Syst. Biol. 46(3):523–536, 1997

GENE TREES IN SPECIES TREES

WAYNE P. MADDISON

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA

Abstract.—Exploration of the relationship between gene trees and their containing species trees leads to consideration of how to reconstruct species trees from gene trees and of the concept of phylogeny as a cloud of gene histories. When gene copies are sampled from various species, the gene tree relating these copies might disagree with the species phylogeny. This discord can arise from horizontal transfer (including hybridization), lineage sorting, and gene duplication and extinction. Lineage sorting could also be called *deep coalescence*, the failure of ancestral copies to coalesce (looking backwards in time) into a common ancestral copy until deeper than previous speciation events. These events depend on various factors; for instance, deep coalescence is more likely if the branches of the species tree are short (in generations) and wide (in population size). A similar dependence on process is found in historical biogeography and host–parasite relationships. Each of the processes of discord could yield a different parsimony criterion for reconstructing the species tree from a set of gene trees: with horizontal transfer, choose the species tree that minimizes the number of transfer events; with deep coalescence, choose the tree minimizing the number of extra gene lineages that had to coexist along species lineages; with gene duplication, choose the tree minimizing duplication and/or extinction events. Maximum likelihood methods for reconstructing the species tree are also possible because coalescence theory provides the probability that a particular gene tree would occur given a species tree (with branch lengths and widths specified). In considering these issues, one is provoked to reconsider precisely what is phylogeny. Perhaps it is misleading to view some gene trees as agreeing and other gene trees as disagreeing with the species tree; rather, all of the gene trees are part of the species tree, which can be visualized like a fuzzy statistical distribution, a cloud of gene histories. Alternatively, phylogeny might be (and has been) viewed not as a history of what happened, genetically, but as a history of what could have happened, i.e., a history of changes in the probabilities of interbreeding. [Biogeography; coalescence; coevolution; evolution; gene duplication; gene genealogy; gene trees; horizontal transfer; hybridization; lineage sorting; parsimony; phylogeny; species concepts; species trees; tree reconciliation.]

A phylogenetic tree of species contains smaller trees descending within its branches: the trees of genes. Recently, the relationship between gene trees and species trees has been the focus of some attention (e.g., Fitch, 1970; Goodman et al., 1979; Avise et al., 1983; Tajima, 1983; Pamilo and Nei, 1988; Takahata, 1989; Roth, 1991; Wu, 1991; Doyle, 1992; Hudson, 1992; Page, 1993; Baum and Shaw, 1995; Maddison, 1995, 1996). One aspect of this relationship is the congruence between the species tree and a tree of gene copies sampled from those species. Imagine that one gene copy was sampled from each species, and the gene tree relating these gene copies is examined. One might expect that two sister species would have sister copies in the gene tree and that other aspects of the gene tree would be congruent with the species tree, but this need not be the case (Fitch, 1970; Avise et al., 1983; Tajima, 1983; Pamilo and Nei, 1988; Doyle, 1992). In this article, I review the processes by which discord can arise and then explore how a species tree can be reconstructed from gene trees by considering these processes of discord. However, discordant gene trees will also provoke me to reconsider precisely what species trees (i.e., phylogenies) are.

GENE TREES AND SPECIES TREES

Genes have gene trees because of gene replication. As a gene copy at a locus in the genome replicates and its copies are passed on to more than one offspring, branching points are generated in the gene tree. Because the gene copy has a single ancestral copy, barring recombination, the resulting history is a branching tree. (Point mutation can cause some of the copies to be imperfect representations of the original, but this process does not compromise the existence of the tree.) Sexual reproduc-

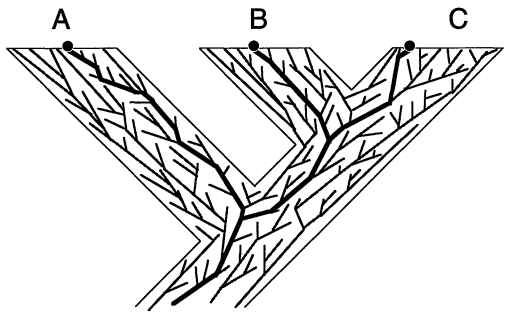

FIGURE 1. A gene tree contained within a species tree leading to three extant species: A, B, and C. Bold branches of gene tree show relationships among the sampled copies of the gene (●). Sampled copies from sister species B and C are sister copies.

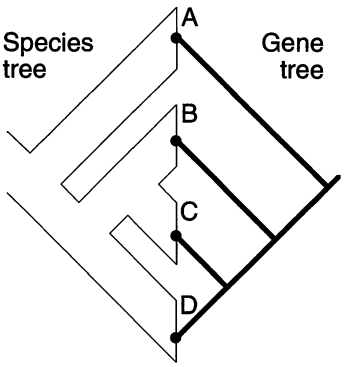

FIGURE 2. Discord between gene and species trees. At left is the species tree of four species, A, B, C, and D, and at right is the tree of a gene sampled one copy per species. Species B and C are sister species, but their gene copies are not sister copies.

tion and recombination within populations may appear to but actually do not cause genetic history to be reticulating. Rather, these processes break up the genomic history into many small pieces, each of which has a strictly treelike pattern of descent (Hudson, 1983; Hein, 1990; Maddison, 1995). Thus, within a species, many tangled gene trees can be found, one for each nonrecombined locus in the genome.

A phylogenetic (species) tree might be defined as the pattern of branching of species lineages via the process of speciation. When reproductive communities are split by speciation, the gene copies within these communities likewise are split into separate bundles of descent. Within each bundle, the gene trees continue branching and descending through time. Thus, the gene trees are contained within the branches of the species phylogeny (Fig. 1) (note, however, that this description rests upon certain concepts of phylogeny and species that I challenge, or at least reconsider, here).

Gene trees within species trees, therefore, are analogous to species trees within area cladograms in biogeography or to parasite trees within host trees in coevolutionary studies (Page, 1988, 1993, 1994a; Doyle, 1992; Maddison, 1996). In each case a *containing* tree descends and branches, while within its branches a *contained* tree itself descends and branches. The processes involved in the descent and containment of the contained tree are expected to be different in each case, however. To make historical interpretations about contained and containing trees, we must take a close look at the processes that determine the relationship between trees of these two types.

THE PROCESSES OF DISCORD

A gene tree can disagree in form with its containing species tree. Let us return to our imaginary example, in which a single gene copy was sampled from each of several species. If we knew the true species tree and the true gene tree relating those gene copies, we might see that sister gene copies are not in sister species (Fig. 2). (I assume through most of this discussion that the true gene trees are known without error. Of course, there will be errors in practice, and these errors will mean that reconstructed gene trees and species trees will have additional sources of discord.)

In the simple example of Figure 2 with one gene copy sampled per species, it was easy to define agreement between the trees; the gene copies must show precisely the same branching topology as their containing species. Figure 1 shows an example of agreement. The gene copy highlighted with a spot in each of the three species is the one sampled. As can be seen from the gene tree, the two sister gene copies fall within the two sister species (B and C), so there is agreement between the species

tree and the tree relating these three gene copies. But if we were to sample more gene copies from each species or if we were to imagine all of the extant gene copies in each of the species, then the gene tree has many more terminal taxa than the species tree, and therefore the branching topologies cannot possibly match. A modification of the definition of agreement is therefore necessary: for the gene and species trees to agree, the sets of gene copies from each species and from each monophyletic group of species on the species tree must form respective monophyletic groups on the gene tree. When all gene copies are considered, the gene tree of Figure 1 disagrees with the species tree, in that the gene copies from the monophyletic species group of B + C are not monophyletic on the gene tree. This example shows that whether or not a gene tree agrees with the species tree may depend on what gene copies have been sampled and included in the gene tree.

Here, I generally refer to discord between a species tree and a tree of sampled gene copies, one from each species, rather than the full tree of all extant gene copies. However, my conclusions do not depend on this restriction (although the simplicity of the explanation does).

Horizontal Transfer

In the example of Figure 2, what could be the cause of such discord between the gene tree and the species tree? One possible cause would be that renegade genes have somehow broken the confines of the species lineages and moved horizontally across the phylogeny (Fig. 3). Horizontal transfer might be accomplished by a vector such as a virus or mite (Kidwell, 1993; Cummings, 1994). Isolated hybridization events across the phylogeny can have a similar genetic effect (Doyle, 1992) and so might be considered examples of horizontal transfer. (However, one could argue that such hybridizations do not generate discord between species trees and gene trees, but rather they indicate that the species tree was more complex than originally thought.)

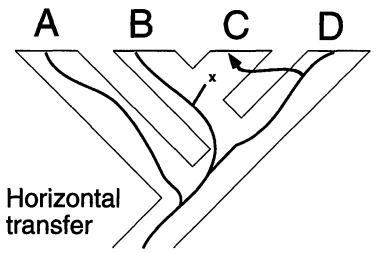

FIGURE 3. Horizontal transfer. A branch of the gene tree jumps between species lineages. If the indigenous gene copy in the receiving species lineage goes extinct or is not sampled (✖), then the gene tree will disagree with the species tree, as shown in Figure 2.

How likely is horizontal transfer? Successful transfer by means other than hybridization requires not only a vector or other means of transfer but also incorporation of the transferred genes to become functioning members of the receiving genome (Cummings, 1994). Generally, it might be expected that successful transfer would be less likely the more phylogenetically distant the original and receiving species. The same would be expected of transfer by hybridization.

Lineage Sorting or Deep Coalescence

Genes do not have to cross species boundaries for their trees to disagree with the containing species tree. It has been realized for some years that when ancestral polymorphisms persist through several speciation events, the subsequent loss or failure to sample some of the gene forms in the various species can give a gene tree with a topology different from that of the species tree (i.e., lineage sorting; Avise et al., 1983; Tajima, 1983; Takahata and Nei, 1985; Neigel and Avise, 1986; Nei, 1987). For instance, if the dashed and solid gene copies existed in the ancestral species shown in Figure 4 and neither were lost from the population by the time of the speciation event marked by an asterisk, then by chance only the dashed gene form might survive and be sampled in species B and only the solid form might be found in species C. Because the solid form was sampled from species D, the resulting gene tree would show the gene copies C and *D*

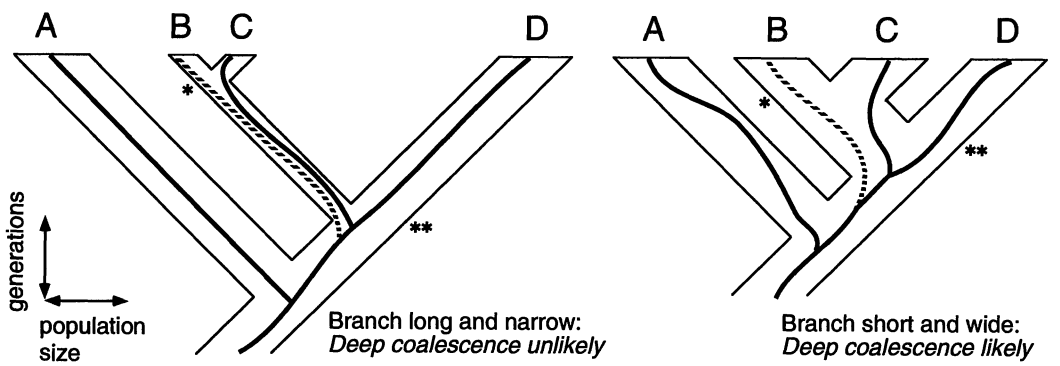

FIGURE 4. Lineage sorting (deep coalescence). Described in a time-forward sense as lineage sorting, an ancestral polymorphism at ** is retained through a lineage to the next speciation event at *, where different forms are sampled in different descendant species. Described in a time-backward sense as deep coalescence, two gene copies from species B and C meet at * but fail to coalesce until deeper than the speciation event at **, at which point the gene from C coalesces first with the gene from D. Failure to coalesce is more likely the shorter (in generations) and wider (in effective population size) the branch is between ** and *.

(from species C and D, respectively) as most closely related. This process can occur, of course, whether or not the various gene copies differ in nucleotide sequence, although without differences the process would be undetectable.

It may be easier to visualize and describe this process as if it operated backwards in time, in the tradition of coalescence theory in population genetics (Kingman, 1982; Hudson, 1990). One advantage of a coalescence perspective is that its implicit focus on a sample of gene copies obviates the need to equivocate about loss versus failure to sample. The source of discord between gene tree and species tree would then be viewed as a problem of *deep coalescence* instead of lineage sorting; i.e., common ancestry of gene copies at a single locus extends deeper than speciation events. Going back in time, the ancestors of the sampled gene copies *B* and *C* in Figure 4 find themselves in the same ancestral species at the point marked by the single asterisk. Chances are that the ancestral copies of *B* and *C* will not share a common ancestral gene copy in the first generation in which they find themselves together in the ancestral species (i.e., in a time-forward sense, in the generation immediately before the speciation event). In fact, if the population is large, the ancestors of these gene copies may take many generations before they happen to find each other and "coalesce" into a common ancestral copy (Tajima, 1983; Hudson, 1990). If by chance they have not yet found their common ancestor by the previous speciation event, marked by two asterisks, then suddenly they find themselves sharing the gene pool with the ancestral copy of *D*. At that point, before *B* and *C* coalesce, one of them might first coalesce with *D*. This would generate discord between the species and gene trees, for the gene copies from sister species B and C would not be sister copies.

The larger the effective population size and the shorter the phylogenetic branch, the greater the chances are that the ancestral copies will fail to coalesce before reaching the deeper speciation event (Pamilo and Nei, 1988). Thus, looking at Figure 4, one can say that the probability of deep coalescence generating discord is greater as the width of the branches (measured as effective population size) approaches the length of the branches (measured in numbers of generations). Long narrow trees are nearly immune to deep coalescence (barring balancing selection and other such processes), whereas short wide trees may show many genes with deep coalescence "problems."

Gene Duplication/Extinction

Like deep coalescence, the process of gene duplication generates multiple gene

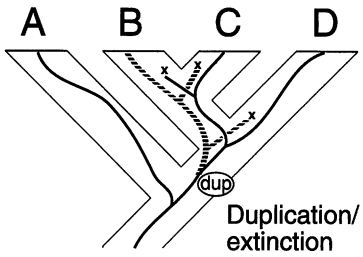

FIGURE 5. Gene duplication and extinction (or paralogous sampling). The gene is duplicated to a different locus, indicated by the dashed lines. If in descendant species one or the other locus goes extinct or is not sampled (✖), then the gene tree will disagree with the species tree, as shown in Figure 2.

lineages coexisting in a species lineage (Page, 1993), and likewise it can result in gene tree–species tree discord (Fitch, 1970; Goodman et al., 1979). When a gene duplication event yields a second locus, the first and second gene loci will have their gene copies evolving and descending independently of each other. In Figure 5, a gene duplication yields a new locus, shown by dashed lines. If some of the surviving or sampled copies in the extant species come from the dashed locus and others come from the solid locus (i.e., they are paralogous instead of orthologous; Fitch, 1970), then the tree of genes can disagree with the tree of species (Goodman et al., 1979).

Unlike deep coalescence, gene duplication and paralogous sampling do not depend in a simple way on population sizes. With deep coalescence, the different gene copies are competing for the same locus in the genome, and the probability that two copies will find themselves sitting on the same chair each time the music stops will depend in a fairly simple way on the number of chairs, i.e., the effective population size. However, with gene duplication, two gene copies at different loci in the genome are not competing for the same site. One of the copies could go extinct (i.e., become unrecoverable) if gene conversion occurred, or if it decayed into a pseudogene, or if it evolved a new function and diverged. The last outcome is more likely in large populations (Walsh, 1995), but in general the rules governing the fate of duplicated gene copies (Walsh, 1987, 1995) are rather different from those governing neutral alleles sharing a locus.

Reconstructing a Species Tree Using Gene Trees

Because gene trees can disagree with their containing species tree, a research program that sequences copies of a single gene from various species to reconstruct the species tree can yield an erroneous species tree even if the gene tree is reconstructed correctly (Goodman et al., 1979; Tajima, 1983; Pamilo and Nei, 1988; Takahata, 1989; Roth, 1991; Wu, 1991; Doyle, 1992). If there is indication that population sizes have been small relative to the length of phylogenetic branches (as might be the case, for instance, with higher level phylogenies), then a gene tree might be a reasonably faithful indicator of species trees. However, near the species level it may be necessary to combine data from many gene copies or multiple genes to arrive at a good estimate of the species tree (Pamilo and Nei, 1988; Takahata, 1989; Wu, 1991; Doyle, 1992). Takahata (1989) showed that if many gene copies are sampled from each species, sister species will show enough of the shallowest interspecific coalescences to allow the species tree to be reconstructed correctly. Wu (1991) and Doyle (1992) both discussed the use of multiple genes, which I consider here.

Assume that there are four species A, B, C, and D and 10 unlinked genes. A single copy of each gene is sampled and sequenced from each species. For each gene, its gene tree of four copies (one from each species) is reconstructed correctly. For three of the genes the gene tree has the form shown in Figure 6a, for three of the genes the gene tree is as shown in Figure 6b, and for four of the genes the gene tree is as shown in Figure 6c. Can these 10 gene trees be used to reconstruct the species tree?

The simplest procedure might be to choose the commonest gene tree found, that in Figure 6c. This may be a reasonable approach in this simple example, but in

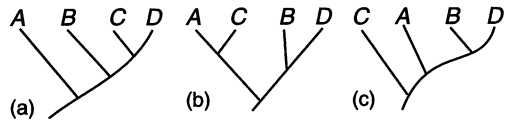

FIGURE 6. Hypothetical example of gene trees used to construct a species tree. Ten gene loci have been sampled, each represented by one copy from each of four species, A, B, C, and D. In three of the genes, the copies from the species are related as shown in tree a, in three of the genes they are related as shown in tree b, and in four of the genes they are related as shown in tree c.

other examples it could fail to reflect the overall support for another tree if a series of less common trees were nearly in agreement and together were much more numerous than the modal tree.

Another approach would be to use a sort of parsimony procedure, which would assess the various possible species trees and for each tree ask what evolutionary events the species tree requires to explain the observed gene trees. I have already outlined three different classes of evolutionary processes by which discordance between gene trees and species trees arise: horizontal transfer, deep coalescence, and gene duplication. I now consider possible parsimony procedures, assuming each process in turn.

Horizontal Transfer

If the only process by which the disagreement among the 10 gene trees (Fig. 6) arose were horizontal transfer, how would we assess the parsimony of a species tree? A simple measure would be to count the minimal number of transfer events needed by the species tree to explain all the gene trees. We therefore need to be able to take a species tree and gene tree and ask how many transfer events are needed to map the gene tree onto the species tree (much as we ask about "steps" for mapping character states onto a tree).

Obtaining this number may not be easy; the algorithms to find quickly the minimal number of transfers have not yet been developed. There are some calculations for fitting one tree into another that seem relevant: Brooks parsimony analysis (BPA; Brooks, 1981, 1990), nearest neighbor interchange metric (NNI; Robinson, 1971; Day, 1983), and the number of branches pruned to yield the greatest agreement subtree (Finden and Gordon, 1985). Doyle (1992) suggested the use of BPA to recode gene trees as character state trees to reconstruct the species tree. BPA uses algorithms designed for mapping character state trees onto species trees and therefore would be analogizing character state changes with transfer events. However, given that the algorithms were not designed to count transfers, it is not surprising that they do not (e.g., their cost for a gene transfer depends on the distance in the gene tree between the entering and resident genes). The NNI metric counts how many branch moves are needed to convert one tree into another, but horizontal transfer in nature is not necessarily restricted to a series of nearest neighbor events. The number of pruned branches from the agreement subtree likewise does not simply count transfer events. What is needed is a method that counts the minimal number of branch moves needed to convert one tree into another, where branch moves are restricted so as not to violate a linear time order (one can imagine a series of branch moves that cannot possibly happen together, e.g., one move from branch A to branch B and then another move from a descendant of B to an ancestor of A). Page (1994b) has made progress in exhaustively specifying alternative events to fit one tree into another, but transfer events are not yet counted separately.

Lacking an exact algorithm, I examined species trees and gene trees by eye and attempted to judge the minimal number of transfer events required of the gene tree by the species tree. The results for each of the 15 possible rooted species trees for the four species A, B, C, and D and for each of the three classes of gene tree are shown in Table 1. The number of transfer events was 0 (if the gene and species trees matched) or 1 or 2 (if not) for all examples. The only complication concerned mapping gene tree (C(*A*(*B*, *D*))) onto species trees (A(D(B, C))) and (A(B(C, D))). In those cases, two trans-

TABLE 1. Parsimony calculations to choose a species tree from observed gene trees. The 10 observed gene trees fall into three categories: three trees have the form (*A*(*B*(*C*, *D*))), three trees have the form ((*A*, *C*)(*B*, *D*)), and four trees have the form (*C*(*A*(*B*, *D*))). Values are the minimal numbers of evolutionary events required by a candidate species tree to explain the gene tree under three alternative models: horizontal transfer (HT), counting number of transfers; deep coalescence or lineage sorting (DC), counting number of extra lineages along species branches; and gene duplication and extinction (D/E), counting number of duplication and extinction events.

	Gene tree											
	(*A*(*B*(*C*,*D*)))			((*A*,*C*)(*B*,*D*))			(*C*(*A*(*B*,*D*)))			Totals		
Species tree	HT	DC	D/E	HT	DC	D/E	HT	DC	D/E	HT	DC	D/E
(A(B(C,D)))	0	0	0/0	1	2	1/4	1 + g[a]	2	1/4	7 + 4g	14	7/28[b]
(A(C(B,D)))	1	1	1/3	1	1	1/3	1	1	1/3	10	10[b]	10/30
(A(D(C,B)))	1	1	1/3	1	2	1/4	1 + g	2	1/4	10 + 4g	17	10/37
(B(A(C,D)))	1	1	1/2	1	2	1/4	2	3	2/7	14	21	14/46
(B(C(A,D)))	2	2	1/4	1	2	1/4	1	3	2/7	13	24	14/52
(B(D(C,A)))	2	2	1/4	1	1	1/3	2	3	2/7	17	21	14/49
(C(A(B,D)))	1	3	2/7	1	1	1/3	0	0	0/0	6[b]	12	9/30
(C(B(A,D)))	2	3	2/7	1	2	1/4	1	1	1/2	13	19	13/41
(C(D(B,A)))	2	3	2/7	1	2	1/4	1	1	1/2	13	19	13/41
(D(A(B,C)))	1	3	2/7	1	2	1/4	2	3	2/7	14	27	17/61
(D(B(A,C)))	2	3	2/7	1	1	1/3	2	3	2/7	17	24	17/58
(D(C(B,A)))	2	3	2/7	1	2	1/4	1	3	2/3	13	27	17/45
((A,B)(C,D))	1	1	1/3	2	2	1/4	1	2	2/6	13	17	14/45
((A,C)(B,D))	1	2	2/6	0	0	0/0	1	1	1/3	7	10[b]	10/30
((A,D)(C,B))	1	2	2/6	2	2	1/4	1	2	2/6	13	20	17/54

[a] g = "ghost" species lineage, which is unobserved and survived long enough to effect transfer.
[b] Preferred species tree(s) for each model.

fer events were required unless one was willing to imagine a now-extinct species lineage that split off early and survived just long enough to transfer its gene into species C when that species became distinct. Fortunately, it made no difference to the final results whether one or two events were counted in these cases.

Overall, the most-parsimonious species tree is the one that has the same form as the gene tree of Figure 6c, (*C*(*A*(*B*, *D*))). It requires six transfer events; other species trees require 7–17 transfer events.

Lineage Sorting (Deep Coalescence)

If the only process by which the disagreement among the 10 gene trees (Fig. 6) arose were deep coalescence, how would we assess the parsimony of a species tree? When a hypothesis of deep coalescence is invoked to explain gene tree disagreement, then this ad hoc hypotheses should be counted against the species tree, but how do we assess the severity of the deep coalescence required? For instance, for gene tree (*A*(*B*(*C*, *D*))), the species tree (A(C(B, D))) requires that two gene lineages fail to coalesce along one branch of the species tree (Fig. 7a). The species tree (C(B(A, D))), however, requires for the same gene tree that two gene lineages fail to coalesce along one branch and three gene lineages fail to coalesce along another (Fig. 7b). In a parsimony framework, one possible measure of the severity of deep coalescence is the number of "extra" gene lineages on species branches. Thus, the tree in Figure 7a has a branch with two gene lineages, i.e., one extra. The tree in Figure 7b has one branch with one and another branch with two extra gene lineages, for a total of three extra gene lineages.

This number of "extra" gene lineages along branches is relatively easy to count, once the gene tree has been fit onto the species tree using the same methods as used with gene duplication and extinction (Goodman et al., 1979: appendix A-2). To fit a node of the gene tree onto the species tree, first find all of the terminal species that contain sampled gene copies descended from that node and then find the most recent common ancestor of those species

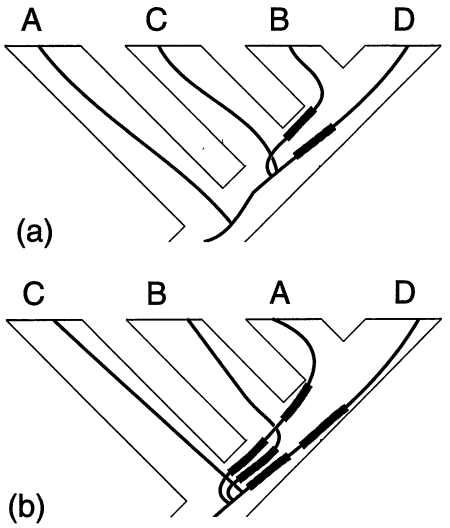

FIGURE 7. Assessing two different species trees for deep coalescence. Both species trees require lineage sorting (deep coalescence) to explain the gene tree (*A*(*B*(*C*, *D*))), where gene copy *A* was sampled from species A and so on. Bold lines mark coexisting gene lineages that fail to coalesce. (a) The species tree (A(C(B, D))) requires two gene lineages to coexist along a branch, i.e., there is one extra lineage. (b) Tree (C(B(A, D))) requires one branch with two extra lineages and another branch with one extra lineage, for a total of three extra gene lineages along branches.

on the species tree. The gene tree node can be placed at that ancestor in the species tree; it needs to have occurred at least that deep in the species tree but need not have occurred any deeper. Once the gene tree has been fit onto the species tree, visit all of the branches of the species tree and for each branch count the number of gene lineages minus one (to count "extra" lineages). The sum of extra gene lineages for each of the 15 possible species trees for the four species A, B, C, and D and for each of the three classes of gene tree are shown in Table 1. The number of extra gene lineages for a gene and species tree is 0–3.

There are two most-parsimonious species trees, both requiring 10 extra gene lineages counted over all of the gene trees. One has the same form as the gene tree of Figure 6b ((A, C)(B, D)), and the other is the tree (A(C(B, D))). Other trees require 12–27 extra gene lineages on branches of the species tree.

Duplication/Extinction

How would we assess the parsimony of a species tree if the only process by which the disagreement among our 10 gene trees arose were gene duplication and extinction (paralogous sampling)? There has already been considerable work on this question. Goodman et al. (1979) developed an algorithm that counts duplication and extinction (failure to sample) events when a gene tree is fitted onto (i.e., reconciled with) a species tree. Page (1988, 1994a) has used these methods also in biogeography and coevolution.

I counted the minimal number of duplication and extinction events for the 15 species trees and three classes of gene trees (Table 1). The species tree requiring the fewest duplication events over all of the gene trees is (A(B(C, D))), which has the same form as the gene tree shown in Figure 6a. This species tree requires 7 duplication events, whereas other species trees require 9–17 duplication events. Counting extinction events gives the same chosen tree, with 28 extinctions required for (A(B(C, D))) and 30–61 required for the other species trees.

Which Process(es) to Invoke?

For these examples, I constructed the data, the 10 gene trees, specifically to illustrate that different assumed processes can lead to different species trees as most-parsimonious explanations of the set of gene trees. Is it possible to construct a mixed method that does not assume only one of the processes is occurring but rather allows each to occur? For instance, a gene tree may be mapped onto a species tree by invoking a bit of deep coalescence, a gene duplication here and there, and a horizontal transfer event. This idea is certainly plausible, but it immediately brings up two difficulties. The first is the algorithmic difficulty of assessing the multitude of possible scenarios that could be used to fit any given gene tree onto a species tree. The second, and more important, is the difficulty of weighting these different events. Is a horizontal transfer worth one, two, or

three failures to coalesce through a branch? What is a gene duplication worth?

Each event of horizontal transfer, deep coalescence, and gene duplication depends upon different circumstances for its occurrence. It will often be possible to restrict which processes can be considered reasonable in different cases. At large phylogenetic scales, deep coalescence may be unlikely, and if vectors and other means of horizontal transfer are apparently unavailable, then gene duplication could be relied on exclusively as the source of gene tree discord. At small scales, near the species level, gene duplication (except with prolific duplicators like movable elements) may be unlikely, and either deep coalescence or a combination of deep coalescence and hybridization (transfer) could be assumed. Depending on the process assumed, the most-parsimonious species tree could be chosen by counting minimal numbers of transfers, extra gene lineages, or duplication and extinction events. (Of course, at any scale an apparent gene tree discord might simply be due to error in reconstructing one or more of the gene trees.)

This reasoning can and probably should be extended to the analogous cases of contained trees within containing trees, such as species trees in area trees (biogeography) and associate trees in host trees (coevolution). In judging a general area cladogram (host tree), should one count minimal number of dispersal events (host shifts), amount of sympatry (host sharing), or hidden speciation and extinction events? Which is appropriate would depend on what processes are expected to be most likely.

If multiple parasites occupying a host, or species occupying an area, coexist without competition, then they are expected to behave more like duplicated genes and might be treated similarly (Page, 1993). However, if there is competition among these coexisting lineages, then they are expected to behave more like alleles at a locus, and their sorting processes are expected to resemble more the process of lineage sorting.

Species Trees from Gene Trees via Maximum Likelihood

Having already begun down the slippery slope of considering evolutionary processes, I now consider for the case of deep coalescence how a species tree could be reconstructed using maximum likelihood techniques that rely upon a full probabilistic model relating gene trees and species trees. Wu (1991) and Hudson (1992) discussed likelihood and other statistical means for using multiple gene trees to reconstruct a species tree for the three-species case. Here I present a general discussion of the issues involved.

If the only process yielding gene tree–species tree discord were deep coalescence, then coalescence theory from population genetics could be used to help us reconstruct species trees. A candidate species tree whose likelihood we wish to consider has a parameter assigned to each branch. This parameter relates to the length versus width of the branch (e.g., the number of generations the branch has existed versus the harmonic mean of the effective population size along the length of the branch). Coalescence theory would then provide the probability that gene copies would coalesce in various ways within this species phylogeny (e.g., Pamilo and Nei, 1988; Takahata, 1989), so we could calculate the probability that a set of sampled gene copies from the extant species would coalesce to yield a particular gene tree. Gene trees with very deep coalescences would be less probable outcomes than those with shallower coalescences. For a set of "observed" gene trees, it would be straightforward to calculate the likelihood of a particular species tree by calculating the probability from coalescence theory of obtaining that set of gene trees from the proposed species phylogeny. By varying the length/width parameters of the branches of the species tree and examining all species trees, the species tree that confers highest probability on the observed gene trees could be found, i.e., the maximum likelihood species tree.

This procedure assumes that the gene

trees were reconstructed without error, but this assumption can be avoided. The likelihood calculations can be made directly from the gene sequences if a model of genetic changes is available. Thus, the likelihood of the species tree would be the probability of obtaining the observed sequences, and this probability would depend both on coalescence theory and on the model of nucleotide sequence evolution.

The likelihood of a given species tree would then be the product, over all loci, of the probability of obtaining the sequences observed at the locus given the species tree:

$$\prod_{\text{loci}} \sum_{\substack{\text{possible} \\ \text{gene trees}}} [P(\text{sequences}\,|\,\text{gene tree}) \cdot P(\text{gene tree}\,|\,\text{species tree})]$$

Because we are not assuming that we have the gene tree reconstructed, the calculations must consider for every locus each possible gene tree and its probability of occurrence given the species tree, P(gene tree|species tree), where the gene tree includes both topology and branch lengths. The probability comes straight from coalescence theory. For each gene tree, the probability of evolving the observed sequences, P(sequences|gene tree), comes from the model of nucleotide evolution.

To search for the maximum likelihood tree using such an approach would be extremely tedious, given that not only do we need to search over species trees but that for every species tree we have to consider all possible gene trees (including branch lengths). In addition, because P(sequences|gene tree) may require gene tree branch lengths independent of population size, the species tree may need to have two parameters (length and width) independently specified for each branch. However, J. Felsenstein (pers. comm.) has pointed out that the search might be made feasible by using approximate likelihoods: one can sample among possible gene trees in proportion to their coalescence probabilities (see Felsenstein, 1992, for a similar approach), thus avoiding an examination of all possible gene trees.

What Is a Species Phylogeny?

Of all the processes generating discord between gene trees and species trees, deep coalescence is perhaps the most problematic because it is expected to be ubiquitous in sexual species, a simple consequence of the fact that species lineages are not simple, indivisible lines but rather that each has a fine structure consisting of many organisms and their genes. Deep coalescence is such a natural outgrowth of our view of evolution that it should not be viewed as pathological in any way. I now expand on my suggestion (Maddison, 1995, 1996) that a full appreciation of this "problem" might lead to our revising the way we view phylogeny itself.

Suppose, to begin with, that a species phylogeny is meant to convey the broad-scale history of genetic descent. That is, it says that the genes of an ancestral species were passed down along the species lineage, but then the genetic connections were sundered into two main lines representing daughter species. Genetic descent continued in these daughter species, which may themselves speciate, and so on, to generate the phylogenetic tree. Thus, the tree is a broad-scale, low-resolution view of the genetic connections from one generation to the next. I occasionally refer to this as the realized genetic history: a summary of the history of the passage of all the genes through the generations.

Phylogeny as a Cloud of Gene Histories

The descent of all of the genes in the genome contributes to this broad history of genetic descent. How do we expect the histories of the individual genes to differ from one another? The simple example in Figure 8a shows a panmictic population descending through time, splitting once and then twice to yield the three populations (so as not to prejudge, I am not saying that this diagram represents the species tree—for now, it is merely a diagram of splitting populations). If the effective population sizes are represented by the branch widths and the durations in generations are represented by the branch

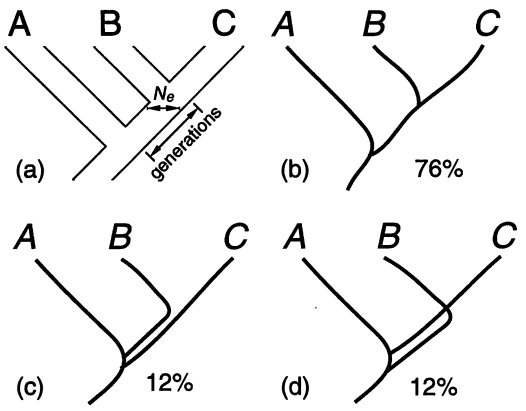

FIGURE 8. A tree of the successive splitting of a population (a) and three sampled gene trees (b–d). If the length (in generations) of the labeled branch is about twice the width of the branch (N_e = effective population size), then the gene trees should occur with the frequences shown.

lengths, then the intermediate branch had a duration about twice its population size. Using coalescence theory, it can be calculated (Pamilo and Nei, 1988) that the ancestors of two sampled gene copies *B* and *C* will coalesce with each other about 76% of the time before either coalesces with copy *A*. Thus, 76% of the genes will have trees like those shown in Figure 8b, 12% will have trees like Figure 8c, and 12% will have trees like Figure 8d. (Because there are many copies of the gene in each of the three species, the full gene trees will be considerably more complex and diverse. However, there will be a probability distribution for all the different possible gene trees that will place higher probabilities on those that tend to group the copies from species B and species C together.)

What is the species tree in this example? A usual interpretation is that the species phylogenetic tree is as in Figure 8a and that 76% of the gene trees agree with it but 24% disagree. But is genetic history a winner-take-all democracy? Those 24% of the genes are not losers that disagree with genetic history; they are part of the genetic history. If the species tree is to be an accurate summary of broad-scale genetic history, it would be far better to say that the species phylogeny is composed 76% of Figure 8b, 12% of Figure 8c, and 12% of Figure 8d.

When we take a sample from a population and try to understand a statistical distribution by calculating means and variances, we do not single out all of the samples whose values differ from the mean as disagreeing with the mean. They are simply part of the variance, part of the distribution. A simple phylogenetic tree diagram with sticklike branches represents only the mean or mode of a distribution. Phylogeny has a variance as well, represented by the diversity of trees of different genes.

This variance does not represent uncertainty due to ignorance or measurement error; it is an intrinsic part of phylogeny's nature. I have previously used an analogy from physics (Maddison, 1996). Although all of us have seen diagrams of an electron zooming around the nucleus of an atom like a discrete little satellite, physics now tells us that the electron is diffuse. It is not a matter of uncertainty about where the electron is. Rather, in a real sense the electron is in more than one place at once. Likewise, phylogenetic history is in more than one place at once; it is a composite of all the varied histories of all the genes, some of which might place species A next to B, others might place A next to C, etc. Just as an electron can be depicted as a cloud, we might want to view phylogeny as a diffuse cloud of gene histories (Fig. 9). To be sure, the cloud has some form, and in many cases it will have a central tendency that will take the form of a tree.

One might hope that an appropriate delimitation of species might somehow sweep these problems under the rug. If we could somehow delimit our species broadly enough, could we ensure that our interspecific phylogenies would have no fuzz about them? It seems unlikely that we could succeed in this endeavor. For instance, species delimitations would likely broaden considerably if a species concept using gene-coalescence exclusivity (Baum and Shaw, 1995) were applied in its strictest conceivable form, i.e., that each species has exclusive coalescence in all of its genes.

FIGURE 9. Phylogeny as a cloud of gene histories. Phylogeny is more like a statistical distribution than a simple tree of discrete thin branches. It has a central tendency, but it also has a variance because of the diversity of gene trees. Gene trees that disagree with the central tendency are not wrong; rather, they are part of the diffuse pattern that is the genetic history.

Thus, the species' gene copies would be monophyletic (with respect to the copies of other species) for every locus in the genome. This strict criterion would seem to guarantee phylogenetic cleanliness, but even if the terminal units of our phylogeny were broadened to satisfy it, the tree could still misbehave (be fuzzy and cloudlike) in its deeper areas because trees of different genes can still disagree as to whether two species are sister species or not (Maddison, 1995; see Maddison, 1996: fig. 5, for an example). More importantly, labeling the problem as "intraspecific" would not make it go away.

Phylogeny as a Model of Probabilities of Interbreeding

Some readers might take issue with the concept of phylogeny as the composite of gene histories. Surely a species phylogeny is more than that? In most of this discussion I have used a different model of a species phylogeny. Under a model of realized gene histories, the species tree is composed of gene trees (gene trees are to the species tree as parts are to the whole), but I have used a model of a species tree as a fragmenting container that stands apart from and constrains the descent of gene trees (gene trees are to the species tree, at least in a sense, as effect is to cause). When discussing deep coalescence (Fig. 4), I assumed that branches of the species tree represented populations, within which different gene copies could compete and coalesce and between which they could not. In many of the discussions that have incorporated coalescence theory in examinations of disagreements between gene trees and species trees, the branches of the phylogenetic trees are even more specifically defined, i.e., they represent panmictic populations that became isolated from other such populations when branching events occurred. The assumption of full panmixia within and complete isolation between populations is handy for the simplicity of the calculations, but of course these assumptions do not need to be so simple and strict. The details of the assumptions are irrelevant here; what is relevant is that the phylogenetic tree can be viewed as a model of the change of interbreeding probabilities through time. This tree is not a history of realized genetic descent because it does not say what passage of genetic material actually happened. It does not rule out the possibility that random matings just by strange chance might have resulted in the partition of a panmictic population for a few generations. It specifies only the probabilities for various patterns of genetic descent. In one sense, such a phylogeny is more than a "mere" history of genes because it adds a notion of cause, even if it does not indicate what biological process controls the interbreeding probabilities. In another sense, it is less than a history of genes because it does not fully specify the genetic outcome.

It is worthwhile to take a moment to examine the strange beast that such a phylogeny is. Some authors (myself included) have characterized the biological species concept ("groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups"; Mayr, 1942:120) as being prospective, focusing on the future (Maddison in Vlijm, 1986; Kluge, 1990; O'Hara, 1993; Baum and Shaw, 1995), and thus inappropriate for use in interpreting evolutionary history. I argued that "dreams of the future will not help us; since all of our

data are of the present and past, our units by which we interpret these data must also be strictly historical" (Maddison in Vlijm, 1986:44). From this point of view, the smallest units depicted in a phylogeny should be strictly retrospective (historical), not defined in terms of their expected future behavior.

However, it now seems that phylogeny can be viewed so that its more basal units, its branches, are defined with reference to interbreeding potentials. The coalescence theorists have used phylogeny as a model of the breakup of panmictic populations (e.g., Pamilo and Nei, 1988). This breakup could be due to the evolution of reproductive isolating mechanisms or to geographic separation (in this respect the branches do not satisfy the original biological species concept), whatever changes probabilities of interbreeding. Even though the model concerns potentials and probabilities and would thus seem to be prospective, coalescence theory successfully treats it as a historical model, asking "What if interbreeding potentials had fragmented like this?" and then following its consequences. In this view, the phylogeny is a history of what interbreeding could and could not have happened. At each moment of history, the phylogeny says what was most likely to happen next, genetically. It is a history of genetic potentialities (Maddison, 1995), a history of what could have been.

Strange as such a concept of phylogeny may seem, a concept such as this is what many of us use when we discuss processes of evolution and imagine genes sorting themselves out with various probabilities within species following speciation events. However, I am drawn toward the concept of phylogeny as the realized genetic history, a bare history of what happened to genes. These two concepts of phylogeny find their parallels in concepts of species; the probability model is companion to species concepts based on interbreeding ability, and the genetic history model is companion to those concepts based on gene genealogies (Baum and Shaw, 1995). As with "species" (de Queiroz and Donoghue, 1988), we probably will find ourselves using "phylogeny" in both senses, genetic history or interbreeding model, depending on our needs in the particular context.

Other concepts of phylogeny are possible. For example, we could view phylogeny as an extended pedigree of individual organisms, a summary of realized matings, implicit in Hennig's well-known diagram (Hennig, 1966: fig. 6). This model is also one of potentials with respect to the realized descent of genes because of the chance process of meiosis. It is, however, more constrained than a model that fails to specify the realized matings and leaves them to probabilities. Yet other concepts of phylogeny can be found implicit in various other species concepts.

Regardless of to what we attach the name "phylogeny," we are still faced with the fact that the history of genetic descent does not take the form of a simple tree with sticklike branches. Given the centrality of genetics in our explanation of evolutionary diversity, we need to confront the composite, cloudlike nature of genetic history.

ACKNOWLEDGMENTS

I am grateful to Mark Siddall and Richard O'Grady for organizing the 1995 SSB symposium in Montreal at which these ideas were presented. I thank David Maddison for helping to clarify alternative concepts of phylogeny and for discussions over the years about the importance of considering its fine structure, Alan de Queiroz for convincing me not to be too sure about phylogeny as a history of realized genetic descent, and David Cannatella for suggesting the term "cloudogram" for Figure 9 (which I was, alas, too reserved to use). Helpful comments on the paper were given by Paul Chippindale, Keith Crandall, and Kerry Shaw. This work was supported by a David and Lucile Packard Fellowship.

REFERENCES

AVISE, J. C., J. F. SHAPIRO, S. W. DANIEL, C. F. AQUADRO, AND R. A. LANSMAN. 1983. Mitochondrial DNA differentiation during the speciation process in *Peromyscus*. Mol. Biol. Evol. 1:38–56.

BAUM, D. A., AND K. L. SHAW. 1995. Genealogical perspectives on the species problem. Pages 289–303 *in* Experimental and molecular approaches to plant biosystematics. Monographs in systematics, Volume 53 (P. C. Hoch and A. G. Stevenson, eds.). Missouri Botanical Garden, St. Louis.

BROOKS, D. R. 1981. Hennig's parasitological method: A proposed solution. Syst. Zool. 30:229–249.

BROOKS, D. R. 1990. Parsimony analysis in historical biogeography and coevolution: Methodological and theoretical update. Syst. Zool. 39:14–30.

CUMMINGS, M. P. 1994. Transmission patterns of eukaryotic transposable elements: Arguments for and against horizontal transfer. Trends Ecol. Evol. 9:141–145.

DAY, W. H. 1983. Properties of the nearest neighbor interchange metric for trees of small size. J. Theor. Biol. 101:275–288.

DE QUEIROZ, K., AND M. J. DONOGHUE. 1988. Phylogenetic systematics and the species problem. Cladistics 4:317–338.

DOYLE, J. J. 1992. Gene trees and species trees: Molecular systematics as one-character taxonomy. Syst. Bot. 17:144–163.

FELSENSTEIN, J. 1992. Estimating effective population size from samples of sequences: A bootstrap Monte Carlo integration approach. Genet. Res. 60:209–220.

FINDEN, C. R., AND A. D. GORDON. 1985. Obtaining common pruned trees. J. Classif. 2:255–276.

FITCH, W. M. 1970. Distinguishing homologous from analogous proteins. Syst. Zool. 19:99–113.

GOODMAN, M., J. CZELUSNIAK, G. W. MOORE, A. E. ROMERO-HERRERA, AND G. MATSUDA. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. 28:132–163.

HEIN, J. 1990. Reconstructing evolution of sequences subject to recombination using parsimony. Math. Biosci. 98:185–200.

HENNIG, W. 1966. Phylogenetic systematics. Univ. Illinois Press, Urbana.

HUDSON, R. R. 1983. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23:183–201.

HUDSON, R. R. 1990. Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. 7:1–44.

HUDSON, R. R. 1992. Gene trees, species trees, and the segregation of ancestral alleles. Genetics 131: 509–512.

KIDWELL, M. G. 1993. Lateral transfer in natural populations of eukaryotes. Annu. Rev. Genet. 27:235–256.

KINGMAN, J. F. C. 1982. The coalescent. Stochast. Proc. Appl. 13:235–248.

KLUGE, A. G. 1990. Species as historical individuals. Biol. Philos. 5:417–431.

MADDISON, W. P. 1995. Phylogenetic histories within and among species. Pages 273–287 *in* Experimental and molecular approaches to plant biosystematics. Monographs in systematics, Volume 53 (P. C. Hoch and A. G. Stevenson, eds.). Missouri Botanical Garden, St. Louis.

MADDISON, W. P. 1996. Molecular approaches and the growth of phylogenetic biology. Pages 47–63 *in* Molecular zoology: Advances, strategies, and protocols (J. D. Ferraris and S. R. Palumbi, eds.). Wiley-Liss, New York.

MAYR, E. 1942. Systematics and the origin of species. Columbia Univ. Press, New York.

NEI, M. 1987. Molecular evolutionary genetics. Columbia Univ. Press, New York.

NEIGEL, J. E., AND J. C. AVISE. 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. Pages 515–534 *in* Evolutionary processes and theory (E. Nevo and S. Karlin, eds.). Academic Press, New York.

O'HARA, R. 1993. Systematic generalization, historical fate, and the species problem. Syst. Biol. 42:231–246.

PAGE, R. D. M. 1988. Quantitative cladistic biogeography: Constructing and comparing area cladograms. Syst. Zool. 37:254–270.

PAGE, R. D. M. 1993. Genes, organisms, and areas: The problem of multiple lineages. Syst. Biol. 42:77–84

PAGE, R. D. M. 1994a. Maps between trees and cladistic analysis of historical associations among genes, organisms and areas. Syst. Biol. 43:58–77.

PAGE, R. D. M. 1994b. Parallel phylogenies: Reconstructing the history of host–parasite assemblages. Cladistics 10:155–173.

PAMILO, P., AND M. NEI. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

ROBINSON, D. F. 1971. Comparison of labeled trees with valency three. J. Combin. Theory 11:105-119.

ROTH, V. L. 1991. Homology and hierarchies: Problems solved and unresolved. J. Evol. Biol. 4:167–194.

TAJIMA, F. 1983. Evolutionary relationships of DNA sequences in finite populations. Genetics 123:229–240.

TAKAHATA, N. 1989. Gene genealogy in three related populations: Consistency probability between gene and population trees. Genetics 122:957–966.

TAKAHATA, N., AND M. NEI. 1985. Gene genealogy and variance of interpopulation nucleotide differences. Genetics 110:325–344.

VLIJM, L. 1986. Ethospecies: Behavioral patterns as an interspecific barrier. Actas X Congr. Int. Aracnol. 2: 41–45.

WALSH, J. B. 1987. Sequence-dependent gene conversion: Can duplicated genes diverge fast enough to escape conversion? Genetics 117:543–557.

WALSH, J. B. 1995. How often do duplicated genes evolve new functions? Genetics 139:421–428.

WU, C. I. 1991. Inferences of species phylogeny in relation to segregation of ancestral polymorphisms. Genetics 127:429–435.

Received 24 September 1996; accepted 4 February 1997
Associate Editor: John J. Wiens