"PRINCIPLES OF PHYLOGENETICS: ECOLOGY AND EVOLUTION"

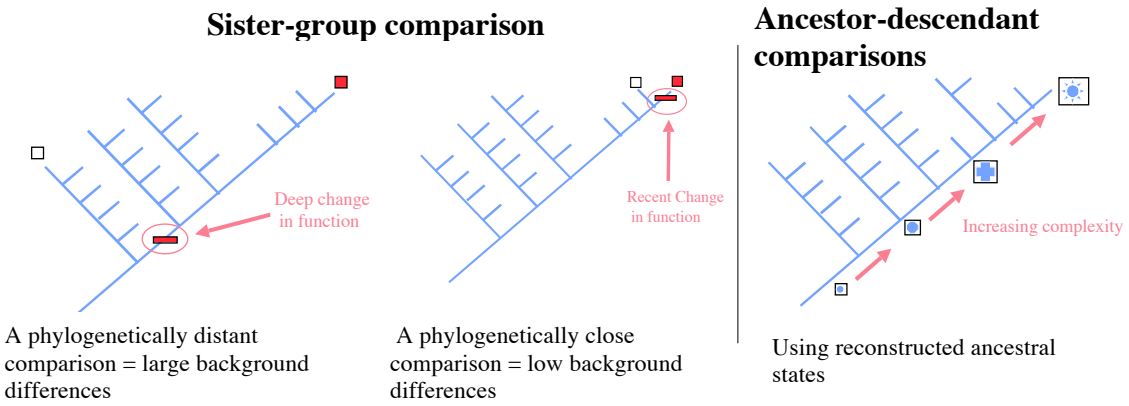
Integrative Biology 200B Spring 2009

University of California, Berkeley B.D. Mishler

Feb. 26, 2009. **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.



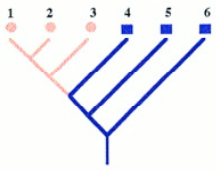
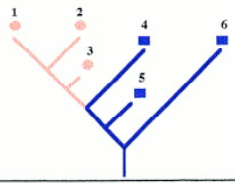
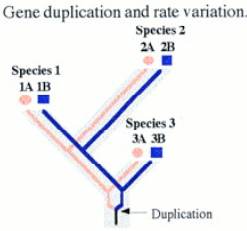
Most importantly for the systematist, the new comparative genomic data should also greatly increase the accuracy of reconstructions of the Tree of Life. Even though nucleotide sequence comparisons have become the workhorse of phylogenetic analysis at all levels, there are clearly phylogenetic problems 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

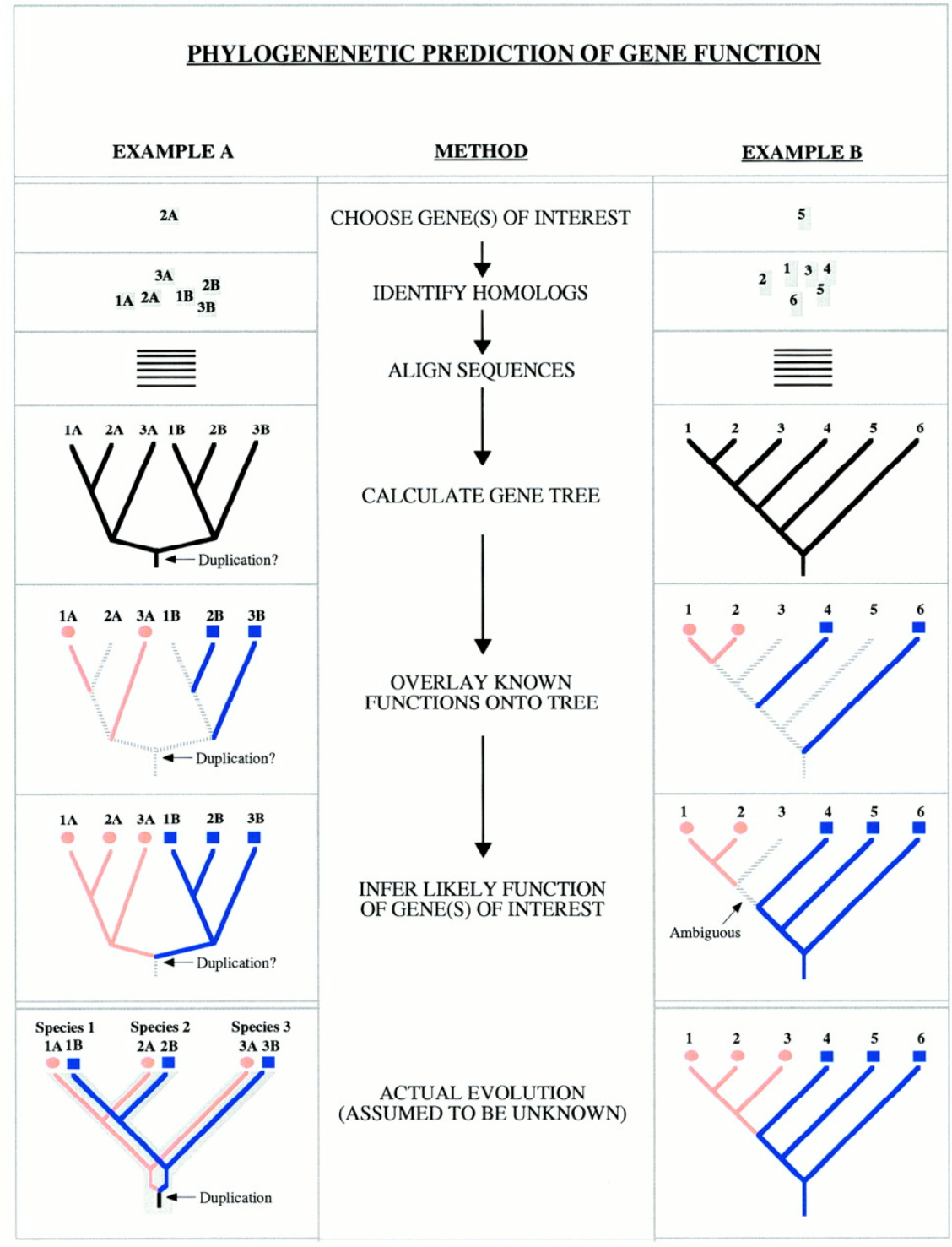
Table 4 Examples of Conditions in Which Similarity Methods Produce Inaccurate Predictions of Function

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 ■	●	-	■	+	

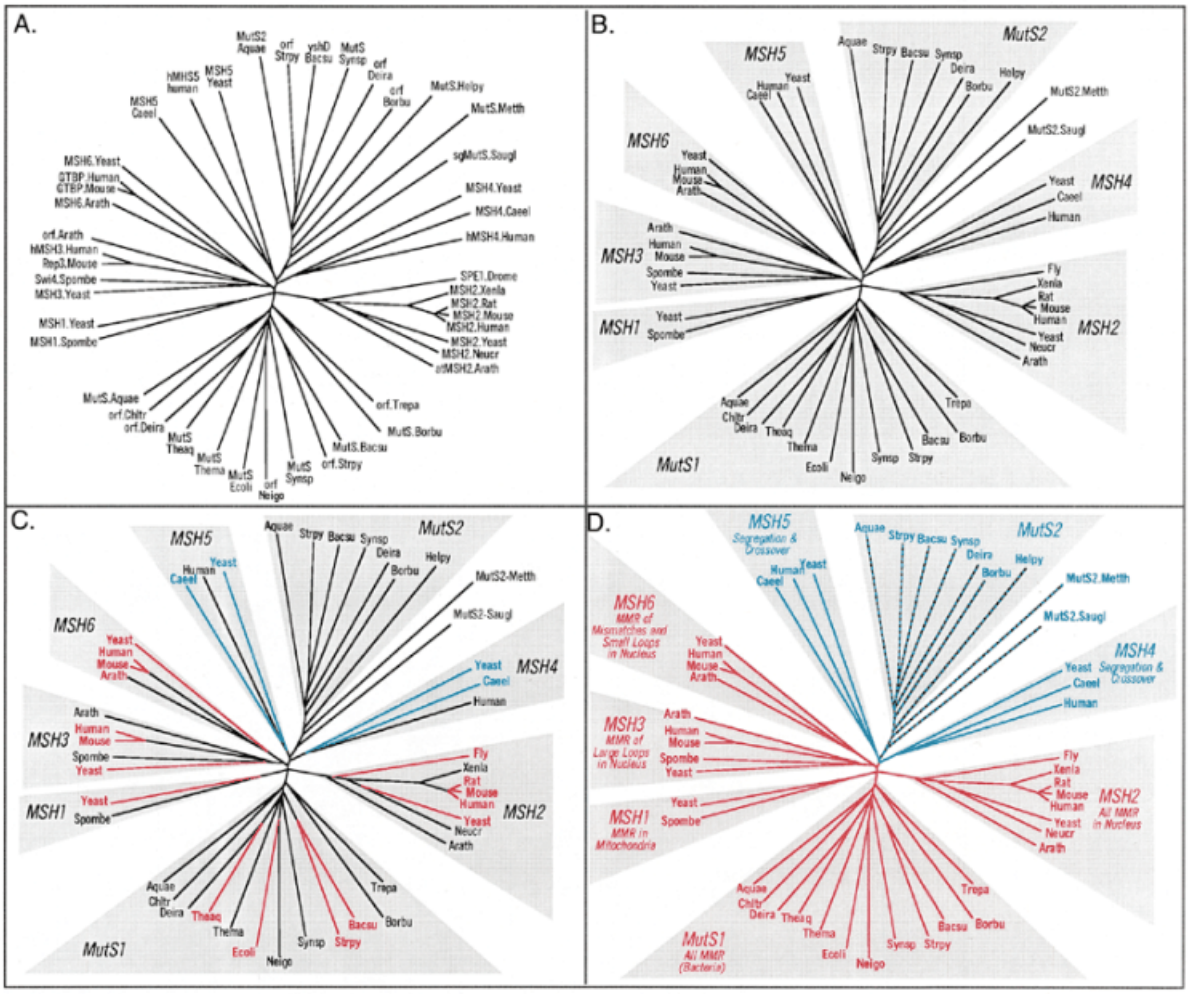
[1] The true tree is shown but it is assumed that it is not known. Different colors and symbols represent different functions. Numbers correspond to different species.
[2] The function of all other genes is assumed to be known.
[3] The top hit can be determined from the tree by finding the gene is the 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

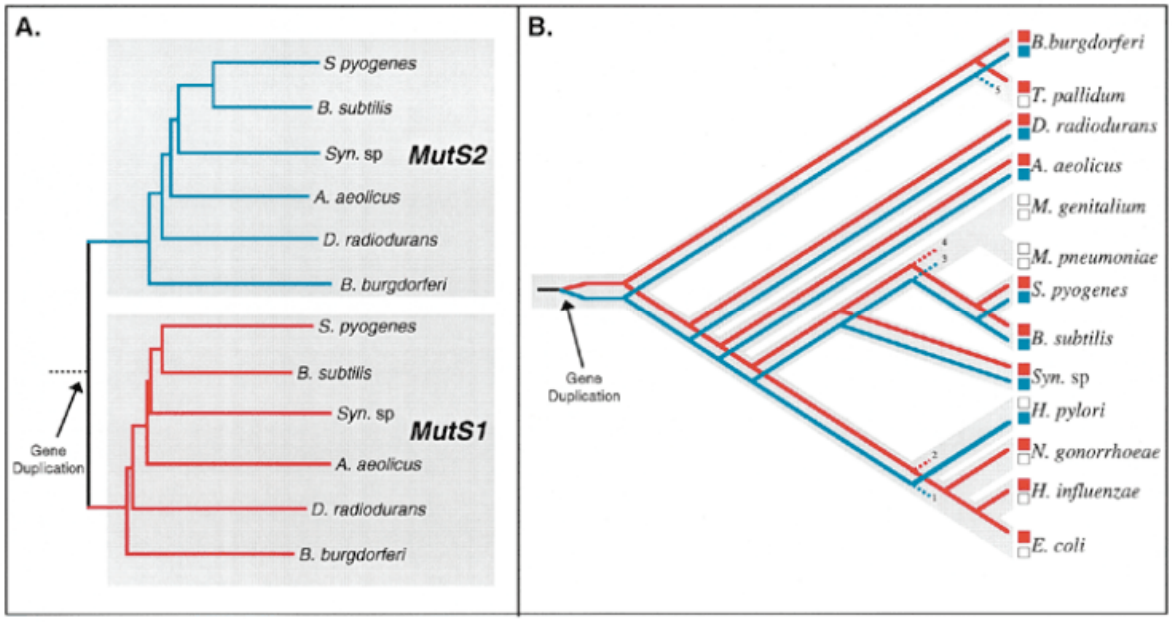
represent the species from which these genes come) and letters (which in *A* represent different genes within a species). The thin branches in the evolutionary trees correspond to the gene phylogeny and the thick gray branches in *A* (*bottom*) correspond to the phylogeny of the species in which the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



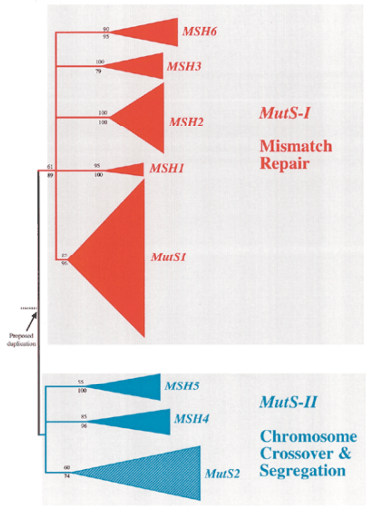
Example below taken from: **JA Eisen "A phylogenomic study of the MutS family of proteins"** Nucleic Acids Research, Vol 26, Issue 18 4291-4300.



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.

Evolution and development ("evo-devo")

The last frontier in our understanding of biological forms is an understanding of their developmental origins. Much of the ultimate control of form resides in the genome, yet much also resides in the environment (at levels from the internal cellular environment to the external habitat). The highly interactive and complex nature of developmental processes make it impractical to deduce phenotype from genotype based on first principles. We need to carefully keep in mind what we mean by "homology" as well. The phenotype is an emergent property and its origin can be studied most efficiently by backtracking from the phenotype itself to its structural, physiological, developmental, ecological, and genetic causes.

Ontogeny and phylogeny revisited

The relation between ontogeny and phylogeny has been of longstanding interest to biologists, and continues to be a timely topic. It is important of course to take a comparative approach to development, within a phylogenetic framework. Our aims are to reconstruct both the developmental pathway taken by a given species for a given structure, and the manner in which the developmental system evolved. Some terminology (see Humphries 1988 for details):

Heterotopy -- evolutionary change in the position of development

Heterochrony -- evolutionary change in the timing of development (see figure later)

Peramorphosis (Hypermorphosis vs. Acceleration vs. Predisplacement)

Paedomorphosis (Progenesis vs. Neoteny vs. Postdisplacement)

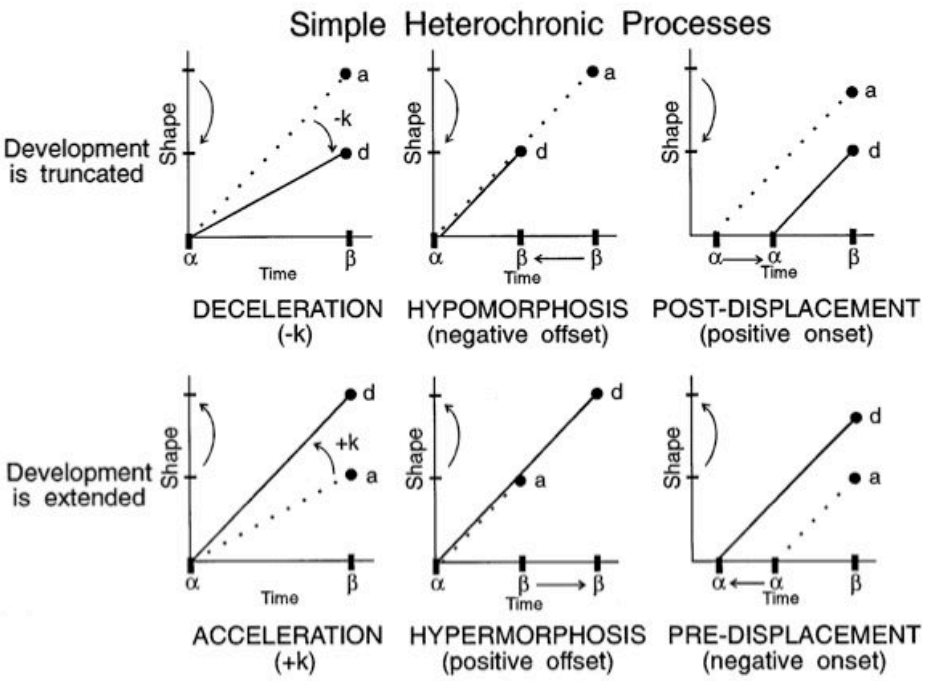


Figure 1. Six simple heterochronic processes identified by comparing ontogenetic trajectories of ancestral (a) versus descendant (d) ontogenies. Ontogenetic trajectories are defined by rate of shape development (k) from age of onset of growth (α) to the age when the offset shape is attained (β). Arrows on the shape axis indicate patterns of truncated (top) or extended (bottom) development. The terms deceleration and hypomorphosis are formally proposed to replace the inappropriate terms neoteny and progenesis, respectively, used by Alberch *et al.* (1979). Although originally defined for comparing species (Alberch *et al.*, 1979) this scheme can be used to categorize both inter- and intraspecific heterochronic phenomena.

http://www.usm.maine.edu/bio/courses/bio205/Lab_3.html

Differences between plants and animals in development

Differences in plant development, as compared to animals:

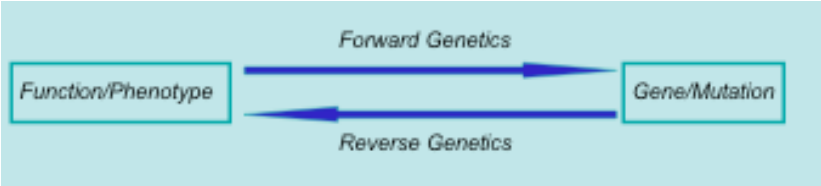
Modular growth, at several hierarchical levels
Growth from an apical meristem (or single apical cell)
Cells don't move (rigid cell wall)
Plants do not have a segregated germ line

Ontogeny and genetics

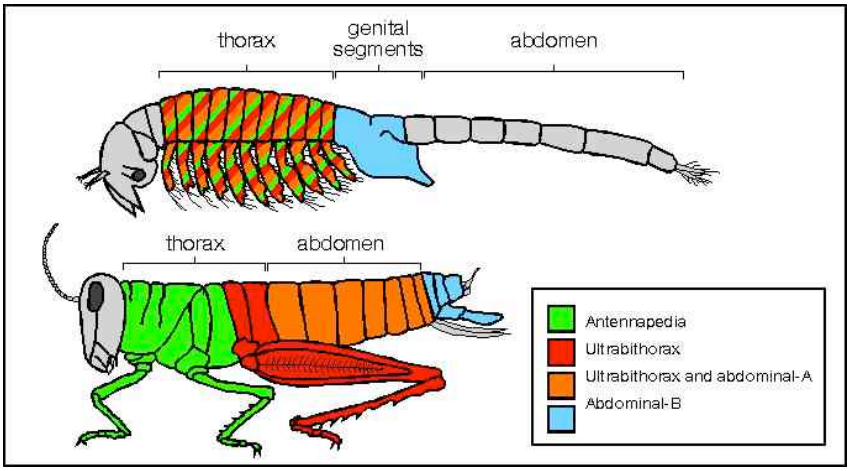
1) Expression studies
-- use of reporter genes
-- EST studies (cDNAs from target tissues)

2) Forward genetics
-- starts with a phenotype and moves towards the gene
-- screen for & isolate relevant mutants
-- map locus through genetic crosses
-- isolate gene & sequence

3) Reverse genetics
-- Starts with a particular gene and assays the effect of its disruption
-- Knockouts of candidate genes by transformation, observe change in phenotypes



4) Gene family evolution
A. Hox genes in animals

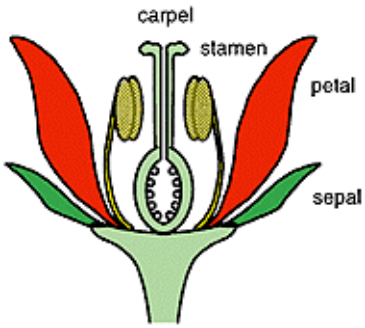


Hox genes are a subset of homeobox genes. Might have arisen by rounds of duplication of an ancestral gene, followed by a quaduplication of the cluster in mammals. Partially

overlapping zones of expression which vary in the anterior extent of their expression define distinct regions. Tandem gene duplication can allow retention of gene while new functions are adopted by one copy. Hox gene cluster arose from rounds of tandem duplication. Vertebrates have four Hoxgene complexes. *Amphioxus*, a vertebrate-like chordate, has one Hox cluster which may be close to ancestral Hox complex. (taken from http://www.mun.ca/biology/desmid/brian/BIOL3530/DB_Ch15/BIOL2900_EvoDevo.html)

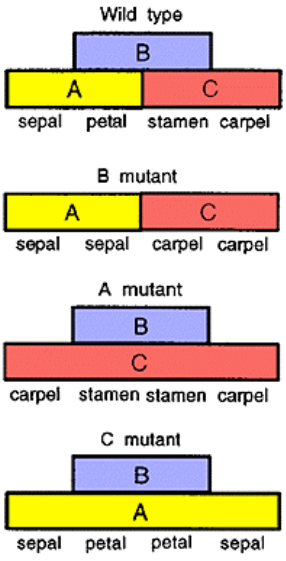
B. The ABC model in flowering plants

The MADS box is a highly conserved sequence motif found in a family of transcription factors. By now, more than hundred MADS box sequences have been found in species from all eukaryotic kingdoms. The family of MADS domain proteins has been subdivided into several distinct subfamilies. Most MADS domain factors play important roles in developmental processes. Most prominently, the MADS box genes in flowering plants are the "molecular architects" of flower morphogenesis (source: The MADS-box Gene Home Page; http://www.mpizkoeln.mpg.de/mads/).



MADS-box genes and the ABC model of organ identity determination

The basic structure of a complete flower consists of four concentric whorls. A simple model has been proposed to predict organ formation in flowers, where three classes of homeotic genes, the so-called ABC-class genes, act alone or together to give rise to sepals (A), petals (A+B), stamens (B+C), and carpels (C).



According to the ABC-model, organ determination in the whorls depends on the combinatorial action of three regulatory functions. A mutation disrupting one of the functions causes a homeotic change in organ identity. Note that the A and C functions are negatively regulating each other: Mutation in one causes expansion in the expression domain of the other. Molecular cloning has indicated that most of these ABC homeotic genes encode a well conserved DNA binding domain, the MADS box, and that this domain has been shown to be capable of binding to specific DNA sequence motifs known as CArG boxes. Because of their essential roles in flower development, and due to the high degree of conservation in the MADS box domain, MADS box genes have been cloned from diverse angiosperm plant species, including petunia, tomato, maize, white campion, sorrel, gerbera, and even one gymnosperm species, spruce. Although the ABC model has been shown to apply in several species other than the model species *Arabidopsis* and *Antirhhinum*, the precise functions of most MADS box genes remain unclear. It seems that, in addition to their essential roles during floral development, MADS box genes act also as regulators for various other aspects of plant development.

Source Gerbera Lab, Univ Helsinki http://honeybee.helsinki.fi/MMSBL/Gerberalab/abc.html