"PRINCIPLES OF PHYLOGENETICS: ECOLOGY AND EVOLUTION"

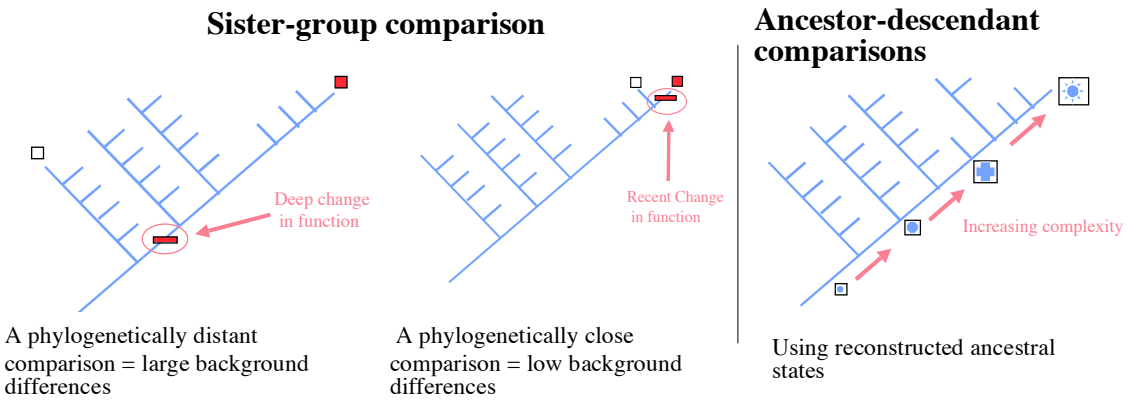
Integrative Biology 200B Spring 2011

University of California, Berkeley Nick Matzke, revised from B.D. Mishler

March 3, 2011. **Comparative genomics; Evolution and development**

This is the era of whole-genome sequencing; molecular data are becoming available at a rate unanticipated even a few years ago. Sequencing projects in a number of countries have produced a growing number of fully sequenced genomes, providing computational biologists with tremendous opportunities. However, comparative genomics has so far largely been restricted to pair-wise comparisons of genomes. The importance of taking a phylogenetic approach to systematically relating larger sets of genomes has only recently been realized.

A recent synthesis of phylogenetic systematics and molecular biology/genomics – two fields once estranged – is beginning to form a new field that could be called "phylogenomics" (Eisen 1998). Something can be learned about the function of genes by examining them in one organism. However, a much richer array of tools is available using a phylogenetic approach. Close sister-group comparisons between lineages differing in a critical phenotype (e.g., desiccation or freeze tolerance) can allow a quick narrowing of the search for genetic causes. Dissecting a complicated, evolutionarily advanced genotype/phenotype complex (e.g., development of the angiosperm flower), by tracing the components back through simpler ancestral reconstructions, can lead to quicker understanding. Hence, phylogenomics allows one to go beyond the use of pairwise sequence similarities, and use phylogenic comparative methods as discussed in this class to confirm and/or to establish gene function and interactions.



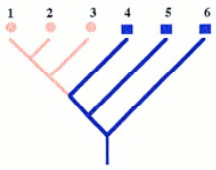
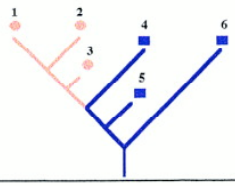
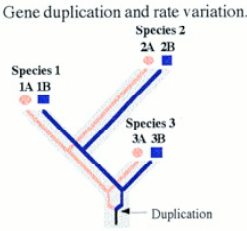
Most importantly for the systematist, the new comparative genomic data should also greatly increase the accuracy of reconstructions of the Tree of Life. Even though nucleotide sequence comparisons have become the workhorse of phylogenetic analysis at all levels, there are clearly phylogenetic problems for which nucleotide sequence data are poorly suited, because of their simple nature (having only four character states) and tendency to evolve in a regular, more-or-less clocklike fashion. In particular, "deep" branching questions (with relatively short internodes of interest mixed with long terminal branches) are notoriously difficult to resolve with DNA sequence data.

It is fortunate therefore, that fundamentally new kinds of structural genomic characters such as inversions, translocations, losses, duplications, and insertion/deletion of introns will be increasingly available in the future. These characters need to be evaluated using much the same

principles of character analysis that were originally developed for morphological characters. They must be looked at carefully to establish likely homology (e.g., examining the ends of breakpoints across genomes to see whether a single rearrangement event is likely to have occurred), independence, and discreteness of character states. Thus close collaboration between systematists and molecular biologists will be required to code these genomic characters properly, and to assemble them into matrices with other data types.

Next two figures from: Jonathan A. Eisen and Claire M. Fraser, Phylogenomics: Intersection of Evolution and Genomics , *Science*, Vol 300, Issue 5626, 1706-1707 , 13 June 2003

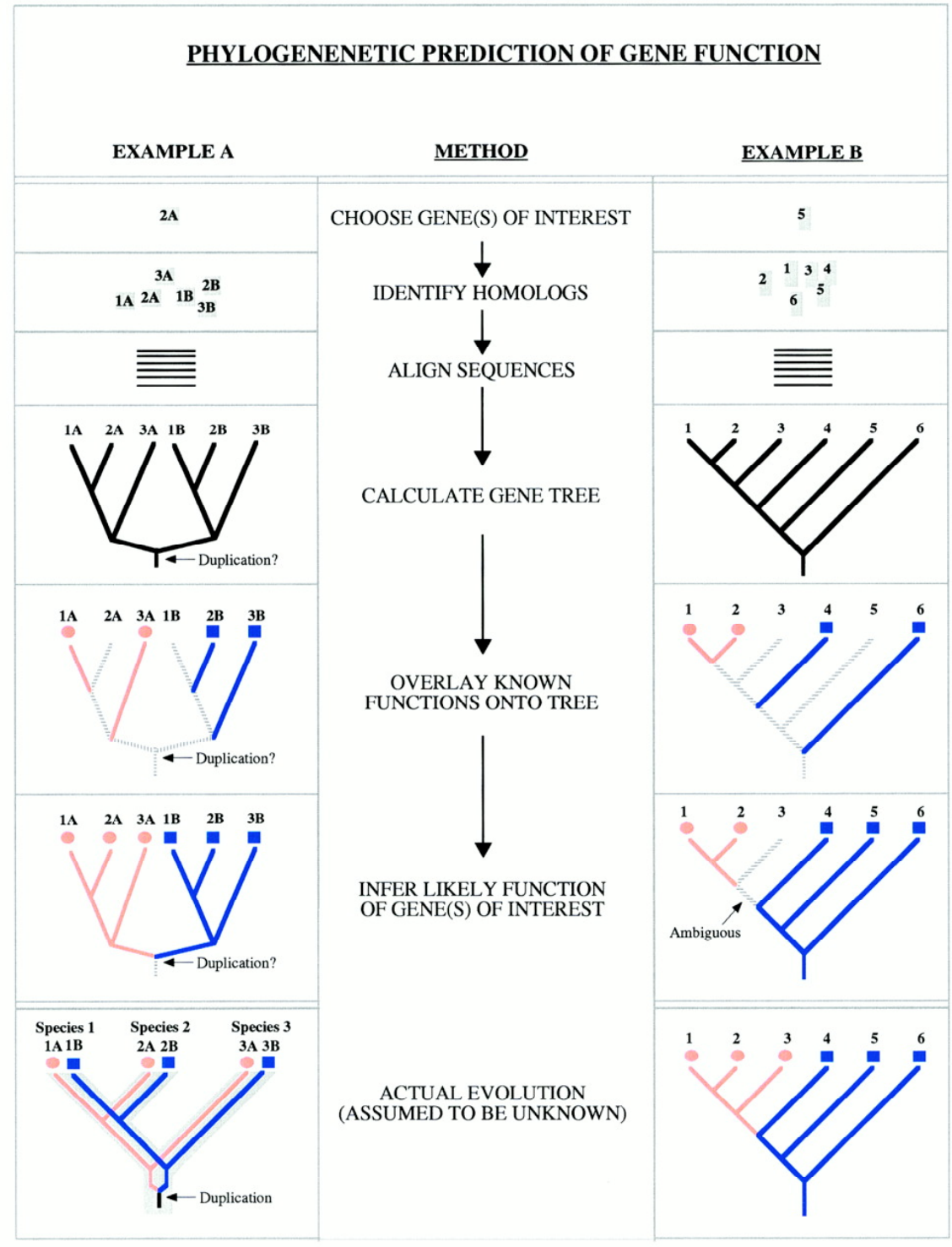
Table 4 Examples of Conditions in Which Similarity Methods Produce Inaccurate Predictions of Function

Evolutionary Pattern and Tree of Genes and Functions[1]	Gene With Unknown Function[2]	Highest Hit Method: Predicted Function[3]	Highest Hit Method: Accurate?	Phylogenomic Method: Predicted Function[4]	Phylogenomic Method: Accurate?	Comments
A. Functional change during evolution. 	1 ●	●	+	●	+	• Phylogenomic method cannot predict functions for all genes, but the predictions that are made are accurate. • Highest hit method is misleading because function changed among homologs but hierarchies of similarity do not correlate with the function (see Bolker and Raff 1996).
	2 ●	●	+	●	+	
	3 ●	●	+	●/■	±	
	4 ■	●	-	●/■	±	
	5 ■	●/■	±	■	+	
	6 ■	●/■	±	■	+	
B. Functional change & rate variation. 	1 ●	●	+	●	+	• Similarity based methods perform particularly poorly when evolutionary rates vary between taxa. • Molecular phylogenetic methods can allow for rate variation and reconstruct gene history reasonably accurately.
	2 ●	●	+	●	+	
	3 ●	■	-	●/■	±	
	4 ■	●	-	●/■	±	
	5 ■	●	-	■	+	
	6 ■	■	+	■	+	
C. Gene duplication and rate variation. 	1A ●	●	+	●	+	• Most-similarity based methods are not ideally set up to deal with cases of gene duplication since orthologous genes do not always have significantly more sequence similarity to each other than to paralogs (Eisen et al. 1995; Zardova et al. 1996; Tatusov et al. 1997). • Similarity-based methods perform particularly poorly when rate variation and gene duplication are combined. This even applies to the COG method (see Table1) since it works by classifying levels of similarity and not by inferring history. Nevertheless, the COG method is a significant improvement over other similarity based methods in classifying orthologs. • Phylogenetic reconstruction is the most reliably way to inferr gene duplication events and thus determine orthology.
	2A ●	●	+	●	+	
	3A ●	■	-	●	+	
	1B ■	■	+	■	+	
	2B ■	■	+	■	+	
	3B ■	●	-	■	+	

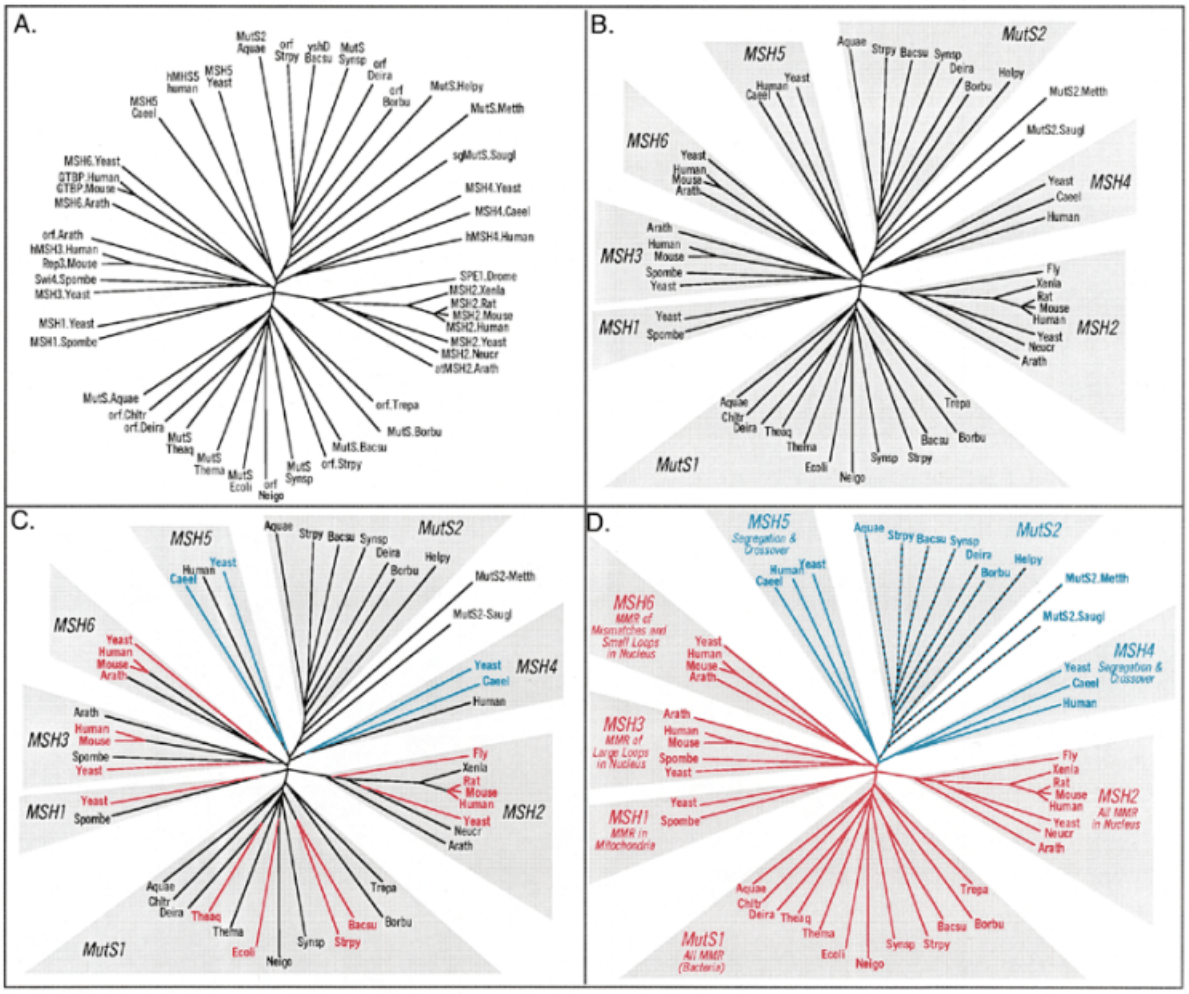
[1] The true tree is shown but it is assumed that it is not known. Different colors and symbols represent different functions. Numbers correspond to different species.
[2] The function of all other genes is assumed to be known.
[3] The top hit can be determined from the tree by finding the gene is the shortest evolutionary distance away (as determined along the branches of the tree).
[4] It is assumed that the tree of the genes can be reproduced accurately by molecular phylogenetic methods (see Fig. 1).

Outline of a phylogenomic methodology (next page). In this method, information about the evolutionary relationships among genes is used to predict the functions of uncharacterized genes (see text for details). Two hypothetical scenarios are presented and the path of trying to infer the function of two uncharacterized genes in each case is traced. (*A*) A gene family has undergone a gene duplication that was accompanied by functional divergence. (*B*) Gene function has changed in one lineage. The true tree (which is assumed to be unknown) is shown at the *bottom*. The genes are referred to by numbers (which

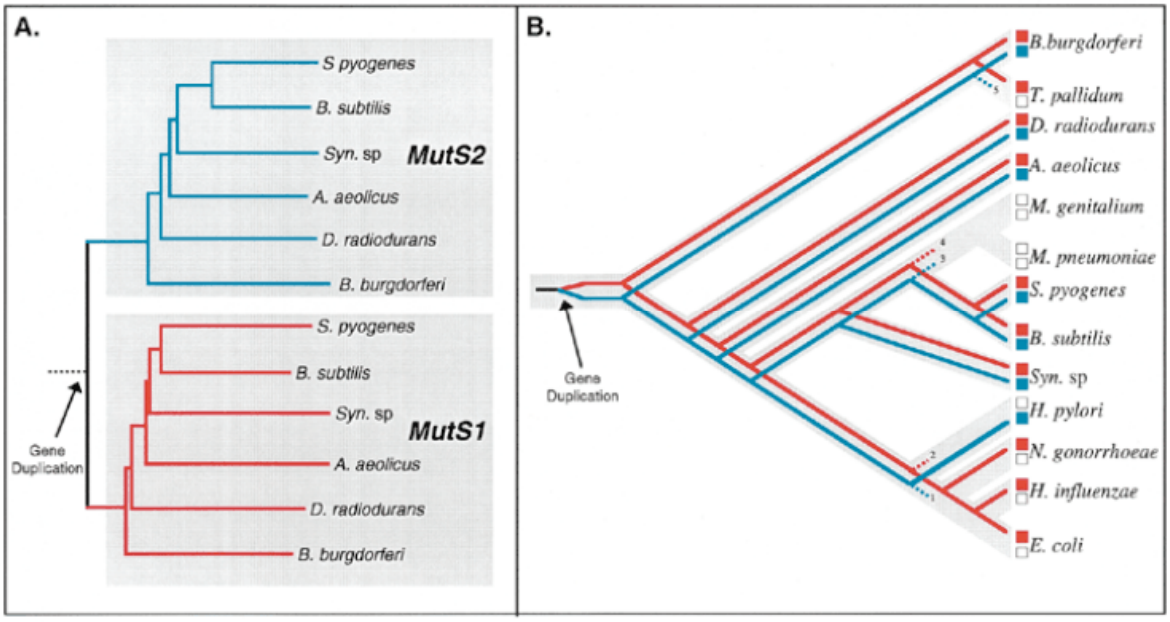
represent the species from which these genes come) and letters (which in *A* represent different genes within a species). The thin branches in the evolutionary trees correspond to the gene phylogeny and the thick gray branches in *A* (*bottom*) correspond to the phylogeny of the species in which the duplicate genes evolve in parallel (as paralogs). Different colors (and symbols) represent different gene functions; gray (with hatching) represents either unknown or unpredictable functions.



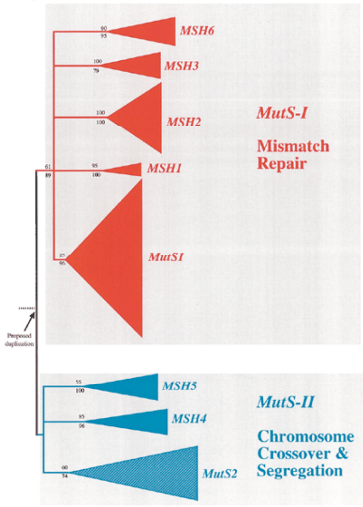
Example below taken from: **JA Eisen "A phylogenomic study of the MutS family of proteins"** Nucleic Acids Research, Vol 26, Issue 18 4291-4300.



Phylogenomic analysis of the MutS family of proteins. (**A**) Unrooted neighbor-joining tree of the proteins in the MutS family. (**B**) Proposed subfamilies of orthologs are highlighted. (**C**) Known functions of genes are overlaid onto the tree. For simplicity, only two colors are used, red for mismatch repair and blue for meiotic-crossing over and chromosome segregation. (**D**) Prediction of functions of uncharacterized proteins based on position in the tree.



Gene duplication and gene loss in the history of the bacterial MutS homologs. (**A**) Neighbor-joining phylogenetic tree of the *MutS1* and *MutS2* subfamilies (using only those proteins from species with both). The identical topology of the tree in the two subfamilies suggests the occurrence of a duplication prior to the divergence of these bacteria. (**B**) Gene loss within the bacteria. Gene loss was determined by overlaying the presence and absence of MutS1 and MutS2 orthologs onto the tree of the species for which complete genomes are available (since only with a complete genome sequence can one be relatively certain that a gene is absent from a species). The thick gray lines represent the evolutionary history of the species based on a combination of the MutS and rRNA trees for these species. The thin colored lines represent the evolutionary history of the two MutS subfamilies (*MutS1* in red and *MutS2* in blue). Branch lengths do not correspond to evolutionary distance. Gene loss is indicated by a dashed line and each loss is labeled by a number: 1, MutS2 loss in enterobacteria; 2, MutS1 loss in *H.pylori*; 3, MutS2 loss in the mycoplasmas; 4, MutS1 loss in the mycoplasmas; and 5, MutS2 loss in *T.pallidum*.



Consensus phylogenetic tree of MutS family of proteins. Branches with low bootstrap values or that were not-identical in trees generated with different methods were collapsed. Only the proposed subfamilies are shown (sequences in each group are listed in Table 1). In addition, two proteins that are related to the *MutS2* subfamily are grouped with it. The height of each subgroup corresponds to the number of sequences in that group and the width corresponds to the longest branch length within the group. Bootstrap values for specific nodes are listed when >40% (neighbor-joining on the top, parsimony on the bottom). The root of the tree was assigned as discussed in the text between the groups labeled *MutS-I* and *MutS-II*. Conserved functions for the different groups are listed.

Other topics

You could have a whole course on new developments in phylogenomics, comparative genomics, etc. Here are just some of the highlights of recent events.

"Genome-phylogenetics"

I just made up this term, to distinguish this from "phylogenomics", but there is an emerging field devoted to simultaneously estimating gene trees and species trees. Bastien Bousseau, a postdoc in the Huelsenbeck lab, is a leader here.

Boussau, B. and V. Daubin (2010). "Genomes as documents of evolutionary history." Trends in Ecology & Evolution **25**(4): 224-232.

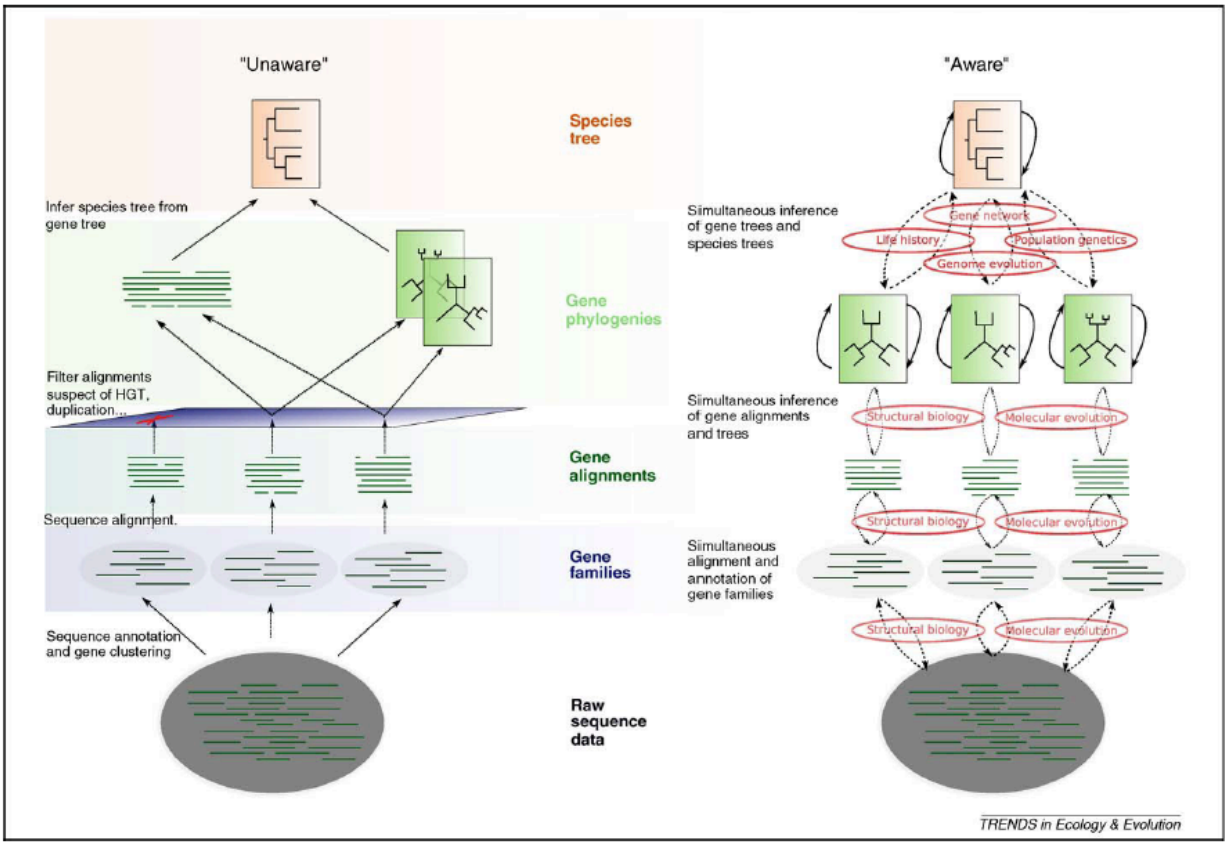


Figure 1. Phylogenetic awareness: the two paths from sequences to an organism tree. In the 'unaware' path (the traditional way of inferring species phylogenies) each stage of the phylogenetic inference is essentially independent from the steps upstream and downstream. In addition, sequence alignments have to pass different filters in order to make gene trees readily understandable as organism trees (absence of duplicates, lateral gene transfer (LGT), etc.). In contrast, the 'aware' path models the dependency and degree of complexity between each step using knowledge from different fields of biology (red ellipses, the list is not exhaustive). Alignments can be statistically estimated simultaneously with gene trees using models of sequence evolution that incorporate insertion and deletion events; and models of gene family evolution incorporating LGT, duplication and/or incomplete lineage sorting specify the dependency between gene trees and organism tree. Two-way arrows represent these dependencies, and solid arrows represent gene tree and organism tree searches. The dependency between gene family annotation, alignment and phylogenetics has not yet been explored, but could theoretically be modeled (see text for discussion). The schematic representation of the synchronous search for organism trees, gene trees, gene alignments and others suggests an obvious architecture for parallelizing this search.

Table 1. Soft-aware

Software name	Description	Web link	Ref.
Alignment and phylogeny			
BAli-Phy (Bayesian Alignment and Phylogeny estimation)	Bayesian program to reconstruct alignments and phylogenetic trees.	http://www.biomath.ucla.edu/msuchard/bali-phy/index.php	[9]
StatAlign	Bayesian program to reconstruct alignments and phylogenetic trees.	http://phylogeny-cafe.elte.hu/StatAlign/	[69]
SimulFold	Bayesian program to reconstruct RNA structural alignment as well as phylogenetic trees.	http://www.cs.ubc.ca/~irmtraud/simulfold/	[70]
SAPF (Statistical Aligner, Phylogenetic Footprinter)	Bayesian program that samples alignments of non-coding sequences given a phylogenetic tree and predicts functional regions, i.e. regions that are particularly well conserved.	http://www.stats.ox.ac.uk/~satija/SAPF/	[16]
Dart (DNA, Amino and RNA Tests)	Software package to build and analyze alignments and phylogenetic trees through transducers notably, for sequences as well as RNA secondary structures.	http://biowiki.org/DART	[5]
Prank (Probabilistic Alignment Kit)	Phylogenetic-aware tool permitting the alignment of multiple sequences given a phylogenetic tree. Contrary to classical heuristics, it distinguishes insertions from deletions and thus has shown higher alignment accuracy.	http://www.ebi.ac.uk/goldman-srv/prank/prank/	
SATé (simultaneous alignment and tree estimation)	An automated method to quickly and accurately estimate both DNA alignments and trees with the maximum likelihood criterion [9].	http://www.cs.utexas.edu/~kliu/public/sate_journal.html	[2]
Species and gene trees			
Best (Bayesian Estimation of Species Tree):	Bayesian program to reconstruct species trees from gene alignments accounting for trans-specific polymorphisms.	http://www.stat.osu.edu/~dkp/BEST/	[22]
Bucky (Bayesian Untangling of Concordance Knots)	Bayesian program permitting analysis of several gene families simultaneously, accounting for some correlations between gene histories through gene-to-trees maps.	http://www.stat.wisc.edu/~larget/bucky.html	[39]
Prime	Set of software applications that can be used to analyze gene families in the presence of duplications and losses given a known species tree.	http://prime.sbc.su.se/	[25]
Inversions and phylogeny			
Badger (Bayesian Analysis to Describe Genomic Evolution by Rearrangement)	Badger is a Bayesian program to analyze genomic evolution through inversions.	http://badger.duq.edu/	[41]
Character evolution			
Sifter (Statistical Inference of Function Through Evolutionary Relationships)	Sifter predicts the function of genes in a gene family based on a model of function evolution and a phylogenetic tree of the gene family.	http://sifter.berkeley.edu/	[26]
BayesTraits	Bayesian program allowing one to analyze the evolution of discrete or continuous characters on a distribution of phylogenies.	http://www.evolution.reading.ac.uk/BayesTraits.html	[57]
Ape (Analysis of Phylogenetics and Evolution):	Package of functions to use in the R statistical software. Ape notably permits analyzing the evolution of discrete or continuous characters on a phylogeny, or studying shapes of phylogenies.	http://ape.mpl.ird.fr/	[71]

Box 2. The myth of 'orthologous gene families'

Coined by Walter Fitch [62], the term 'ortholog' designates genes that are related through speciation events, as opposed to 'paralogs', which are the result of duplications. Therefore, to reconstruct a phylogeny of species, one could use orthologous genes. However, the identification of orthologous genes is not always unequivocal (Figure I). First, phylogeneticists usually rely on the absence of duplicated copies in the datasets under study, but duplications could have occurred during the history of a gene family without leaving obvious traces. This is particularly dramatic in the event of reciprocal losses, when two species lose different copies of an ancestrally duplicated gene. The impact of this phenomenon, known as a hidden paralogy, is difficult to estimate on a large scale, but reciprocal losses have been shown to be frequent after whole genome duplications in yeasts and fish [63]. Second, lateral gene transfer (LGT) has been shown to be pervasive throughout the history of life. Therefore, it is unsafe to assume *a priori* that the history of a gene is devoid of such events, whatever its function. Third, even genes that would be considered genuine orthologs might not retrace the history of species; the persistence of different allelic forms of a gene during long periods of time relative to the lapse between speciation events, a phenomenon known as trans-specific polymorphisms (TSP) [20], can result in differences among gene trees (incomplete lineage sorting) even in the absence of paralogy or LGT. The assemblage of these processes makes it difficult to expect that a single gene history would faithfully mirror a tree of species throughout several billion years of evolution. In addition to these biological problems, even the most advanced phylogenetic methods are often unable to accurately model the evolution of biological sequences, which can result in the inference of erroneous trees. There is no, and will never be, a perfect dataset, devoid of lateral gene transfer, incomplete lineage sorting, hidden or apparent paralogies, convergent gene losses or systematically biased or accelerated evolutionary rates. As the impact of most of these processes is only expected to increase with more data, it is necessary to exploit the evolutionary significance of these events rather than discard them.

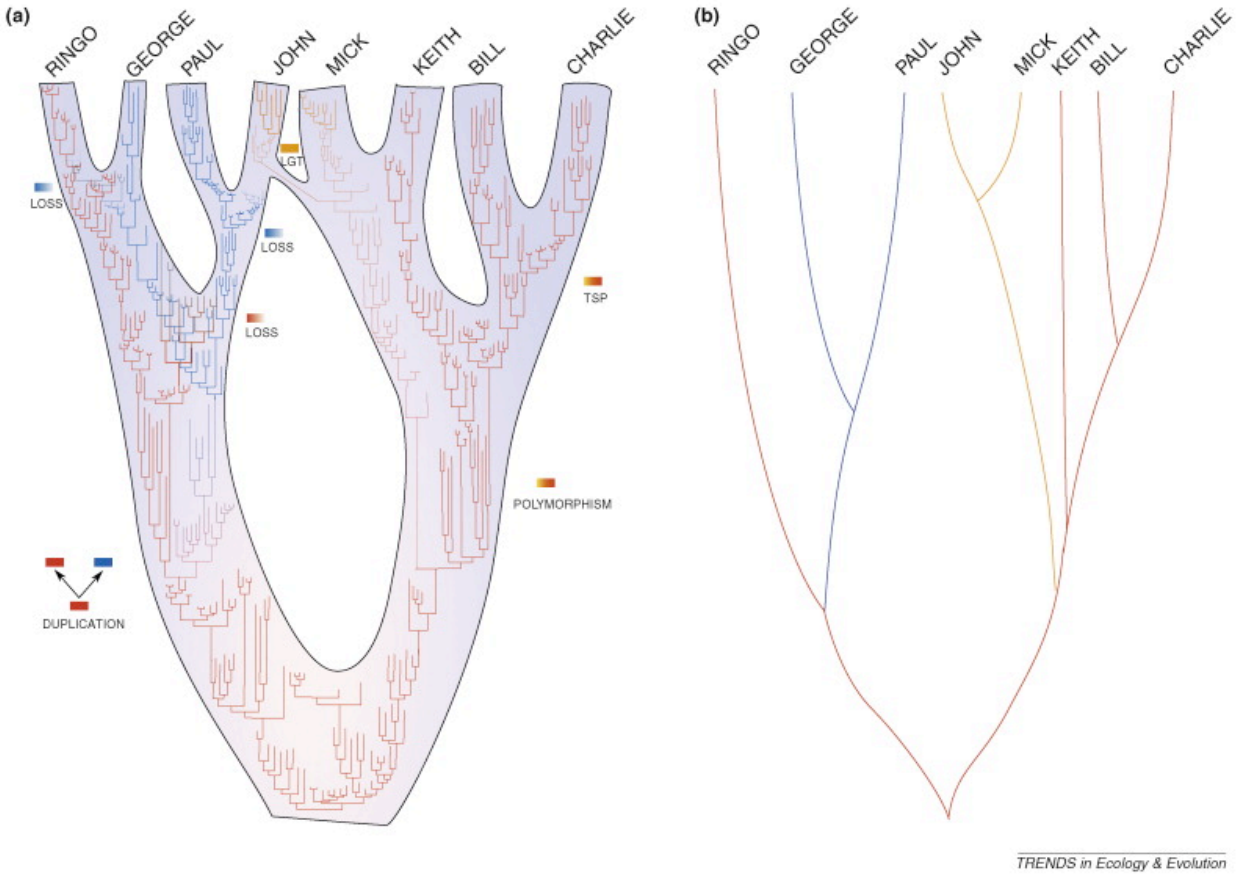
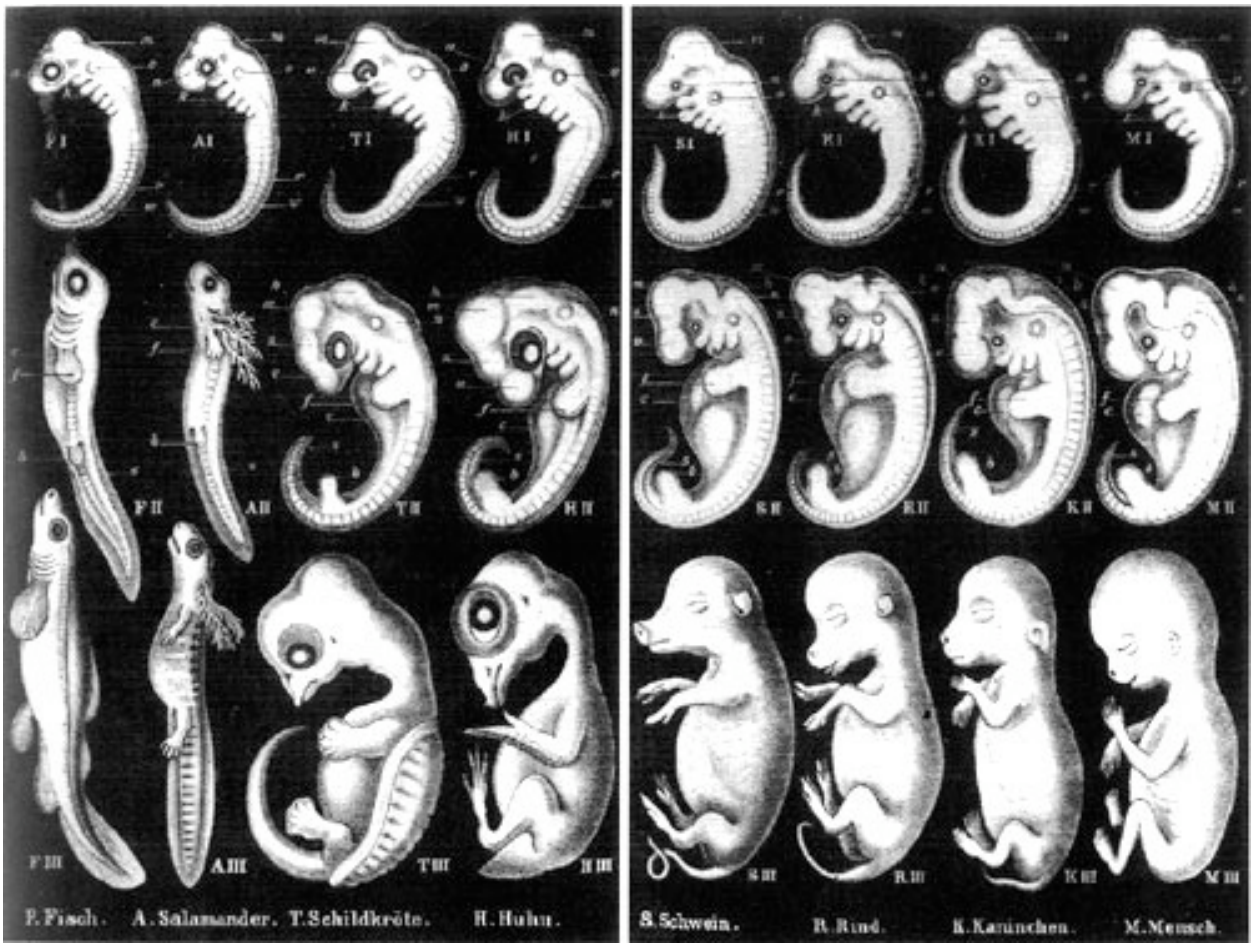


Figure I. Various processes can generate discordance between organism and gene trees. **(a)** A tree depicting the relationships of eight species. Ringo and George, Paul and John are on one side of the root, and Mick and Keith, Bill and Charlie on the other. The history of a gene family is depicted within the bounds of this organism tree, and processes acting at the genome level (duplication, loss, gene transfer) as well as population level (polymorphism) are shown. **(b)** The gene tree reconstructed from this gene family shows a topology that conflicts with the organism tree. Following a duplication and losses, George and Paul are grouped together, a gene transfer groups John and Mick, and trans-specific polymorphism leads to Keith being clustered with Bill and Charlie. Processes from population genetics and from genome dynamics both affect gene histories; models of gene family evolution could help reconstruct gene phylogenies, organism history and genome evolution.

We are rapidly reaching the point where instead of just sequencing a few genes, everything will have its genome sequenced. Dealing with this flood of data is a major issue.

Genomics and evo-devo

Everyone has probably seen Haeckel's famous embryo drawing (This is from Haeckel, Anthropogenie, 1874 1st edition I believe):



Many people have probably also seen claims that Haeckel committed "fraud" with this drawing, and mislead generations of scientists. E.g., *Science* reported this in 1997:

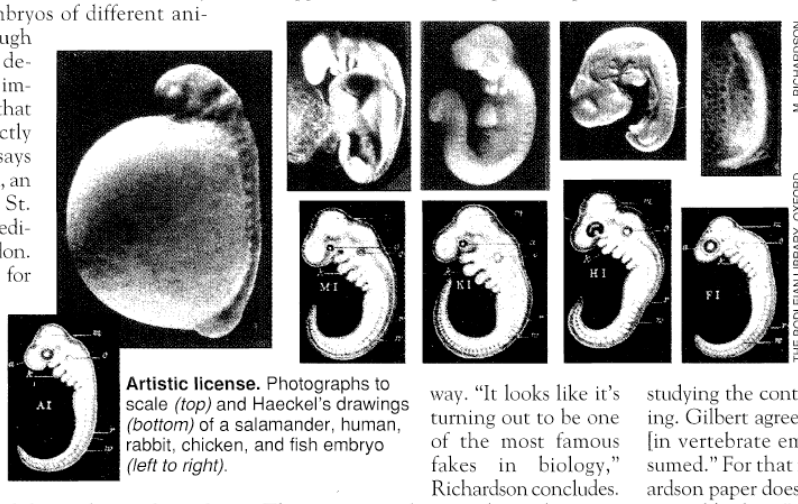
www.sciencemag.org • SCIENCE • VOL. 277 • 5 SEPTEMBER 1997

DEVELOPMENTAL BIOLOGY

Haeckel's Embryos: Fraud Rediscovered

Generations of biology students may have been misled by a famous set of drawings of embryos published 123 years ago by the German biologist Ernst Haeckel. They show vertebrate embryos of different animals passing through identical stages of development. But the impression they give, that the embryos are exactly alike, is wrong, says Michael Richardson, an embryologist at St. George's Hospital Medical School in London. He hopes once and for all to discredit Haeckel's work, first found to be flawed more than a century ago.

Richardson had long held doubts about Haeckel's drawings because they didn't square with his understanding of the rates at which fish, reptiles, birds, and group of animals. In reality, Richardson and his colleagues note, even closely related embryos such as those of fish vary quite a bit in their appearance and developmental pathway. "It looks like it's turning out to be one of the most famous fakes in biology," Richardson concludes.



M. RICHARDSON; THE BODLEIAN LIBRARY, OXFORD

Artistic license. Photographs to scale *(top)* and Haeckel's drawings *(bottom)* of a salamander, human, rabbit, chicken, and fish embryo *(left to right)*.

This news might not have been so shocking to Haeckel's peers in Germany a century ago: They got Haeckel to admit that he relied on memory and used artistic license in preparing his drawings, says Scott Gilbert, a developmental biologist at Swarthmore College in Pennsylvania. But Haeckel's confession got lost after his drawings were subsequently used in a 1901 book called *Darwin and After Darwin* and reproduced widely in English-language biology texts.

The flaws in Haeckel's work have resurfaced now in part because recent discoveries showing that many species share developmental genes have renewed interest in comparative developmental biology. And while some researchers—following Haeckel's lead—like to emphasize the similarities among species, Richardson thinks studying the contrasts may be more interesting. Gilbert agrees: "There is more variation [in vertebrate embryos] than had been assumed." For that reason, he adds, "the Richardson paper does a great service to developmental biology."

–Elizabeth Pennisi

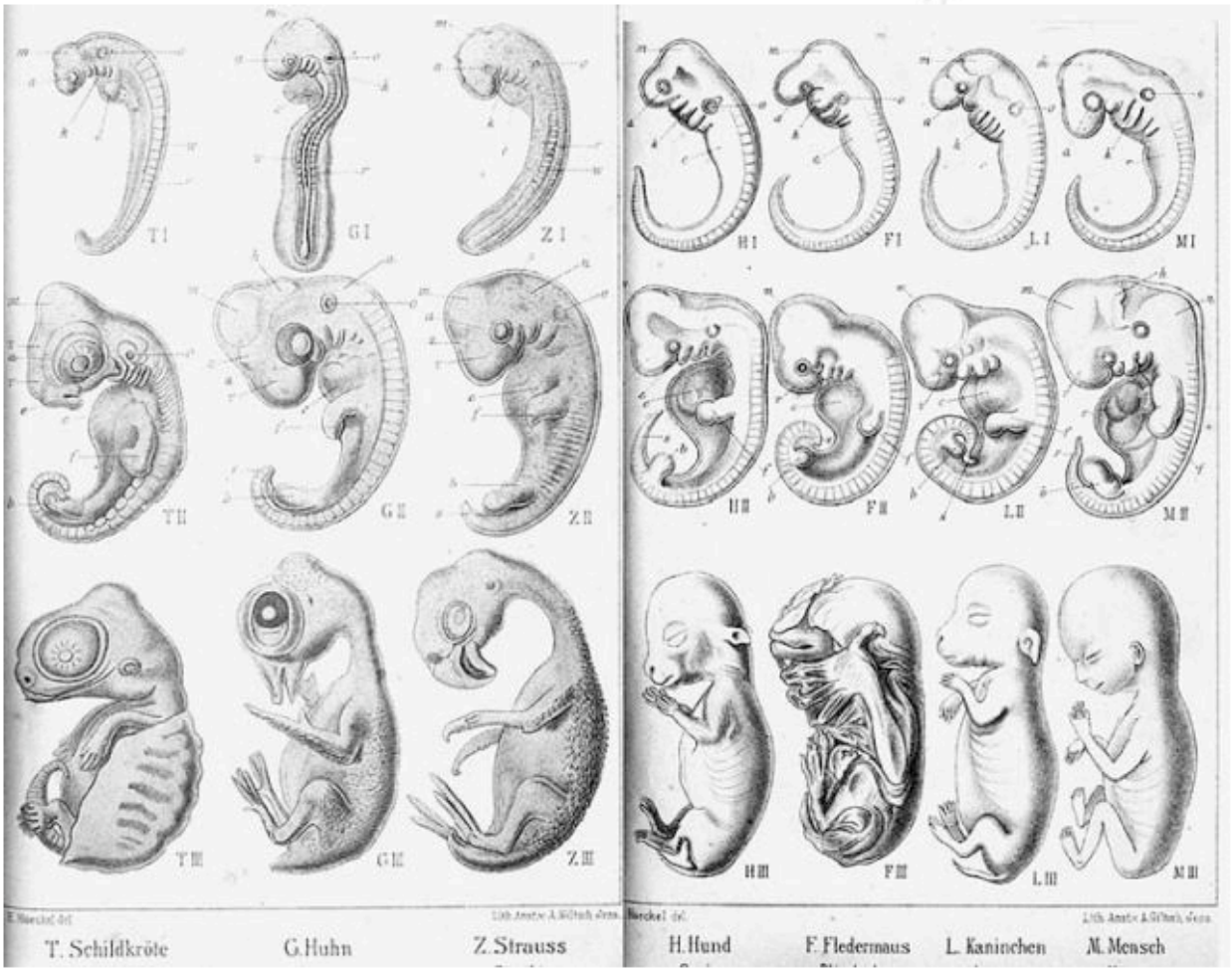
…you can imagine the mileage that creationists got out of this!

However, the truth is much more complicated. Haeckel was limited by the material that was available at the time, he updated his drawings as time went on, and more or less by accident it was one version of an older drawing that got into an English-language textbook and was widely copied. Read:

1. Biography by Robert Richards, *The Tragic Sense of Life*, 2008
2. Hopwood, N. (2006). "Pictures of evolution and charges of fraud." Isis **97**(2): 260-301.

Here is a later drawing, never cited by creationists or other critics:

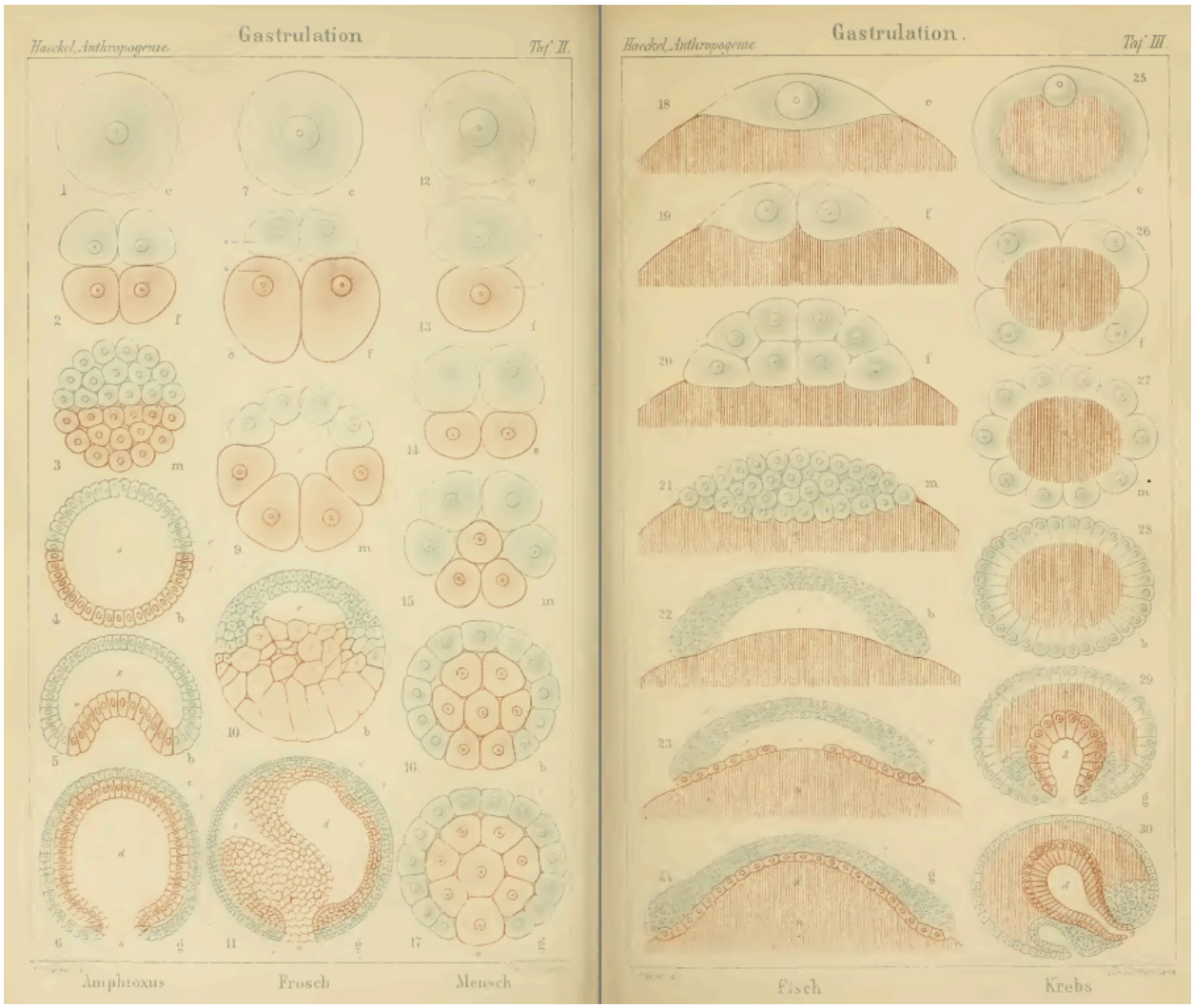
Illustrations from Ernst Haeckel, Anthropogenie, 4th ed. (1891)



http://ncse.com/image/illustrations-from-ernst-haeckel-anthropogenie-4th-ed-1891

…Haeckel has ditched the most inaccurate part of the drawing, which was not any of the tetrapod stages, but the "lower" vertebrates.

Haeckel is also criticized for saying that the earliest stages of development get more and more similar, which is indeed false. But Haeckel knew this! He even diagrammed the quite divergent embryos that occur before gastrulation:



Alan Gishlick, 2003, National Center for Science Education, Figure 8 at:
http://ncse.com/creationism/analysis/icons-evolution-figures

Anyway, the whole situation gets much clearer when photos of different embryo stages are made available, and then (crucially) put in phylogenetic context:

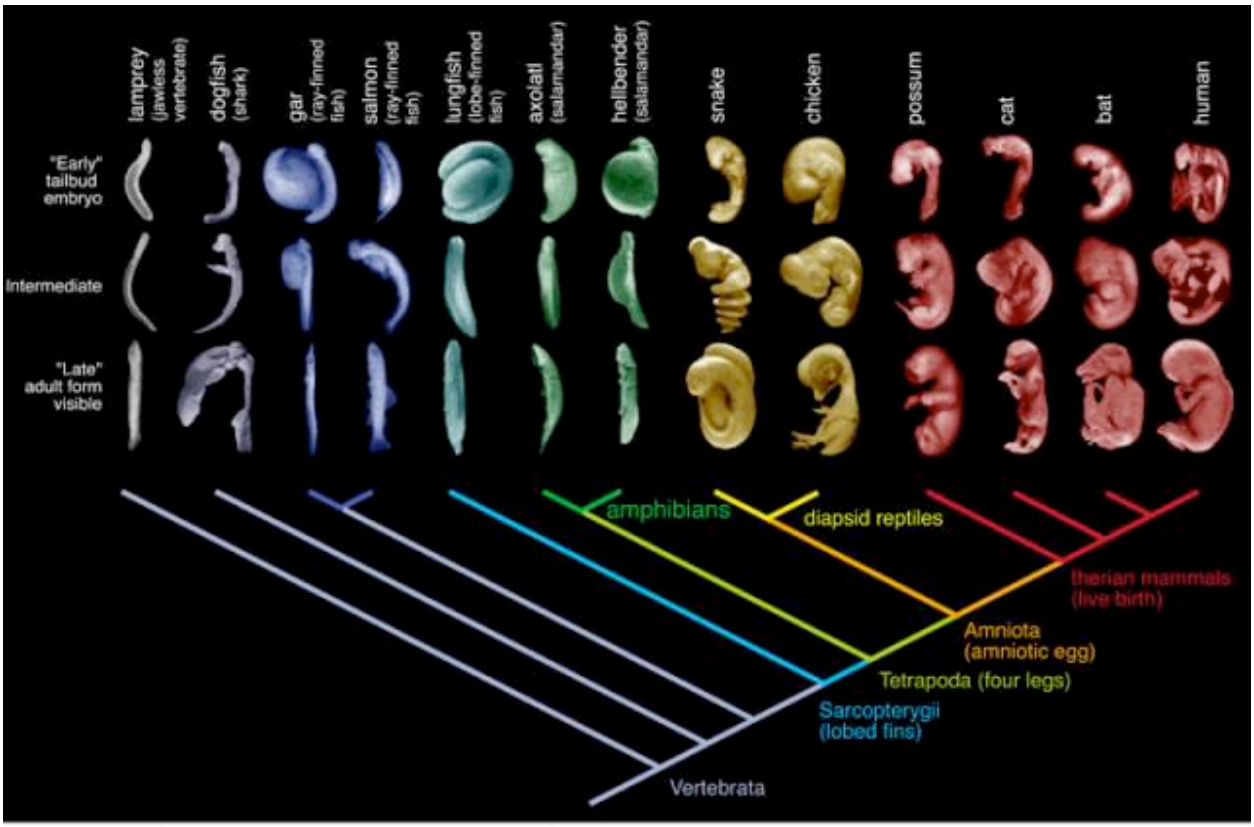


Figure 8. Developmental sequences of various vertebrates shown in phylogenetic context. Note the shared similarities of some closely related taxa, particularly the amniotes (modified from Richardson et al. 1998.)

I go through all of this for background and general education. The generalization that has developed and been widely accepted (with caveats!) is the "developmental hourglass" – that morphology seems to be more conserved at the "phylotypic stage" compared to before or after.

In 2010, we had this Nature cover:

NEWS & VIEWS

EVOLUTIONARY BIOLOGY

Genomic hourglass

Comparative genomics studies reveal molecular signatures of the controversial 'phylotypic' stage — a time when embryos of members of an animal phylum all look more alike than at other embryonic stages. SEE LETTERS P.811 & P.815

BENJAMIN PRUD'HOMME & NICOLAS GOMPEL

Most people would say that lizards and elephants bear little resemblance to each other. But not so the embryologist, for, at a particular stage in development, the embryos of very different species may look much the same. Elsewhere in this issue, papers by Kalinka *et al.*[1] and Domazet-Lošo and Tautz[2] offer a fresh perspective on this intriguing phenomenon.

This is a topic with a long history. In 1828, the German biologist Karl von Baer, one of the fathers of embryology, reported how very similar the early embryos of different species can be[3]: "I have two small embryos preserved in alcohol, that I forgot to label. At present I am unable to determine the genus to which they belong. They may be lizards, small birds, or even mammals." In fact, it was later observed that, over the course of development, the youngest embryos within an animal phylum often look very different, but progressively converge towards a similar form (described by von Baer and later dubbed the phylotypic

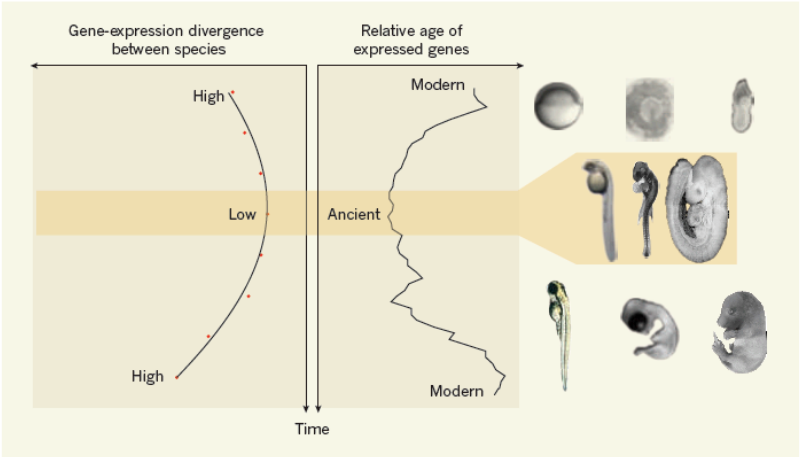


Figure 1 | The developmental hourglass, as revealed by comparative genomics. Mid-embryogenesis is marked by the phylotypic stage, a period of minimal anatomical divergence between species, as illustrated for vertebrate species by the orange band. This stage is now shown by Kalinka *et al.*[1] to display minimal gene-expression divergence between *Drosophila* species (left curve), and by Domazet-Lošo and Tautz[2] to express the oldest gene set of the entire life cycle (right curve). The species depicted, left to right, are zebrafish, chick and mouse. (Images reproduced from refs 12–14.)

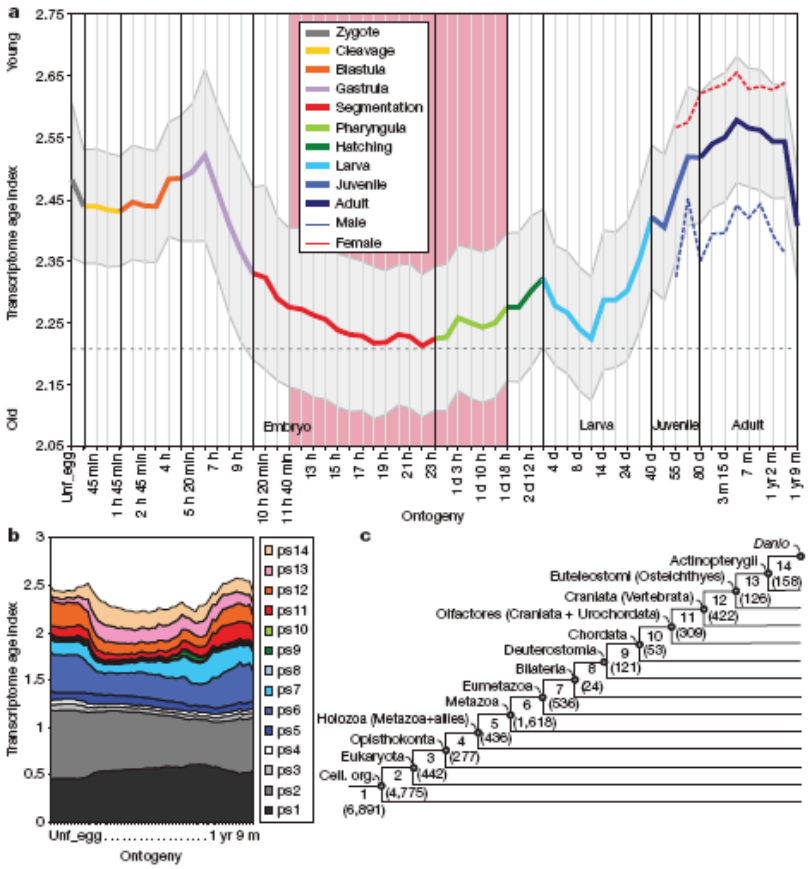


Figure 1 | Transcriptome age profiles for the zebrafish ontogeny. a, Cumulative transcriptome age index (TAI) for the different developmental stages. The pink shaded area represents the presumptive phylotypic phase in vertebrates. The overall pattern is significant by repeated measures ANOVA ($P = 2.4 \times 10^{-15}$, after Greenhouse–Geisser correction $P = 0.024$). Grey shaded areas represent ± the standard error of TAI estimated by bootstrap analysis. b, Transcriptome indices split according to the origin of the genes from the different phylostrata, based on the same developmental series as in a. c, Depiction of the phylostrata analysed; numbers in parentheses denote the number of array probes analysed for each phylostratum.

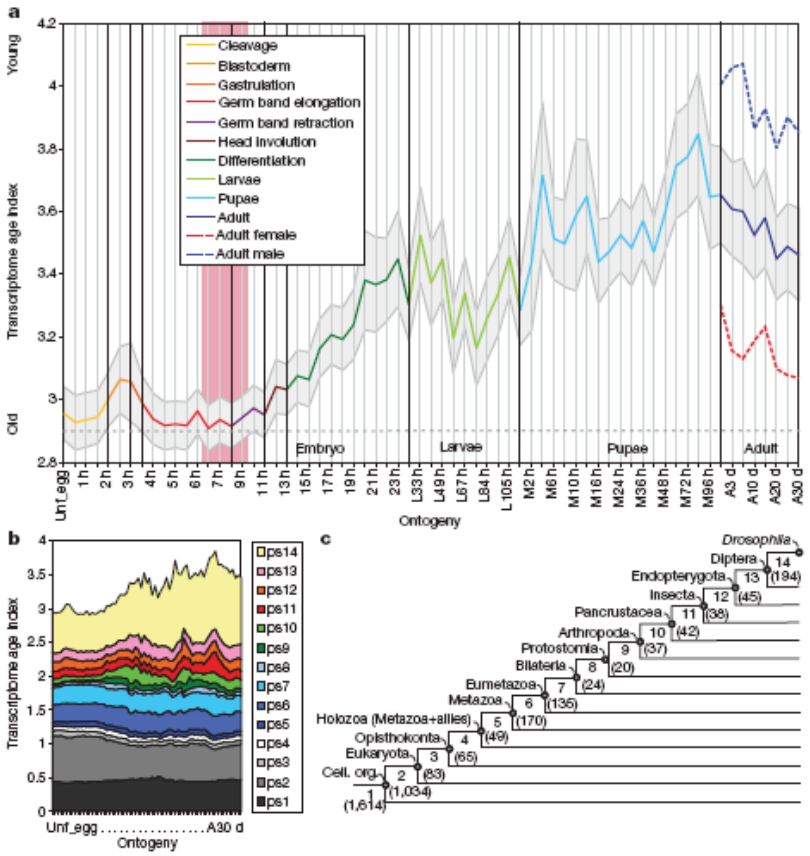


Figure 3 | Transcriptome age profiles for the *Drosophila* ontogeny, based on the data in ref. 23. **a**, Cumulative transcriptome index for the different developmental stages. The pink shaded area represents the presumptive phylotypic phase in insects. The overall pattern of differences in TAI is significant by repeated measures ANOVA ($P = 2.5 \times 10^{-93}$, after Greenhouse–Geisser correction $P = 1.22 \times 10^{-15}$). Grey shaded areas represent ± the standard error of TAI estimated by bootstrap analysis. **b**, **c** Same as for Fig. 1b, c.

References

Prud'Homme, B. and N. Gompel (2010). "Evolutionary biology: Genomic hourglass." Nature **468**(7325): 768-769.

Domazet-Loöo, T. and D. Tautz (2010). "A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns." Nature **468**(7325): 815-818.

Kalinka, A. T., K. M. Varga, et al. (2010). "Gene expression divergence recapitulates the developmental hourglass model." Nature **468**(7325): 811-814.