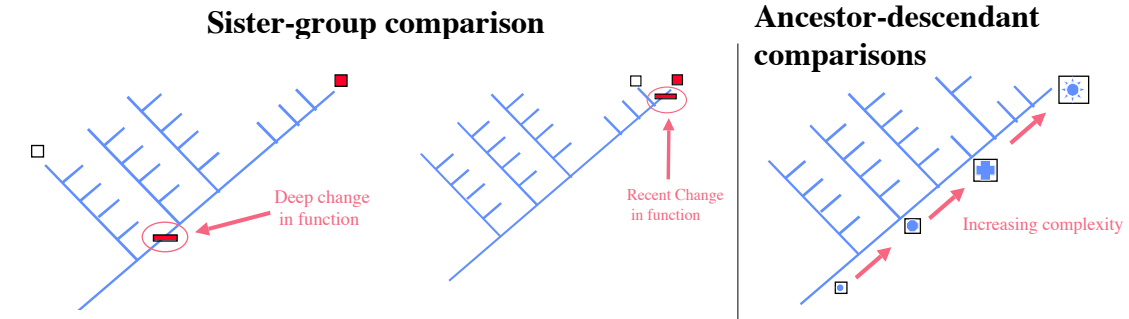
April 13, 2018. **Comparative genomics; evo-devo**

A. Phylogenomics

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.



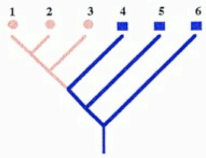
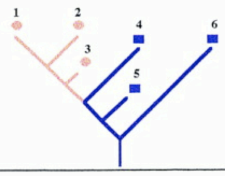
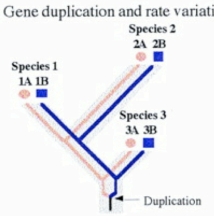
Most importantly for the systematist, the new comparative genomic data should also greatly increase the accuracy of reconstructions of the Tree of Life. Even though nucleotide sequence comparisons have become the workhorse of phylogenetic analysis at all levels, there are clearly phylogenetic problems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

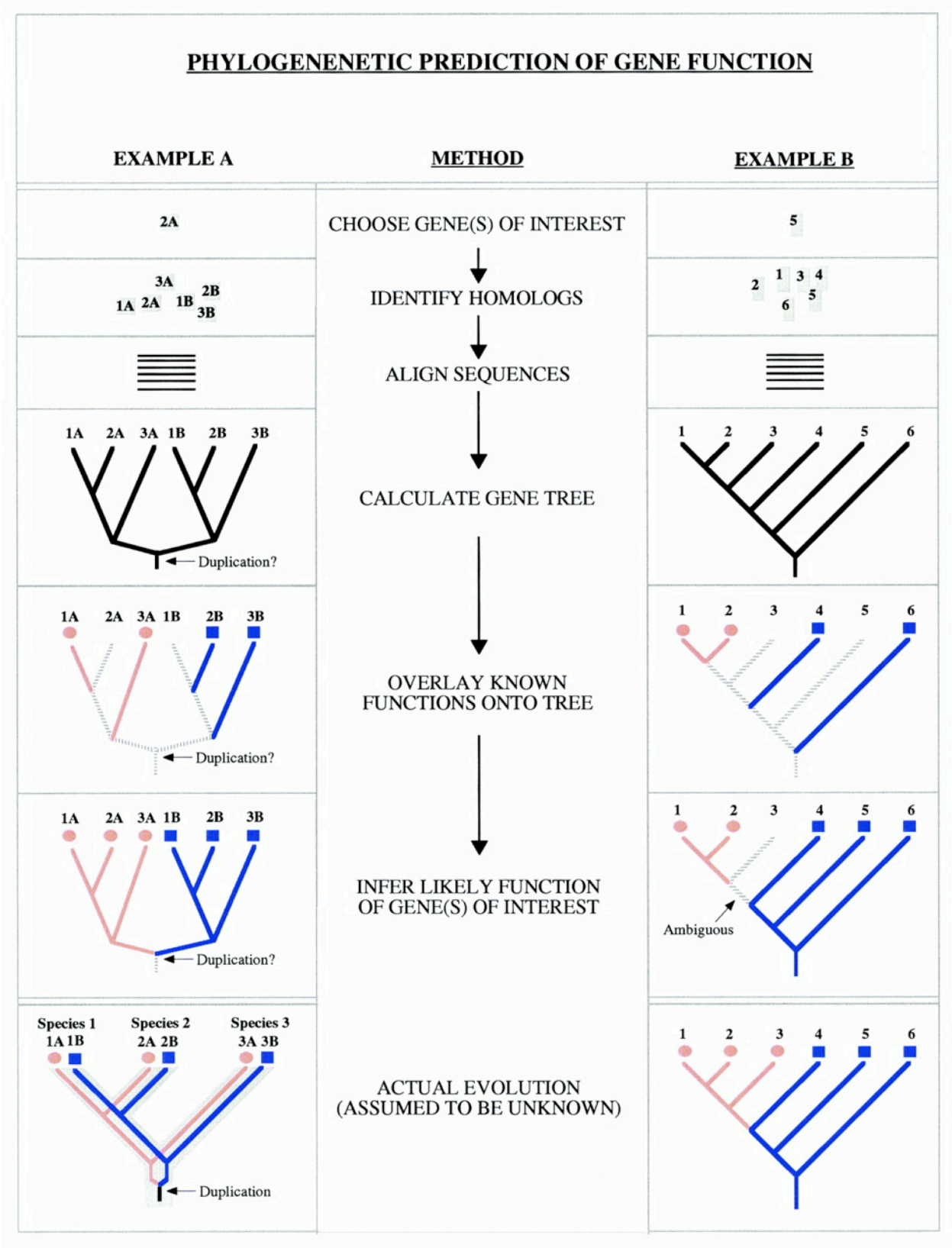
Table 4 Examples of Conditions in Which Similarity Methods Produce Inaccurate Predictions of Function

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. 	1 A ●	●	+	●	+	• 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.
	2 A ●	●	+	●	+	
	3 A ●	■	-	●	+	
	1 B ■	■	+	■	+	
	2 B ■	■	+	■	+	
	3 B ■	●	-	■	+	

[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. 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.



B. 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.

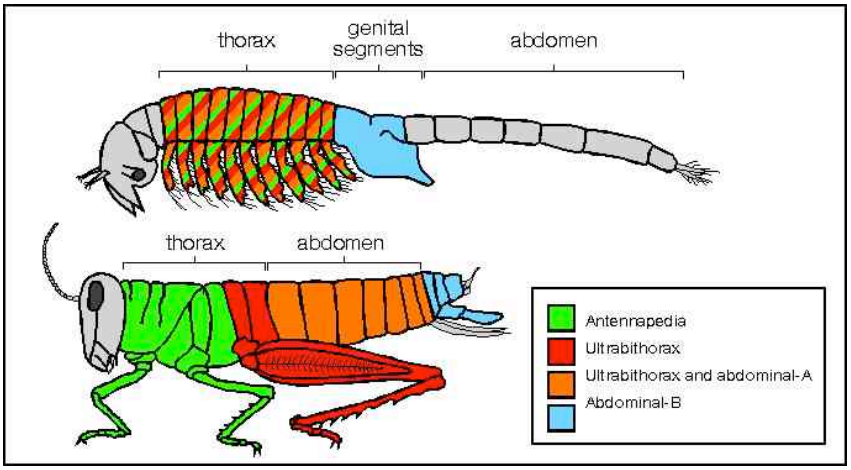
Ontogeny and genetics

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

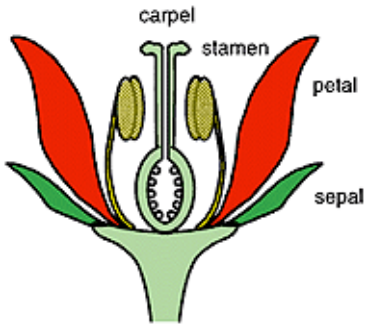
distinct regions. Tandem gene duplication can allow retention of gene while new functions are adopted by one copy. Hox gene cluster arose from rounds of tandem duplication. Vertebrates have four Hoxgene complexes. *Amphioxus*, a vertebrate-like chordate, has one Hox cluster which may be close to ancestral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

B. The ABC model in (some) flowering plants

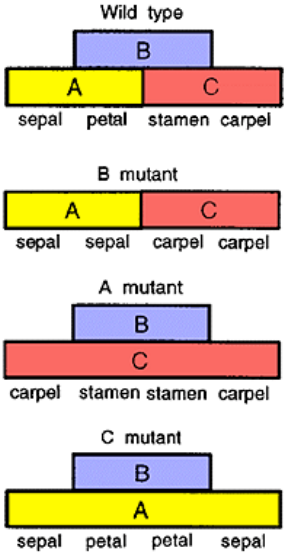
The MADS box is a highly conserved sequence motif found in a family of transcription factors. By now, more than hundred MADS box sequences have been found in species from all eukaryotic kingdoms. The family of MADS domain proteins has been subdivided into several distinct subfamilies. Most MADS domain 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

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



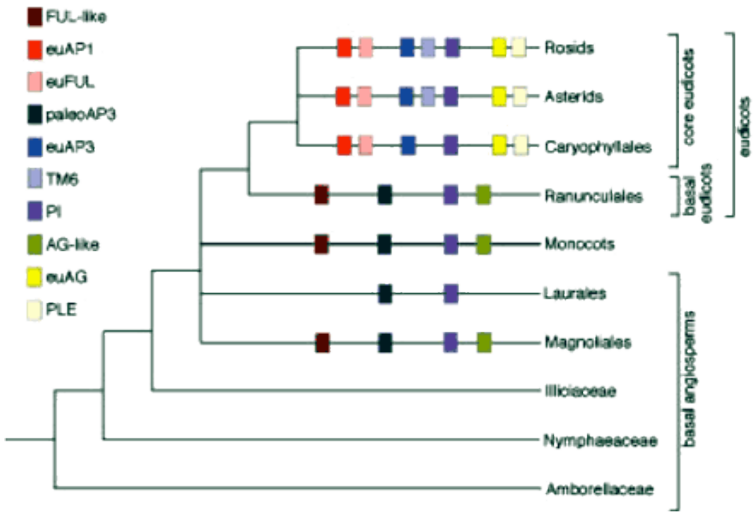
The basic structure of a complete flower consists of four concentric whorls. A simple model has been proposed to predict organ formation in flowers, where three classes of homeotic genes, the so-called ABC-class genes, act alone or together to give rise to sepals (A), petals (A+B), stamens (B+C), and 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. For general reviews see: Kramer & Hall, 2005. [Evolutionary dynamics of genes controlling floral development, *Current Opinion in Plant*

Biology 8: 13-18], and Heijmans, Morel, & Vandenbussche, 2012. [MADS-box genes and floral development: the dark side, *Journal of Experimental Botany* 63:5397–5404].

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; homologs are found in most Eukaryotes! This is a good example of repurposing of genes, probably though duplication and subfuntionalization.



MADS box gene duplications. Genes corresponding to the different MADS box gene lineages are indicated in the clades where they have been identified.
Source: Irish VF (2003) The evolution of floral homeotic gene function. BioEssays 25:637?646.

5) Biochemical pathways

These are great examples of the relationship between ontogeny and phylogeny. By far the easiest way for chemical diversity to evolve is through change in existing biosynthetic pathways. An example from Kip Will's current research:

Bombardier beetles. Geadephaga is the largest clade of organisms that use a single homologous gland system to produce no less than 19 distinct classes of chemical compounds for defense. Four lineages of quinone producing carabid beetles, including four species commonly known as the bombardier beetles, chemically blast their defensive quinones at extremely hot temperatures (up to 100 °C). Transcriptomes for genes involved in quinone production can be examined comparatively to elucidate chemical biosynthetic pathways, and describe the genetic architecture of quinone evolution. The evolutionary history of quinone biosynthesis in carabids can be studied by inferring the phylogenetic history of candidate gene families using the tree topology and branch lengths to test whether genes are ancient and shared among taxa, or if gene diversification is recent and specific to certain lineages. The hypothesis is that the genes up-regulated in secretory cells during quinone synthesis are closely related to those involved in quinone production in the arthropod cuticle.