

Gene trees and “species” trees: problems for comparative genomics and classification

Brent D. Mishler

University Herbarium, Jepson Herbarium, and Department of Integrative Biology
1001 Valley Life Sciences Bldg # 2465
University of California, Berkeley
Berkeley, CA 94720-2465 U.S.A.
E-mail: bmishler@calmail.berkeley.edu Phone: 1-510-642-6810

Cross-genome comparisons using phylogenetic trees are beginning to provide insights into many important functional questions, including understanding the processes underlying genomic evolution, gene regulation, the complex relationship between phenotype and genomic changes, and the evolution of complex physiological mechanisms. Before these comparisons can proceed, however, one has to be able to get the phylogenetic trees right. Several problems exist, most of which are receiving careful attention these days, yet one serious problem remains underappreciated: there is no single phylogenetic tree, instead there is a fractal pattern of lineages nested inside of each other, not always matching each other in branching pattern from level to level. There are four major categories of reasons why two genes in the same genome could be incongruent with each other, or with a higher-order population lineage, and it is important to keep the implications of each distinct:

- reticulation (which has several subcategories itself)
- lineage sorting (i.e., uneven extinction of ancestral polymorphisms)
- convergence due to natural selection
- random effects (i.e., differential evolutionary rates or biases) causing spurious branch attraction

We are faced at the moment with the opportunity to annotate a newly-available genome, *Physcomitrella*. We can look afresh at questions such as: What does it mean to "identify" a gene, place it into a classification, and make inferences about its function and history? These are all phylogenetic issues, and a number of concepts first developed in the field of systematics are highly relevant. Getting clear about concepts of basic taxa ("species?"), homology (including the difficult relationship between orthology and paralogy), functional relationships, and higher-level phylogenetic categories is the key to good gene classifications. We need to add a gene ontology category reflecting history! This would not be to the exclusion of functional ontology categories, but rather an addition to them. We would thus be able to look at function and history in light of each other, i.e., the *evolution* of function. We need a unique identifier for each individual gene from each genome, plus a clearly applicable name for each gene clade (distinct from the associated taxon name!!), registered in a data base (GO associated?). Some insights into how to do this can be gained from the PhyloCode (<http://www.ohiou.edu/phylocode/>), which is currently being developed for phylogenetic taxonomy of organism lineages (an interesting and unanticipated convergence of ideas!).