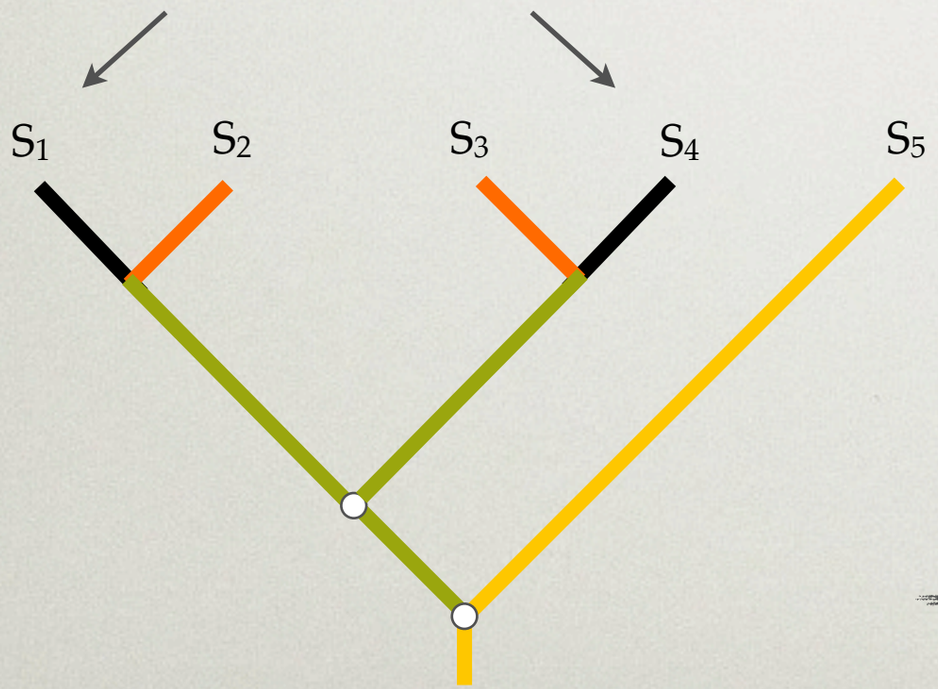


Smilodon



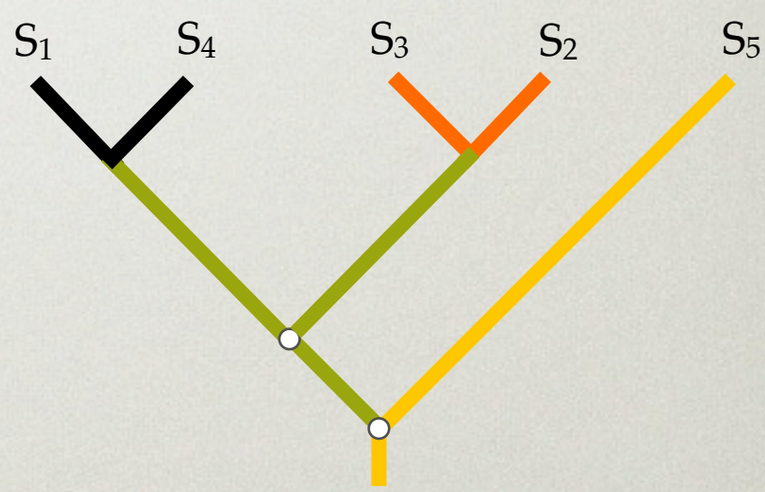
Parallelism

Independent evolution of the same feature from **the same** ancestral condition



True Tree





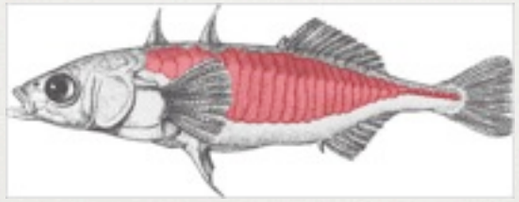
Misleading Reconstruction

Homoplasy: *Reversals*

Homoplasy is the similarity that is not due to common ancestry. If such similarity is misinterpreted as homology, we probably mislead the phylogeny reconstruction.

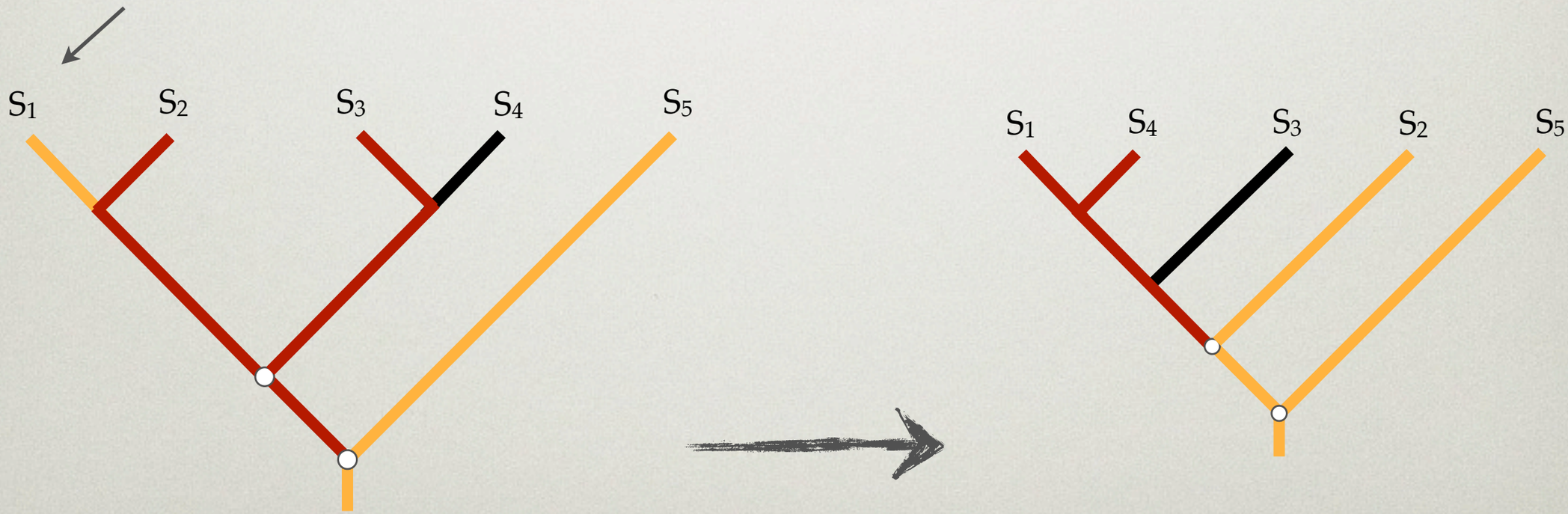
A study in sticklebacks *(Gasterosteus aculeatus*) suggests that the fish may have "evolved in reverse" in Lake Washington, developing more armoured plating for protection after a cleanup of the lake made the fish more visible to predators. The founders of the lake population were the marine sticklebacks with a similar phenotype.

Current Biology (2008) vol. 18 (10) pp. 769-774





Reversals
Reversion to ancestral condition



True Tree

Misleading Reconstruction

Similarity

* For molecular sequence data, **homology** means that two sequences or even two characters within sequences **descend from a common ancestor**.

* This term is frequently **mis-used as a synonym of similarity**.

* as in **two sequences are 70% homologous**.

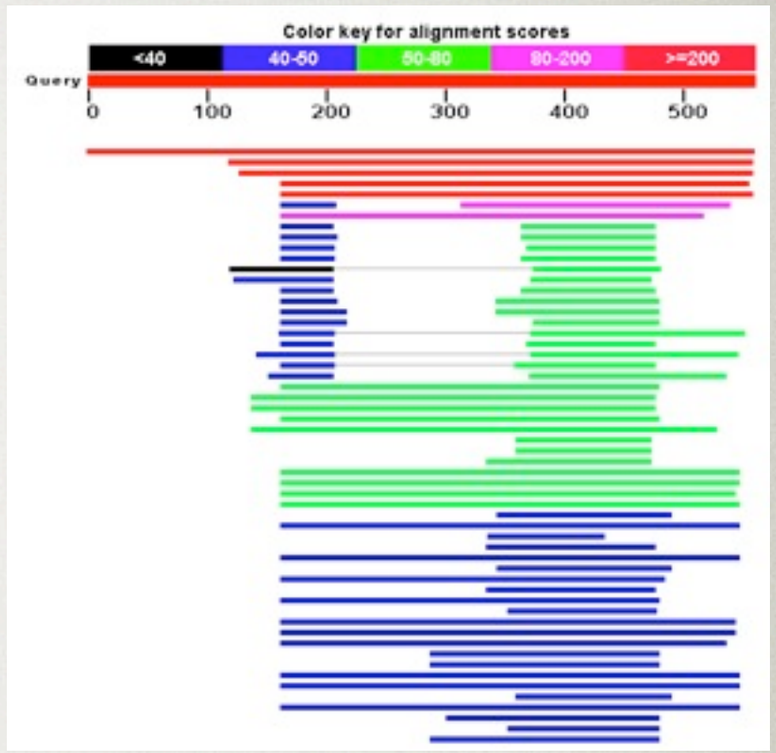
* **This is totally incorrect!**

* Sure sequences show a certain amount of similarity, and it will be high if they "recently" descend form a common ancestor

* From this similarity value, and using the proper phylogenetic methods we can infer if sequences are homologous or not

* **Homology is like pregnancy! You are either pregnant or not!**

* **Two sequences are either homologous or they are not.**



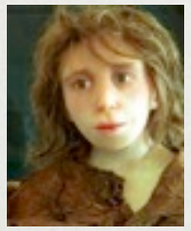


Sequence Homology

In molecular studies it is important to distinguish among kinds of homology

- **Orthologous**: Homologous that diverged from each other after **speciation events** (e.g., human β and chimp β-globin)
- **Paralogous**: Homologous that diverged from each other after **gene duplication events** (e.g., β and γ-globin)
- **Xenologous**: Homologous that diverged from each other after **lateral gene transfer events** (e.g., antibiotic resistance genes in bacteria).







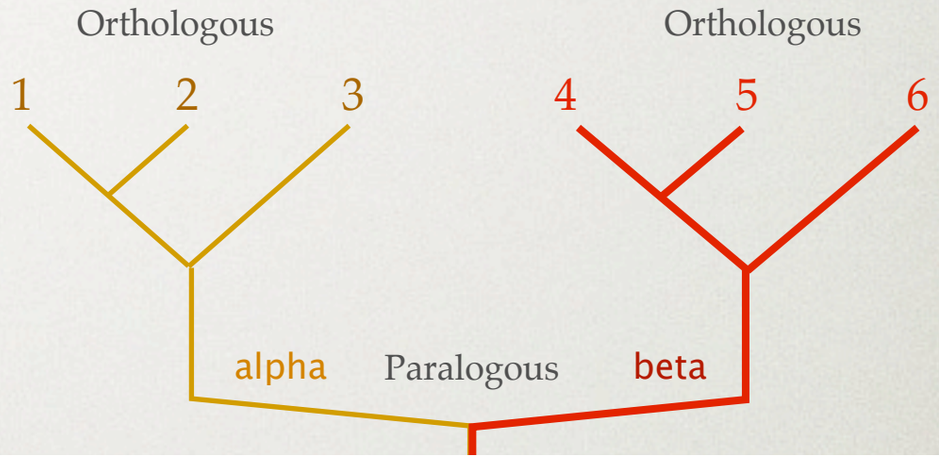
1 4 2 5 3 6

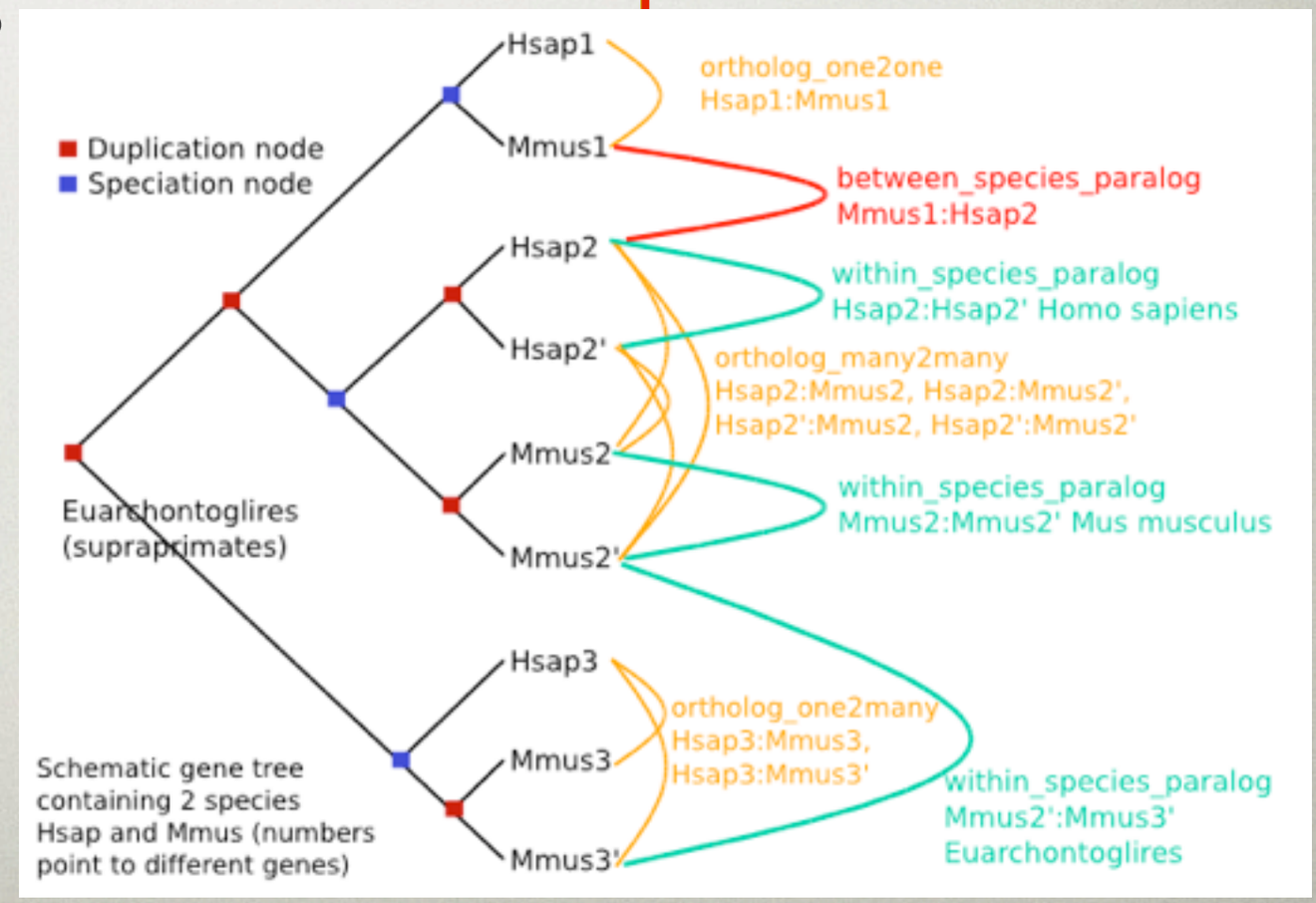
Speciation Event 2

Speciation Event 1

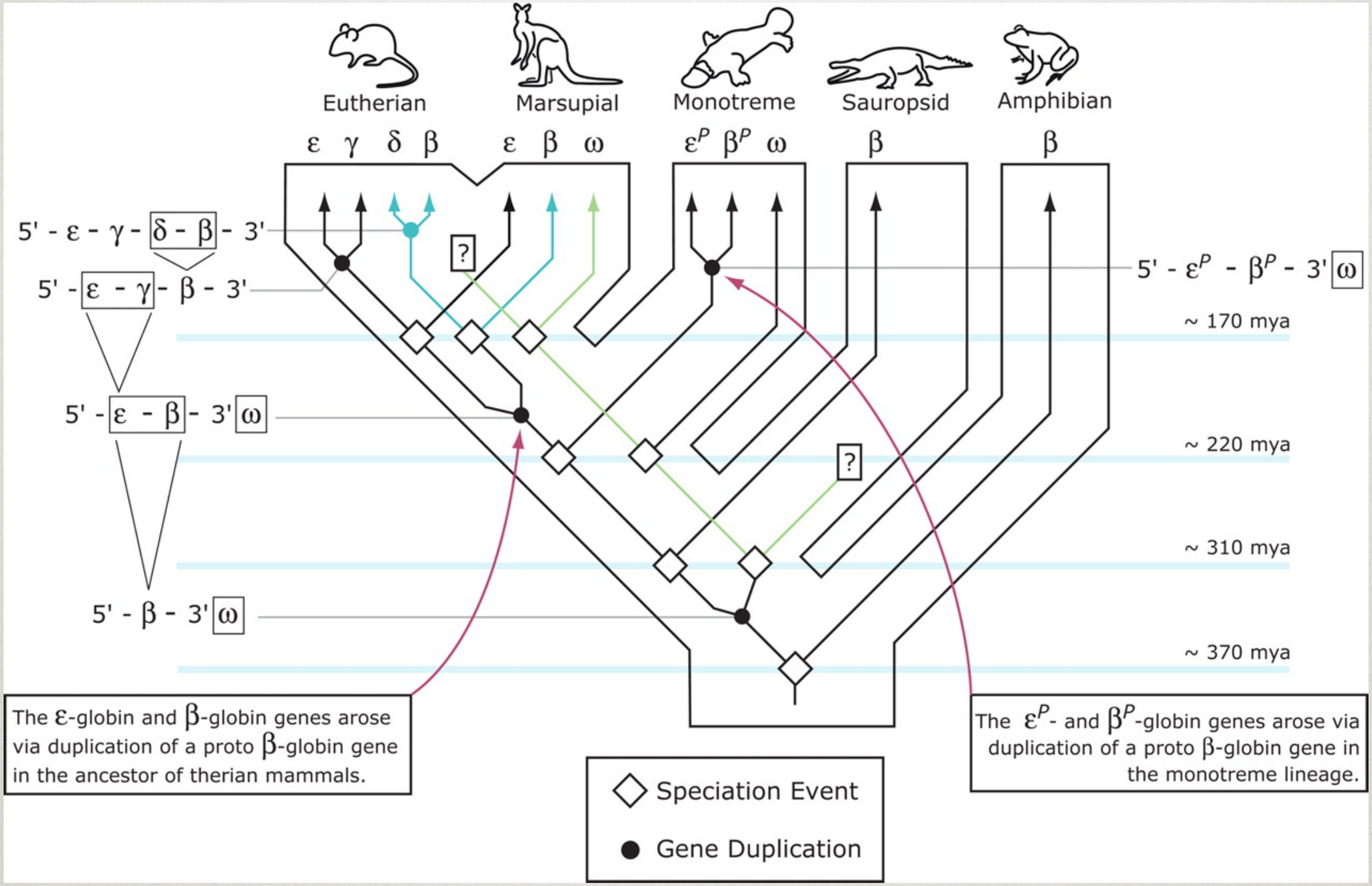
alpha beta

Duplication Event





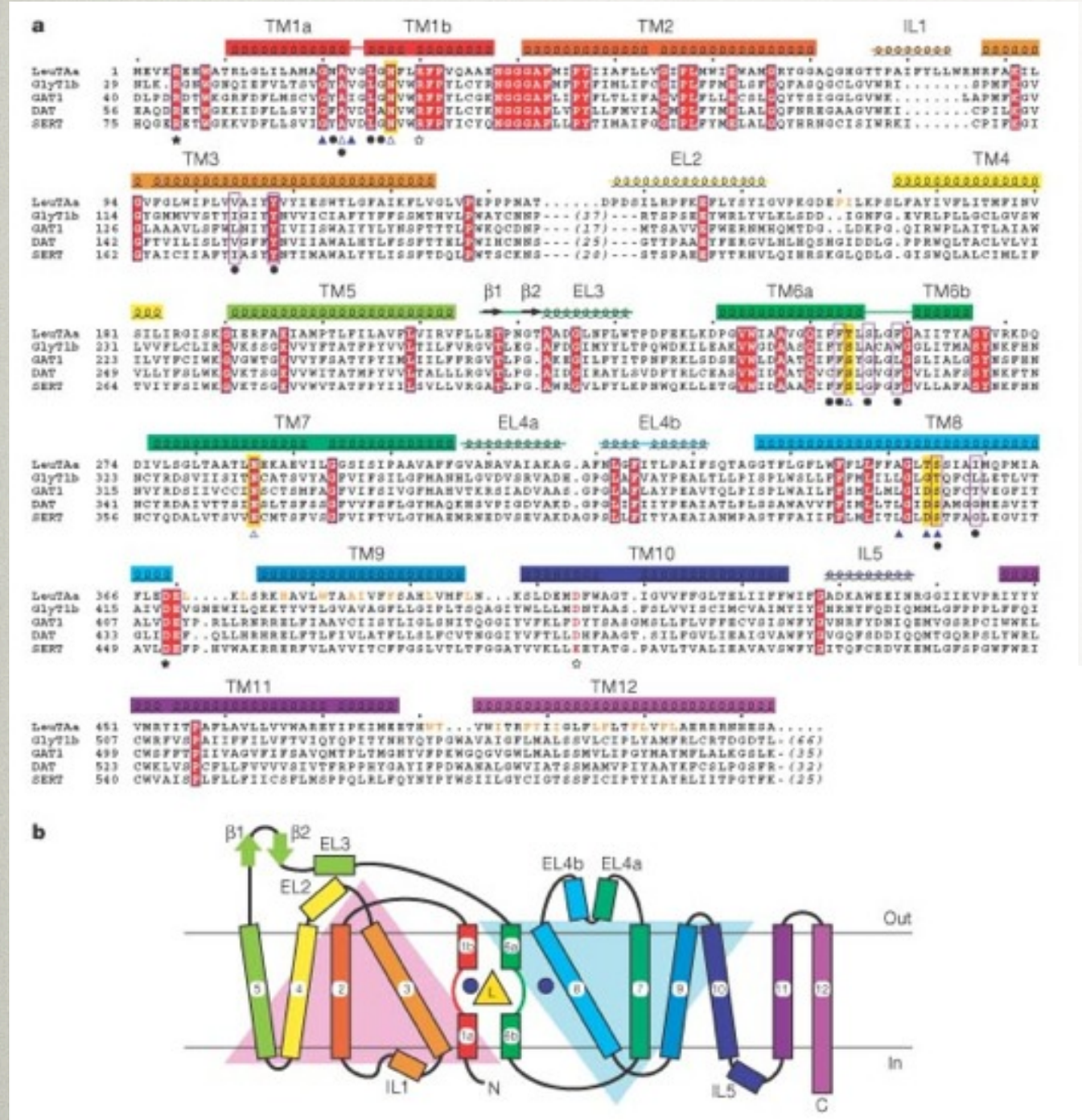
An evolutionary hypothesis regarding the evolution of the β-globin gene family.

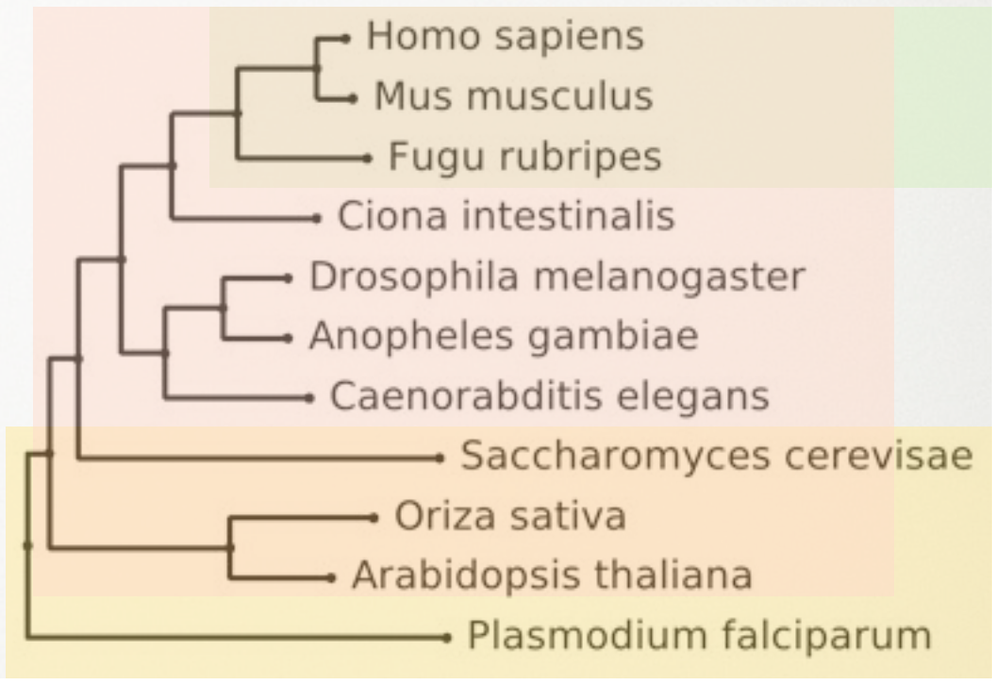


Opazo J C et al. PNAS 2008;105:1590-1595

Positional Homology

Common ancestry of specific amino acid or nucleotide positions in different genes.





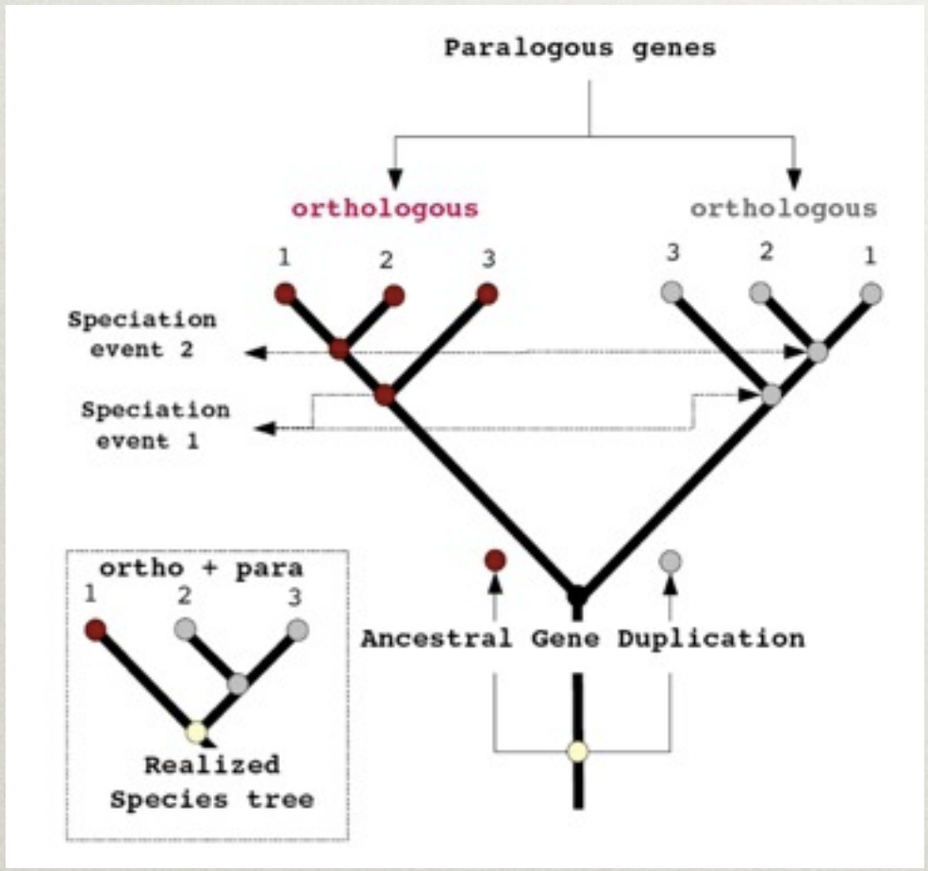

```
                                                  11 50462
Homo.sapie  VGLLGRTGSGKSTLLSAFLRLLNTEG-EIQI:
Mus.muscul  VGLLGRTGSGKSTLLSAFLRMLNIKG-DIEI:
Fugu.rubri  MGLLGRTGSGKSTLLSALLRLASTDG-EISI:
Ciona.inte  VGIVGRTGAGKSSLISTLFRLNEYSKGSVMI:
Droso.mela  VGIVGRTGAGKSSLIGALFRLAHIEG-EIFI:
Anoph.gamb  VGIVGRTGAGKSSLIGALFRLAQVEG-EIRL:
Caeno.eleg  VGIVGRTGAGKSSLTLALFRIIEADGGSIEI:
Sacch.cere  IGIVGRTGAGKSTIITALFRFLEPETGHIKI:
Arabi.thal  IGIVGRTGSGKTTLISALFRLVEPVGGKIVV:
Oryza.sati  IGVVGRTGSGKSTLVQALFRLVEPVEGHIIV:
Plasm.falc  IGIVGKSGAGKSTMILSILGLIGTTRGRITI:
```

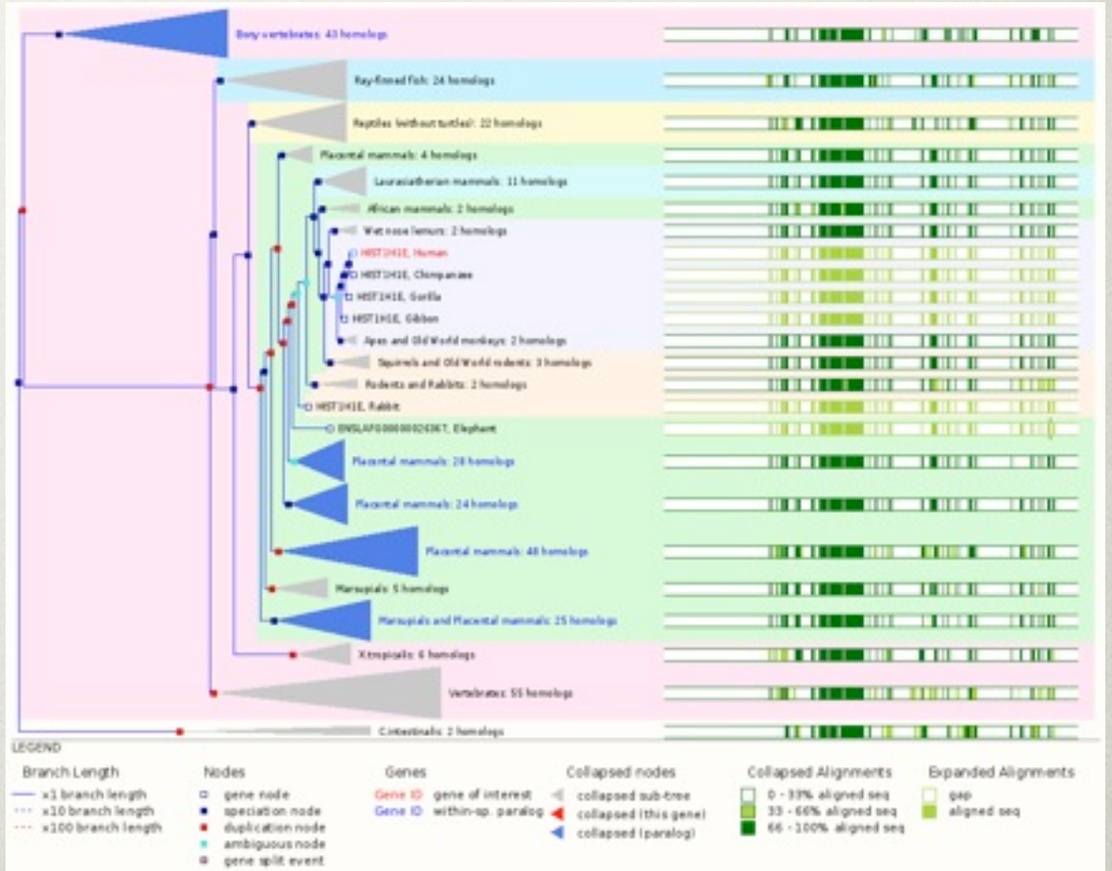
Molecular Evolution

Gene Trees & SpeciesTrees

* Sure, all phylogenetic reconstruction of sequences are genes trees.
* The naive expectation of molecular systematics is that phylogenies for genes match those of the organisms or species (species trees). **There are many reasons why this needs not be so!!**

1. If there were duplications, (gene family) only the phylogenetic reconstruction of orthologous sequences could guarantize **the expected** or true species tree.

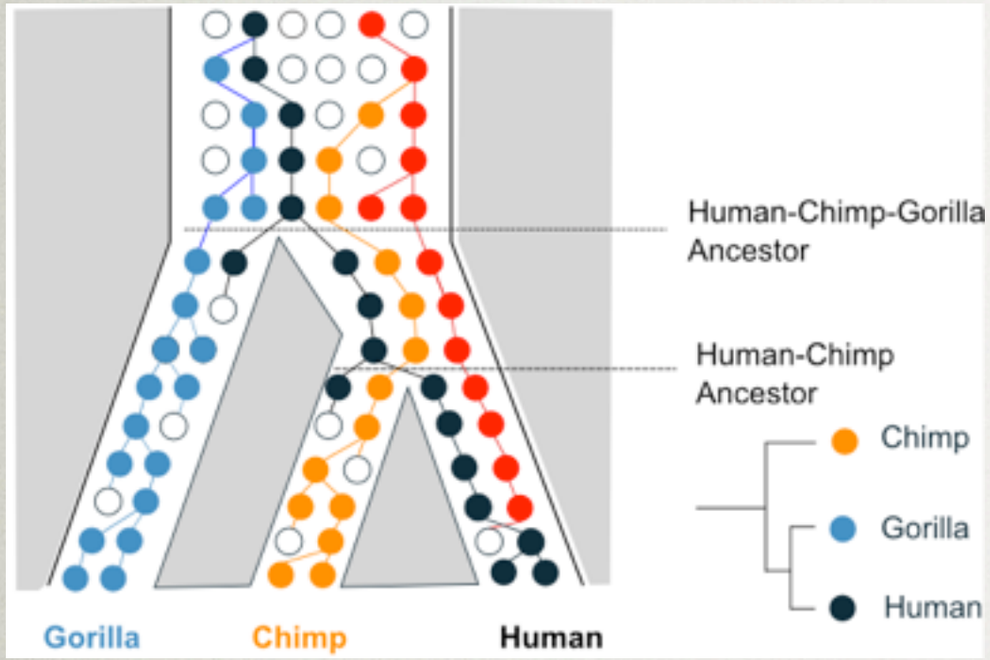


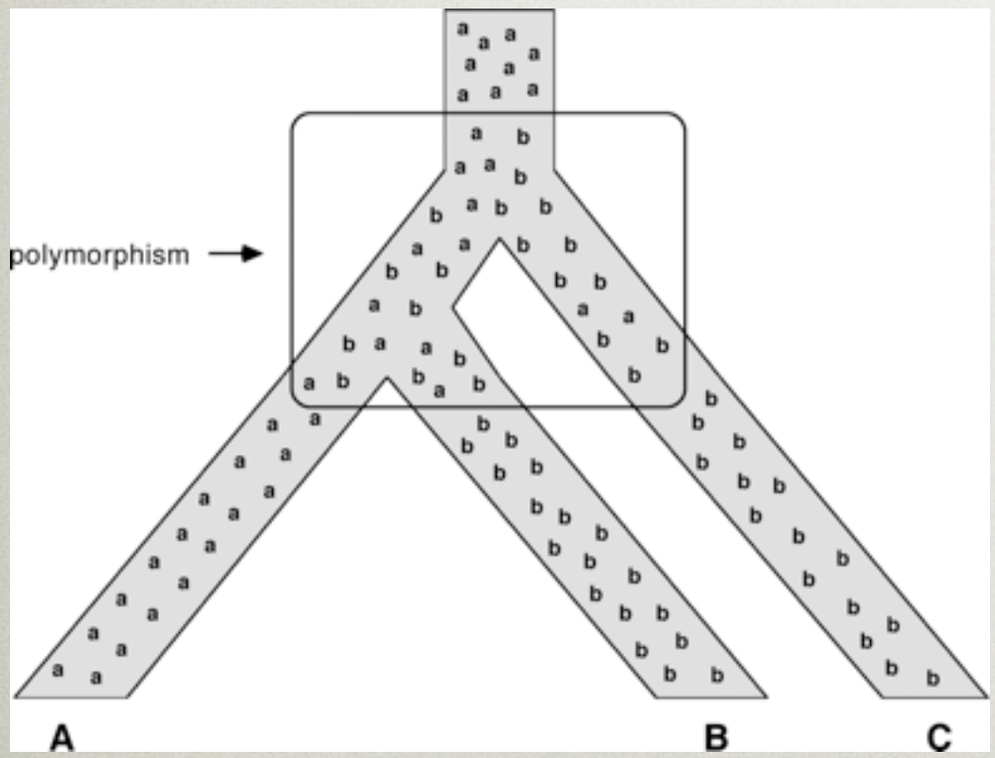


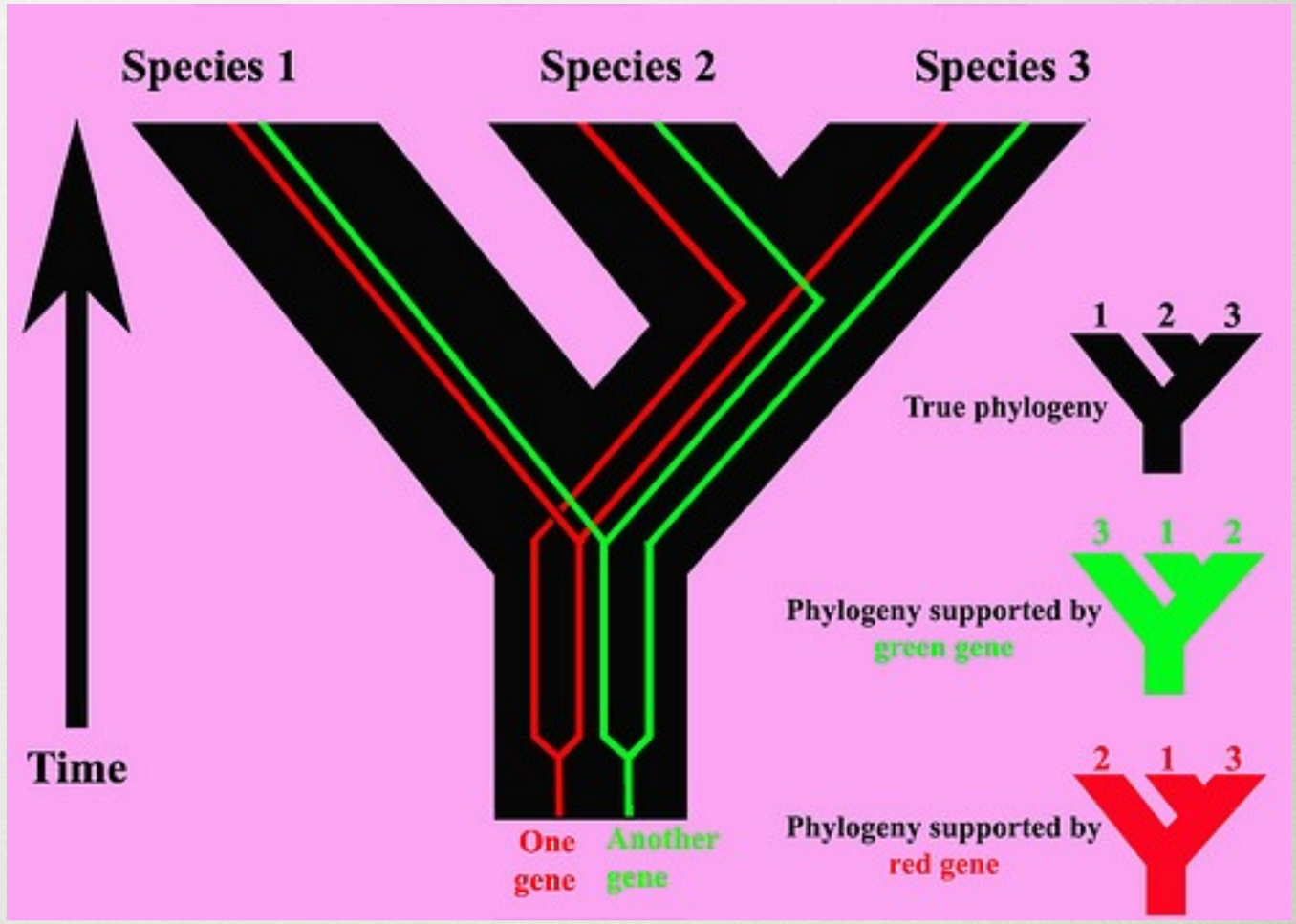
HIST1H1E_HUMAN

- The expected tree is the three that can be constructed using a infinetily long sequences

2. Lineage Sorting. The process by which, following separation of two species, the ancestry of every gene converges to the overall phylogeny of the species. This process takes about 2N generations, and **incomplete lineage sorting** **implies discordance between genealogies.**







- **To obtain a reliable tree of intraspecific populations or closely related species, a large number of unlinked genes need to be used.**

Molecular Clock

- The molecular clock hypothesis postulates that for any given macromolecule (a protein or DNA sequence), the rate of evolution *-measured as the mean number of amino acids or nucleotide sequence change per site per year-* is approximately constant over time in all the evolutionary lineages .

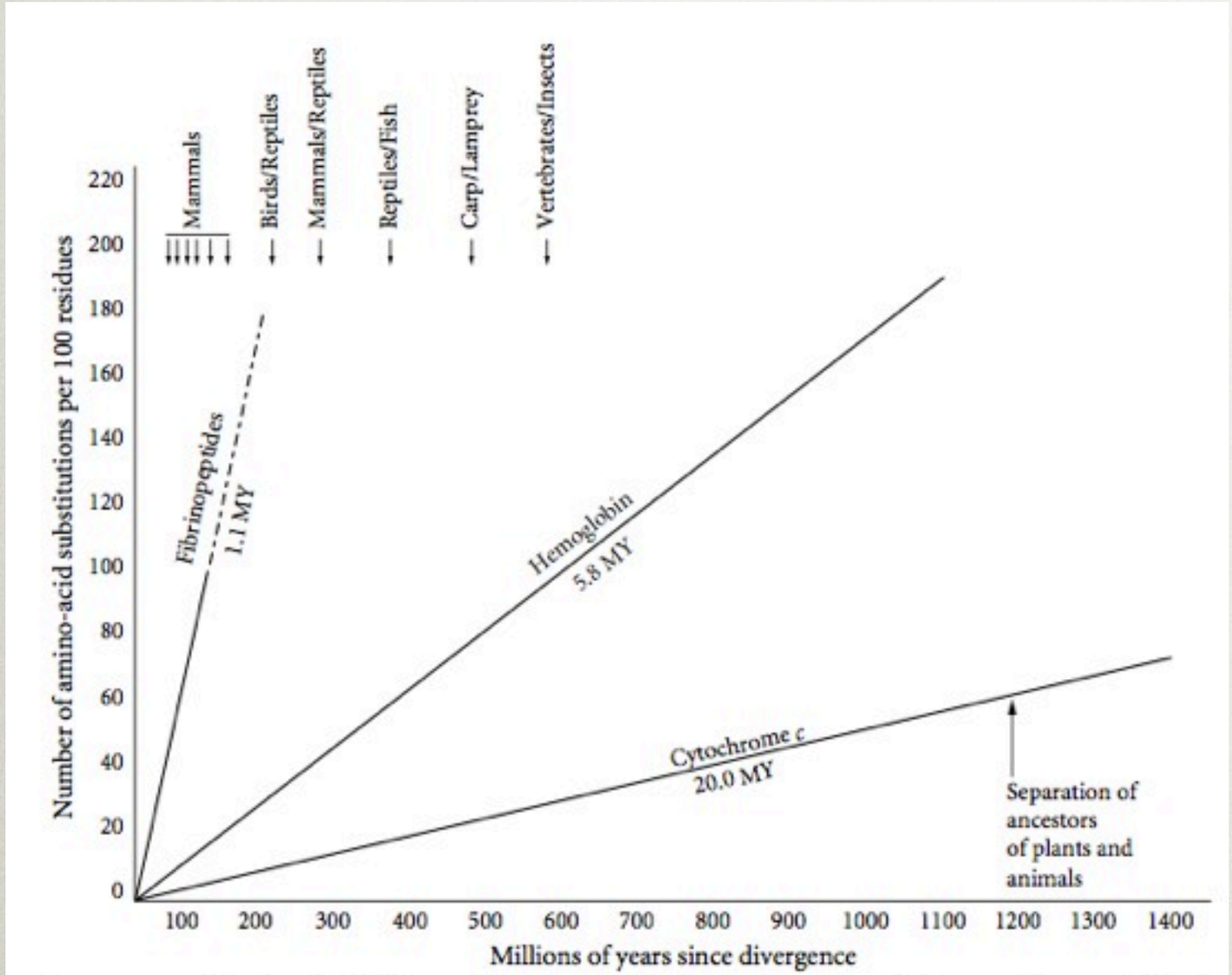


Figure 10.3 Molecular clock ticking at different speed in different proteins. Fibrinopeptides are relatively unconstrained and have a high neutral substitution rate, whereas cytochrome *c* is more constrained and has a lower neutral substitution rate (after Hartl and Clark, 1997).

- This hypothesis has estimulated much interest in the use of macromolecules in evolutionay studies for two reasons:
  - Sequences can be used as molecular markers to date evolutionary events.
  - The degree of rate change among sequences and lineages can provide insights on mechanisms of molecular evolution. For example, a large increase in the rate of evolution in a protein in a particular lineage may indicate adaptive evolution.

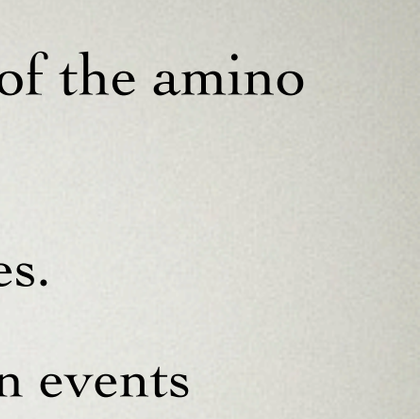
- Zuckerkandl E. and L. Pauling. Molecules as documents of evolutionary history. J Theor Biol, 8:357–366, 1965

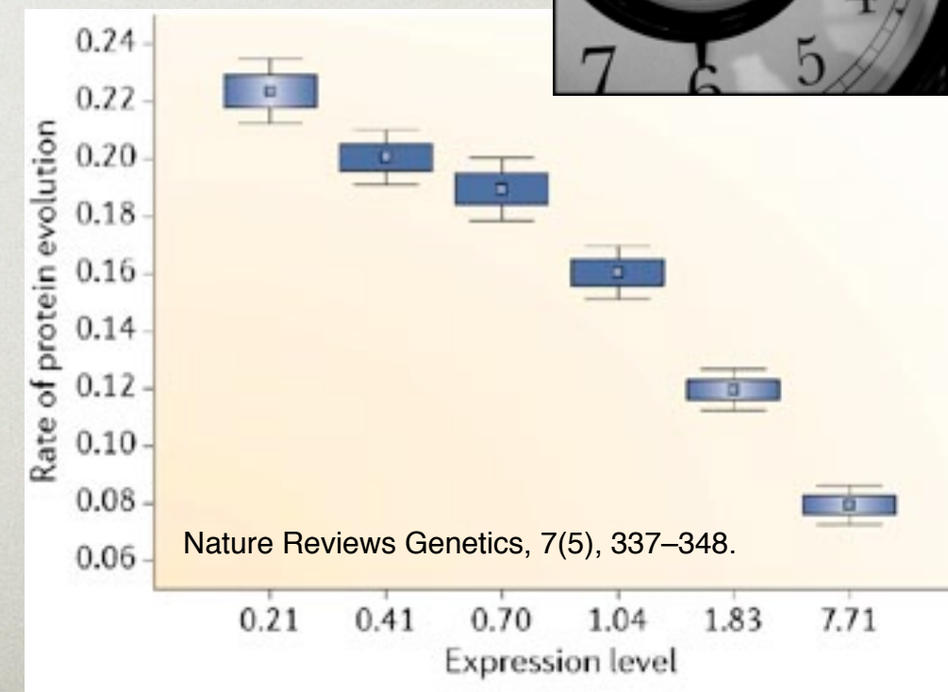
- First, the clock was envisaged as a stochastic clock due to the random nature of the amino acid or nucleotide-substitution process.
  - It 'ticks' at random intervals rather than regularly as a wristwatch normally does.
  - Under the Markov models of nucleotide or amino acid substitution, substitution events ('ticks') arrive at time intervals that are exponentially distributed.

- Second, it was noted from the beginning that different **proteins or regions of a protein evolve at drastically different rates**, so that the clock hypothesis allows for such rate differences among proteins; the different proteins are said to have their own clocks, which tick at different rates.

- Third, rate constancy may **not hold globally for all species** and usually applies to a group of species. For example, we may say that the clock holds for a particular gene within primates.

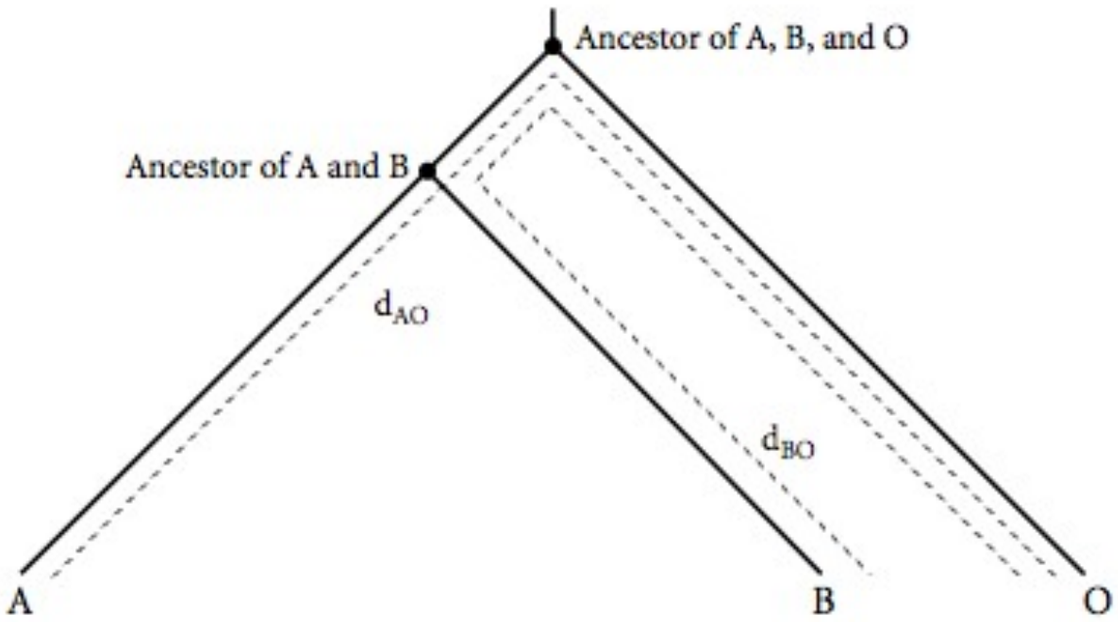
- Fourth, Gene-expression level (measured as mRNA abundance) correlates strongly and negatively with the rate of protein evolution in yeast (R^2 = 0.29 for individual genes). Evolutionary rate (non-synonymous divergence).





Relative Rate Test

- The relative rate test evaluates the molecular clock hypothesis comparing whether dAO – dBO is significantly different from zero.
- The sign of the difference indicates which taxon is evolving faster or slower.
- The relative rate test assumes that the phylogenetic relationships among the taxa are known.



Under a perfect molecular clock,
$d_{AO} = d_{BO}$,
then $d_{AO} - d_{AO} = 0$

Figure 10.4 The relative rate test. Under a molecular clock, the distance from A to O should be the same as the distance from B to O.

RRTree in Phylemon 2.0

treatement of 2 lineages

equally weighted sequences:
> Outgroup:
bull 1
> lineage 1:
hsap 0.5
ptro 0.5
> lineage 2:
mmus 1

4 651
hsap GGGGCCTTGC ACGGAGCCCT GCGGCTACTG CCAACCACCT GGCTTCTATG
ptro GGGGCCCTGC ACCGCCACAC ACGGCTACTC TGTTCCCGCA GAGATACTGA
mmus GGGACCCTGC GCAGAACCCT GTGGCTCTTG TGTCCCACCT GGCTTCTATA
bull GGGACCCTGC GCAGAACCCT GTGGCTCTTG TGTCCCACCT GGCTTCTATA

AACATGGCGA TGGCTGCGTG TCCTGCCCCA CGAGCACCCT GGGGAGCTGT
CTGTGGGACC TGCCTGCCTG GCTTCTATGA ACATGGCGAT GGCTGCGTGT
TACGTGGCAA TGACTGCACG TCCTGCCCCA CGGGCTTCAG CAGCGTTTGC
TACGTGGCAG ATGGCTGCGT GTCCTGCCCC ACGAGCACCC TGGGGAGCTG

.....

mean GC Outgroup: all positions=0.674 position1=0.728 position2=0.737 position3=0.558
mean GC lineage 1: all positions=0.677 position1=0.793 position2=0.694 position3=0.544
mean GC lineage 2: all positions=0.630 position1=0.654 position2=0.668 position3=0.567

mean number of synonymous sites compared: 198.8
mean Ks lineage 1: 0.387166
mean Ks lineage 2: 0.397199
dKs: -0.0100328 standard deviation (sd): 0.0439469 ratio dKs/sd: -0.228294
exact probability (P): 0.81942

mean number of non synonymous sites compared: 452.2
mean Ka lineage 1: 0.360689
mean Ka lineage 2: 0.533611
dKa: -0.172922 standard deviation (sd): 0.0418418 ratio dKa/sd: -4.13276
exact probability (P): 3.77151e-05

Neutral Theory of Molecular Evolution

* At molecular level, the most frequent changes are those involving fixation in populations of neutral selective variants.

- Allelic variants are functionaly equivalent
- Neutralism does not deny adaptive evolution
- Fixation of new allelic variants occurs at a constant rate μ.
- This rate does not depends on any other population parameter, then it's like a clock!!
- $2N\mu * 1/2N = \mu$



Motoo Kimura & Tomoko Otha, 1968

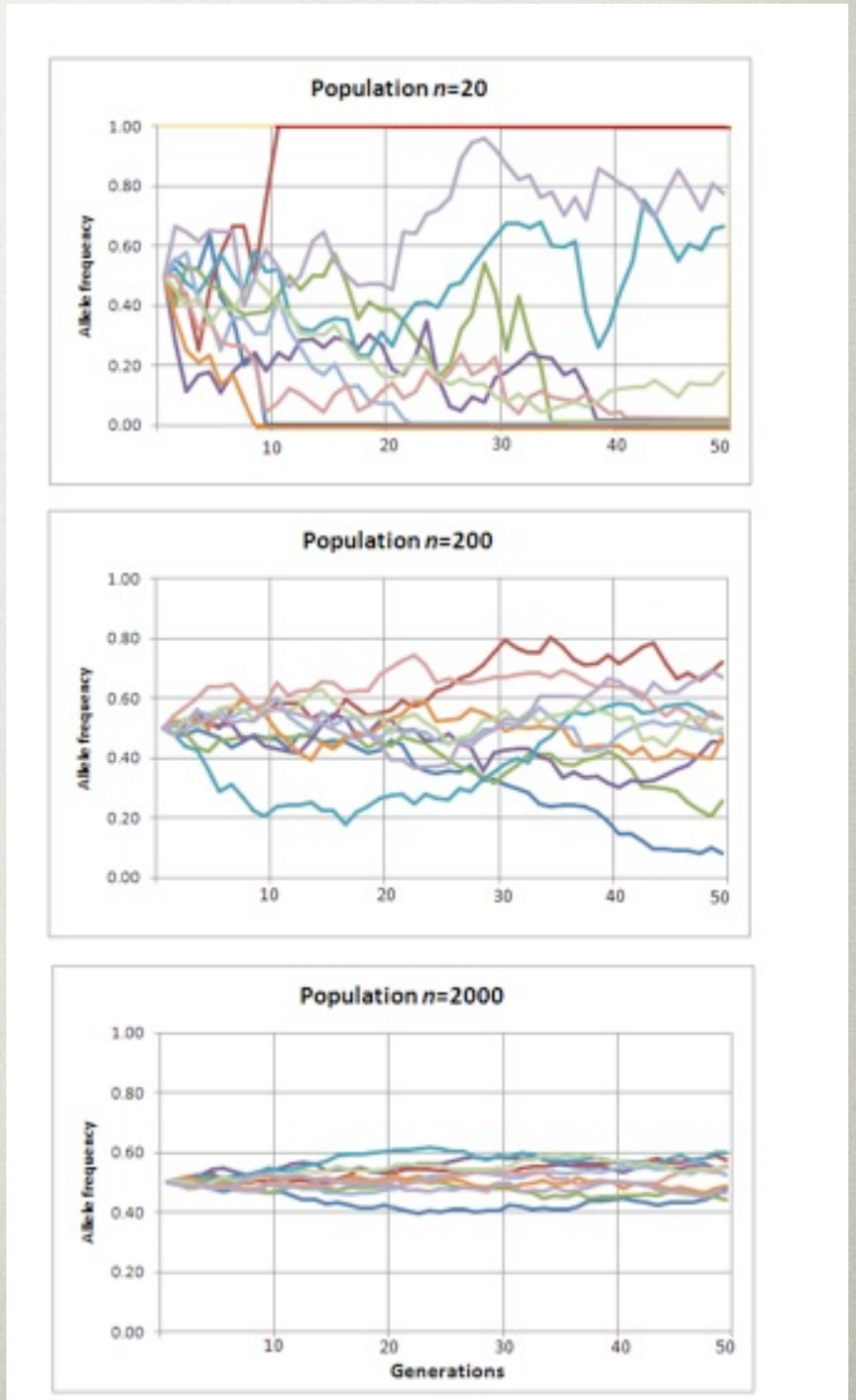
Evolutionary Rate at the Molecular Level

by
MOTOO KIMURA

National Institute of Genetics,
Mishima, Japan

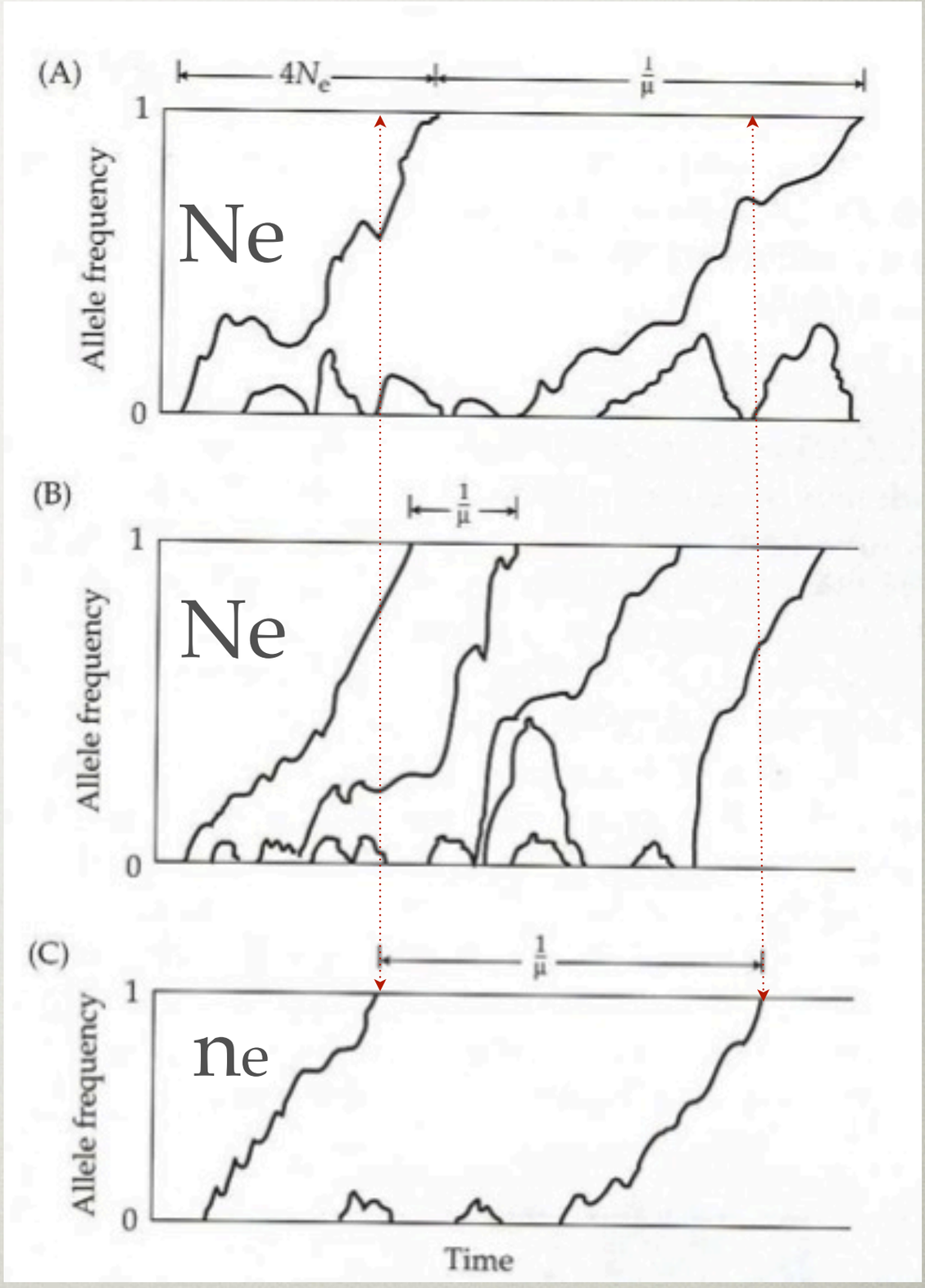
Calculating the rate of evolution in terms of nucleotide substitutions seems to give a value so high that many of the mutations involved must be neutral ones.

COMPARATIVE studies of haemoglobin molecules among different groups of animals suggest that, during the evolutionary history of mammals, amino-acid substitution has taken place roughly at the rate of one amino-acid change in 10^7 yr for a chain consisting of some 140 amino-acids. For example, by comparing the α and β chains of man with those of horse, pig, cattle and rabbit, the figure of one amino-acid change in 7×10^6 yr was obtained[1].



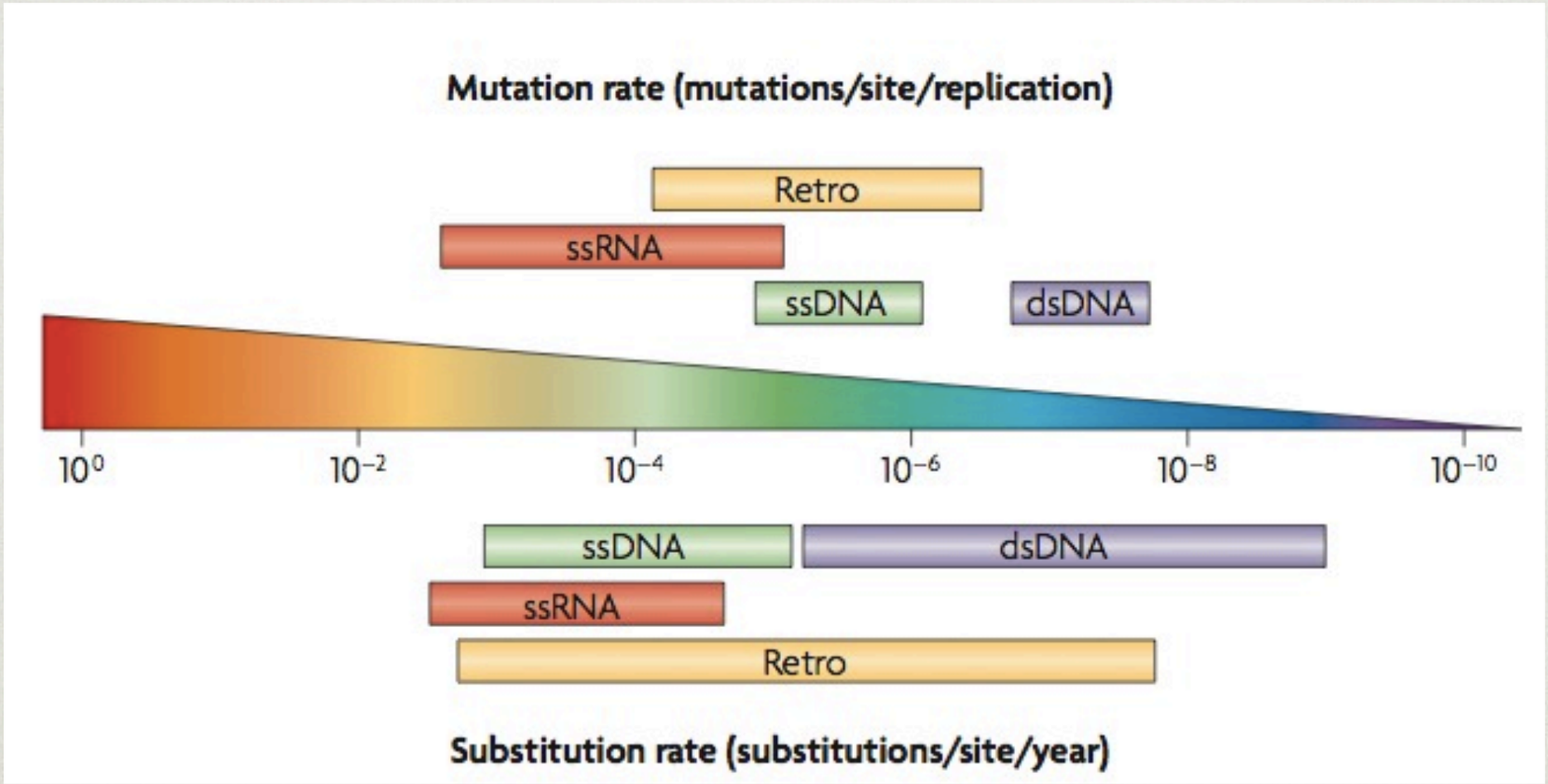
Mutation rate & Substitution rate

- Assuming an "Infinite Allele Model [1]"
- New alleles enter the population by mutation with frequency of 1/2N
- Most alleles are lost. Those that go to fixation take an average of 4Ne generations
- The trajectory of neutral alleles in a population.
  - A: Moderate size population (Ne)
  - B: The same popularion size but a high mutation rate gives a less time between fixations of alleles.
  - C: A smaller population has alleles that go to fixation more rapidly, but the time of fixation is the same 1/u



1. The IAM (*Kimura & Crow, 1964*) assumes that all alleles mutate at equal rates, and when they do, **all give rise only to new alleles** that have never been seen before. When it is used for nucleotide sequences, this model in effect treats all sequences as differing, **without inquiring at which sites they differ or by how many sites.** Any statistical treatment based on this method will thus necessarily discard all information about which alleles are historically close to each other.

Mutation rate & Substitution rate



Nature Reviews Genetics **9**, 267-276 (April 2008)

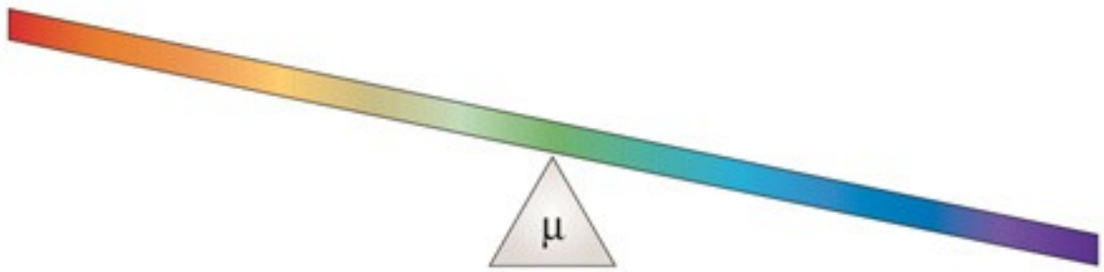
The **ranges of mutation rates**, given as mutations per site per round of replication, for viruses with different genomic architectures are summarized in the upper part of the figure. The **ranges of average substitution rates**, given as substitutions per site per year, are shown in the lower part of the figure.

The ranges are bound by the extremes of average substitution rates: for the retrotranscribing viruses (Retro), human immunodeficiency virus type-1

Mutation rate & Substitution rate

Effects on substitution rate

Generation time	Shorter	Longer
Transmission	Direct	Vector borne
	Horizontal	Vertical
Selection	Positive	Purifying



Effects on mutation rate

Genomic architecture	Single stranded	Double stranded
	RNA	DNA
	Smaller	Larger
Replication speed	Faster	Slower
Viral enzymes	–	DNA repair
Host enzymes	Deamination, oxidation	–
Environmental effects	Deamination, oxidation, UV	–

Nature Reviews Genetics **9**, 267-276 (April 2008) Nature Reviews | Genetics

In the upper portion of the figure, **the factors that increase viral substitution rates are shown in red and those that decrease it are shown in blue. The baseline substitution rate is determined by the neutral mutation rate, μ.** Because increasing mutation rate can increase the substitution rate, the factors increasing mutation rate are also shown in red and those decreasing mutation rate in blue. It is important to note that many of these factors are not independent; for instance, smaller genomes tend to replicate faster, or vertically transmitted viruses sometimes become latent in the host genome and so experience fewer generations compared with obligately lytic, horizontally-transmitted viruses. UV, ultraviolet irradiation.

References

History of Phylogenetics

- Camin J. H. and R. R. Sokal. A method for deducing branching sequences in phylogeny. Evolution, 19:311–326, 1965
- Cavalli-Sforza L. L. and A. W. F. Edwards. Analysis of human evolution. In Genetics Today. Proceeding of the XI International Congress of Genetics, The Hague, The Netherlands., volume 3, pages 923–933. Pergamon Press, Oxford, 1965.
- Cavalli-Sforza L. L. and A. W. F. Edwards. Phylogenetic Analysis: Models and estimation procedures. American Journal of Human Genetics, 19:223– 257, 1967.
- Darwin, Charles (1859), On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. London John Murray.
- Eck, R. V. and M. O. Dayhoff. Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Spring, Maryland, 1966
- Farris J. S.. Methods for computing Wagner trees. Systematics Zoology, 19:83–92, 1970.
- Felsenstein J. Inferring phylogenies. Sinauer associates, Inc., Sunderland, MA, 2004
- Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology 20: 406-416
- Fitch, W. M. and E. Margoliash. (1967). Construction of phylogenetic trees. Science 155: 279-284.
- Haeckel Ernst. Generelle Morphologie der Organismen, 1866
- Hennig W. Cladistic analysis or cladistic classification?“ A reply to Ernst Mayr. in: Syst. Zool. 24.1975, 244-256
- Hennig W. Grundzüge einer Theorie der phylogenetischen Systematik. Deutscher Zentralverlag, Berlin 1950.
- Hennig, W. 1965. Phylogenetic Systematics. Ann. Rev. Entomol. 10,97-116
- Jukes T. H. and C. R. Cantor. Evolution of protein molecules. In M. N. Munro, editor, Mammalian protein metabolism, volume III, pages 21–132. Academic Press, N. Y., 1969.
- Kluge and J. S. Farris. A successive approximations approach to character weighting. Systematics Zoology, 18:374–385, 1969.
- Kluge and J. S. Farris. Quantitative phyletics and the evolution of anurans. Systematics Zoology, 18: 1–36, 1969.
- Mayr, Ernst (1942). Systematics and the origin of species, from the viewpoint of a zoologist. Cambridge: Harvard University Press.
- Mayr, Ernst (1982). The Growth of Biological Thought: Diversity, Evolution, and Inheritance. Cambridge, Mass: Belknap Press.
- Simpson George G. Tempo and mode in evolution ,1944
- Sneath P.H.E. and Sokal R.R. (1963) Principles of Numerical Taxonomy. Freeman & Co., San Francisco
- Sokal R.R. (1988) "Genetic, geographic, and linguistic distances in Europe." Proceedings of the National Academy of Sciences USA 85:1722-1726.
- Sokal R.R. and Michener C.D. (1958) "A Statistical Method for Evaluating Systematic Relationships". The University of Kansas Scientific Bulletin 38: 1409-1438.
- Edwards E.. Statistical methods for evolutionary trees. Genetics (2009) vol. 183 (1) pp. 5-12

Phylogenetic Applications I

Evolutionary genetics

- D. A. Buonagurio, S. Nakada, W. M. Fitch, and P. Palese. Epidemiology of influenza C virus in man: multiple evolutionary lineages and low rate of change. Virology, 153:12–21, 1986.
- W. M. Fitch. Evolution of clupeine Z, a probable crossover product. Nat New Biol, 229:245–247, 1971.
- W. M. Fitch. Phylogenies constrained by the crossover process as illustrated by human hemoglobins and a thirteen-cyclel. Genetics, 86:623–644, 1977.
- A. Kurosky, D. R. Barnett, T. H. Lee, et al. Covalent structure of hu- man haptoglobin: a serine protease homolog. PNAS, 77:3388–3392, 1980.
- J. Raymond, J. L. Siefert, C. R. Staples, and R. E. Blankenship. The natural history of nitrogen fixation. Mol Biol Evol, 21:541–554, 2004.
- S. Rudikoff, W. M. Fitch, and M. Heller. Exon-specific gene correction (conversion) during short evolutionary periods: homogenization in a two- gene family encoding the beta-chain constant region of the T-lymphocyte antigen receptor. Mol Biol Evol, 9:14–26, 1992.
- C. Scholtissek, S. Ludwig, and W. M. Fitch. Analysis of influenza A virus nucleoproteins for the assessment of molecular genetic mechanisms leading to new phylogenetic virus lineages. Arch Virol, 131:237–250, 1993.
- P. H. Harvey, A. J. Leigh Brown, John Maynard Smith, and S. Nee. New Uses for New Phylogenies. Oxford Univ Press, Oxford. England, 1996.

Population genetics

- J. Felsenstein. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet Res, 59:139–147, 1992
- B. Golding and J. Felsenstein. A maximum likelihood approach to the detection of selection from a phylogeny. J Mol Evol, 31:511–523, 1990
- J. Hey. The structure of genealogies and the distribution of fixed differences between DNA sequence samples from natural populations. Genetics, 128:831–840, 1991.
- K. Sjolander. Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics, 20:170–179, 2004.
- B. C. Emerson, E. Paradis, and C. Thebaud. Revealing the demographic histories of species using DNA sequences. TREE, 16:707–716, 2001.

Rates & Dates

- W. M. Fitch and F. J. Ayala. The superoxide dismutase molecular clock revisited. Proc Natl Acad Sci U S A, 91:6802–6807, 1994.
- S. B. Hedges. The origin and evolution of model organisms. Nat Rev Genet, 3:838–849, 2002.
- S. B. Hedges, H. Chen, S. Kumar, D. Y. Wang, A. S. Thompson, and H. Watanabe. A genomic timescale for the origin of eukaryotes. BMC Evol Biol, 1:4, 2001.
- S. Kumar and S. B. Hedges. A molecular timescale for vertebrate evolution. Nature, 392:917–920, 1998.
- S. H. Yeh, H. Y. Wang, C. Y. Tsai, C. L. Kao, J. Y. Yang, H. W. Liu, I. J. Su, S. F. Tsai, D. S. Chen, and P. J. Chen. Characterization of severe acute respiratory syndrome coronavirus genomes in Taiwan: molecular epidemiology and genome evolution. Proc Natl Acad Sci U S A, 101:2542– 2547, 2004.

Testing Hypotheses

- T. Gubitz, R. S. Thorpe, and A. Malhotra. Phylogeography and natural selection in the Tenerife gecko Tarentola delalandii: testing historical and adaptive hypotheses. Mol Ecol, 9:1213–1221, 2000.
- M. S. Hafner and R. D. Page. Molecular phylogenies and host-parasite cospeciation: gophers and lice as a model system. Philos Trans R Soc Lond B Biol Sci, 349:77–83, 1995.
- R. D. Page, R. H. Cruickshank, M. Dickens, R. W. Furness, M. Kennedy, R. L. Palma, and V. S. Smith. Phylogeny of Philoceanus complex seabird lice (Phthiraptera: Ischnocera) inferred from mitochondrial DNA se- quences. Mol Phylogenet Evol, 30:633–652, 2004.
- R. D. M. Page. Tangled trees. The University of Chicago Press, Chicago, London, 2001.

Phylogenetic Applications II

- Y. Surget-Groba, B. Heulin, C. P. Guillaume, R. S. Thorpe, L. Kupriyanova, N. Vogrin, R. Maslak, S. Mazzotti, M. Venczel, I. Ghira, G. Odierna, O. Leontyeva, J. C. Monney, and N. Smith. Intraspecific phylogeography of Lacerta vivipara and the evolution of viviparity. Mol Phylogenet Evol, 18:449–459, 2001.

Conservation and Legal /Forensic Applications

- D. M. Hillis and J. P. Huelsenbeck. Support for dental HIV transmission. Nature, 369:24–25, 1994.
- C. Moritz. Strategies to protect biological diversity and the evolutionary processes that sustain it. Syst Biol, 51:238–254, 2002.

Applications using a known tree

- D. R. Brooks and D. A. McLennan. Phylogeny, ecology and behaviour. A research program in comparative biology. The University of Chicago Press, Chicago. USA, 1991.
- P. H. Harvey and M. D. Pagel. The comparative Method in Evolutionary Biology. Oxford Series in Ecology and Evolution, Oxford. England, 1991.
- E. P. Martins. Phylogenies and the comparative method in animal behavior. Oxford University Press, Oxford. England, 1996.
- A. Meyer. Hox gene variation and evolution. Nature, 391:225, 227–8, 1998.
- R. D. Page, R. H. Cruickshank, M. Dickens, R. W. Furness, M. Kennedy, R. L. Palma, and V. S. Smith. Phylogeny of Philoceanus complex seabird lice (Phthiraptera: Ischnocera) inferred from mitochondrial DNA se- quences. Mol Phylogenet Evol, 30:633–652, 2004.
- Blackledge and Gillespie. Convergent evolution of behavior in an adaptive radiation of Hawaiian web-building spiders. Proc Natl Acad Sci USA (2004) vol. 101 (46) pp. 16228-33
- Sebé-Pedrós et al. Ancient origin of the integrin-mediated adhesion and signaling machinery. Proc Natl Acad Sci USA (2010) pp.

Phylogenomics

- F. Delsuc, H. Brinkmann, and H. Philippe. Phylogenomics and the re- construction of the tree of life. Nature Review in Genetics, 6:361–375, 2005.
- H. Dopazo, J. Santoyo, and J. Dopazo. Phylogenomics and the number of characters required for obtaining an accurate phylogeny. Bioinformatics, 20 (Suppl. 1):i116–i121, 2004.
- J. E. Blair, K. Ikeo, T. Gojobori, and S. B. Hedges. The evolutionary position of nematodes. BMC Evol Biol, 2:7, 2002.
- H. Dopazo and J. Dopazo. Genome scale evidence of the nematode-arthropod clade. Genome Biology, 6:R41, 2005.
- Y. I. Wolf, I. B. Rogozin, and E. V. Koonin. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res, 14:29–36, 2004.
- J. A. Eisen. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res, 8:163–167, 1998.
- J. A. Eisen and M. Wu. Phylogenetic analysis and gene functional predictions: phylogenomics in action. Theor Popul Biol, 61:481–487, 2002.
- Arbiza et al. Selective pressures at a codon-level predict deleterious mutations in human disease genes. J Mol Biol (2006) vol. 358 (5) pp. 1390-404
- Capriotti et al. Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. Hum Mutat (2008) vol. 29 (1) pp. 198-204
- Conde et al. PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. Nucleic Acids Res (2006) vol. 34 (Web Server issue) pp. W621-5
- STAR Consortium et al. SNP and haplotype mapping for genetic analysis in the rat. Nat Genet (2008) vol. 40 (5) pp. 560-6
- Huerta-Cepas et al. The human phylome. Genome Biol (2007) vol. 8 (6) pp. R109

Homology

- Owen, R. 1847. On the archetype and homologies of the vertebrate skeleton. London.
- Darwin, Charles (1859), On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. London John Murray.
- Homology. 1994. The hierarchical basis of comparative biology, Academic Press, N. Y.
- Abouheif. 1997. Developmental genetics and homology: a hierarchical approach. Trends in Ecology & Evolution. vol. 12 (10) pp. 405-408
- Shubin N, Tabin C, Carroll S. 1997. "Fossils, genes and the evolution of animal limbs". Nature 388 (6643): 639–48
- Richardson, Michael K, et l. 1998, Haeckel, Embryos, and Evolution. Science, Vol. 280, Iss. 5366, p. 983
- Gilbert, Scott F. 2000. "Homologous Pathways of Development". Developmental biology (6th ed.). Sunderland, Mass: Sinauer Associates.
- Sanetra et al. Conservation and co-option in developmental programmes: the importance of homology relationships. Front Zool (2005) vol. 2 pp. 15
- Fitch W. 1970. Distinguishing homologous from analogous proteins. Syst. Zool., 19:99–113
- Wiley EO, D Siegel-Causey, DR Brooks, VA Funk.1991 The Complete Cladistic: A primer of phylogenetic procedures.
- Fitch W. 2000. "Homology a personal view on some of the problems". Trends Genet. 16 (5): 227–31.
- Cresko et al. Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. Proc Natl Acad Sci USA (2004) vol. 101 (16) pp. 6050-5
- Duret L, Mouchiroud D, Gouy M (June 1994). "HOVERGEN: a database of homologous vertebrate genes". Nucleic Acids Res. 22 (12): 2360–5.
- Penel S, Arigon AM, Dufayard JF, et al. (2009). "Databases of homologous gene families for comparative genomics". BMC Bioinformatics 10 (Suppl 6): S3.
- Koonin EV (2005). "Orthologs, paralogs, and evolutionary genomics". Annu. Rev. Genet. 39: 309–38.
- Chen F, et al (January 2006). "OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups". Nucleic Acids Res. 34 (Database issue): D363–8.
- Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak MK, Douzery EJ (2007). "OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics". BMC Evol. Biol. 7: 241.
- Chiu JC, Lee EK, Egan MG, Sarkar IN, Coruzzi GM, DeSalle R (March 2006). "OrthologID: automation of genome-scale ortholog identification within a parsimony framework". Bioinformatics 22 (6): 699–707.
- Conte MG, Gaillard S, Lanau N, Rouard M, Périn C (January 2008). "GreenPhylDB: a database for plant comparative genomics". Nucleic Acids Res. 36 (Database issue): D991–8.
- Kitano et al. Reverse evolution of armor plates in the threespine stickleback. Current Biology (2008) vol. 18 (10) pp. 769-774