Concordance trees, concordance factors, and the exploration of reticulate genealogy

David A. Baum

Department of Botany, University of Wisconsin, Madison, Wisconsin 53706, U.S.A. dbaum@wisc.edu

Reticulate evolution, whether due to lateral gene transfer, introgression, incomplete lineage sorting, or hybrid speciation, is a ubiquitous phenomenon, yet systematists have few tools for analyzing and summarizing the resulting genealogical discordance. Here I argue that a primary concordance tree, a tree built from clades that are true of a plurality of the genome, provides a useful summary of the dominant phylogenetic history for a group of organisms. Residual historical signals can also be extracted in the form of secondary, tertiary, etc. concordance trees, which are built from clades that are present in the genome but contradict clades on the primary concordance tree. Clades on concordance trees can be annotated with their *concordance factor* (CF), the proportion of the genome for which the clade is true. Concordance trees can potentially be estimated either from population histories or from multilocus molecular datasets. In the latter case one can use the recently developed Bayesian concordance analysis to obtain point estimates and credibility intervals on CF's. I argue that the concepts of concordance trees and concordance factors inform the debate on how to integrate information from multiple independent datasets and help clarify the nature of the boundary between reticulate and divergent genealogy.

KEYWORDS: concordance, consensus, genealogical discordance, phylogeny, reticulate evolution, taxonomy

INTRODUCTION

A divergent phylogenetic tree is a useful tool for communicating information about evolutionary history and for quantitative analyses of evolutionary patterns. However, relationships are not always strictly tree-like, which raises the question of whether the tree metaphor can be adapted to cases where there is considerable reticulation. In the face of reticulate evolution, must we abandon tree conventions (e.g., Rivera & Lake, 2004; Bapteste & al., 2005), or can we find effective means to use tree concepts to analyze genealogical histories that are not fully divergent?

The aim of this paper is to develop a conceptual framework for summarizing the intermixing of divergence and reticulation. I start by considering the theoretical relationship between population trees and genealogical histories and suggest that a key consideration is the proportion of the genome for which a given clade is true, the *concordance factor* (CF) of the clade. I propose that a valuable summary of the dominant phylogenetic history for a group of organisms is the primary concordance tree, a tree composed of clades with higher concordance factor than any contradictory clade. I then discuss how concordance factors and concordance trees can be estimated in practice and consider whether these concepts help clarify the aims of taxonomy.

THE NATURE OF RETICULATION

Reticulation, the genealogical pattern that predominates within sexual populations, and divergence, the pattern that predominates in asexual organisms and among non-hybridizing higher-level taxa, are sometimes thought of as being distinguished by the pattern of organismic descent relations. However, it can be fruitful to instead differentiate these concepts based on their expected effects on gene genealogies (Avise & Ball, 1990; Baum & Shaw, 1995; Maddison, 1997). Considering a sufficiently small stretch of genetic material, in the limit a single base-pair, the underlying genealogy will always be strictly tree-like: lineages, once diverged, will never fuse. Given this, we can define reticulation as applying whenever different parts of the genome have different genealogical histories and divergence can be understood as applying whenever there is concordance among gene genealogies. It should be noted that I will use the term concordance to refer to agreement between gene trees rather than agreement between gene trees and "species trees" (as in Rosenberg, 2002).

All kinds of reticulation ultimately depend upon recombination, but three kinds of mechanism can be distinguished (Table 1): (1) lateral gene transfer, where a small number of genes are passed between distantly related organisms by mechanisms other than normal sexual reproduction; (2) reproduction within populations, incomplete

lineage sorting, and introgression, where discordant genealogies arise due to normal sexual reproduction; and (3) hybrid speciation and vertically inherited endosymbiosis, where two genomes come together and form a permanent association. In the case of horizontal gene transfer, most of the genome shares a common history (assuming the other mechanisms have not occurred), the *primary history*, whereas those genes transmitted horizontally have tracked an alternative path, a *minor history*. In the case of "normal" sexual reproduction there will either be multiple minor histories (e.g., within a panmictic population) or there will be one primary history (reflecting the history of population divergence and restricted gene flow) and a number of minor histories (gene trees that have been subject to alternative patterns of lineage sorting). Such minor histories that arise due to incomplete lineage sorting may be distinguished from the minor histories that arise from horizontal gene transfer because, at least in the case of lineage sorting involving three lineages, one expects the alternative genealogies to occur at equal frequency (Pamilo & Nei, 1988; Rosenberg, 2002). To emphasize this point, in cases where there are multiple conflicting gene histories that have the same underlying frequency, I will call the component histories, *cominor histories*. In the case of hybrid speciation one expects two *coprimary histories*, corresponding to the two fused genomes, and no minor histories (except as a consequence of the other processes).

Table 1. Mechanisms of reticulation and their genealogical consequences.

Mechanism	Dominant Genealogy	Secondary Genealogy
Horizontal transfer	Primary history	Minor history
Lineage sorting	Primary history	Cominor histories
Hybrid speciation	Coprimary histories	–

CONCORDANCE TREES FROM POPULATION HISTORIES

Imagine that one had access to the complete population history for a group of extant populations, which is to say, for all times in the past, one knew the assignment of individuals to interbreeding populations, the descent relations of those populations, and the extent of gene flow between populations. Due to incomplete lineage sorting, reticulation could apply even if the population tree was fully divergent. How could we summarize the reticulate structure within such a known population history?

For simplicity of exposition, let us begin by considering a case in which the only sources of reticulation are ones that generate primary, minor, and cominor histories (see Table 1), not mechanisms that generate coprimary histories (e.g., hybrid speciation). Given the complete population history and an appropriate population genetic model, one could obtain the expected distribution of gene genealogies, i.e., the probability that a randomly sampled gene from a set of extant tips would have each possible topology. Figure 1 illustrates this with a simple example in which one internal population lineage has been subject to incomplete lineage sorting and there has been one recent lateral gene transfer event. For such a simple case it is relatively easy to enumerate the possible gene genealogies and their expected frequencies (Fig. 1).

How can we summarize the distribution of potential gene genealogies in such a way that we can readily visualize the primary history? One popular approach is consensus network analysis, as implemented in the program *SplitsTree* (Holland & al., 2004; 2006; Huson & Bryant, 2006), which combines topologies representative of each gene into a single network with the frequency of a topology shown by the length of edges consistent with that topology. While a consensus network may provide a useful snapshot of the distribution of genealogies, the choice of which trees should be used to represent each gene is problematic. Furthermore because consensus network approaches do not yield fully divergent trees (except in cases of complete concordance), downstream tasks that use divergent structures, most notably the generation of hierarchical taxonomies, are confounded. I wish to propose that a *primary concordance tree*, a tree composed of clades that are true for a plurality of the genome, provides a useful first order summary of the primary history.

A standard majority-rule consensus procedure can be used to construct a primary concordance tree. All clades that occur in any of the predicted gene trees are ranked in order of their frequency in the genome, their concordance factor (CF). Table 2 illustrates this step for the simple example shown in Fig. 1. The next step is to build a tree by considering the clades from highest to lowest CF, adding them to the tree if and only if they are compatible with all higher ranked clades. The procedure will continue until either a fully resolved tree has been obtained or all compatible clades have been added to the tree. It is worth noting that this procedure allows clades (partitions in the unrooted case) with CF's below 50% to be included on the tree, provided that they are not contradicted by any clade/partition with higher CF.

In the case of conflicting clades with equal CF's, as might arise when there is a hard polytomy in the population tree, one could decide that clades with exactly equal CF's would cancel each other out such that none of them would appear on the primary concordance tree. One virtue of this approach is that the primary concordance tree will depict polytomies at nodes that correspond to polytomies on the

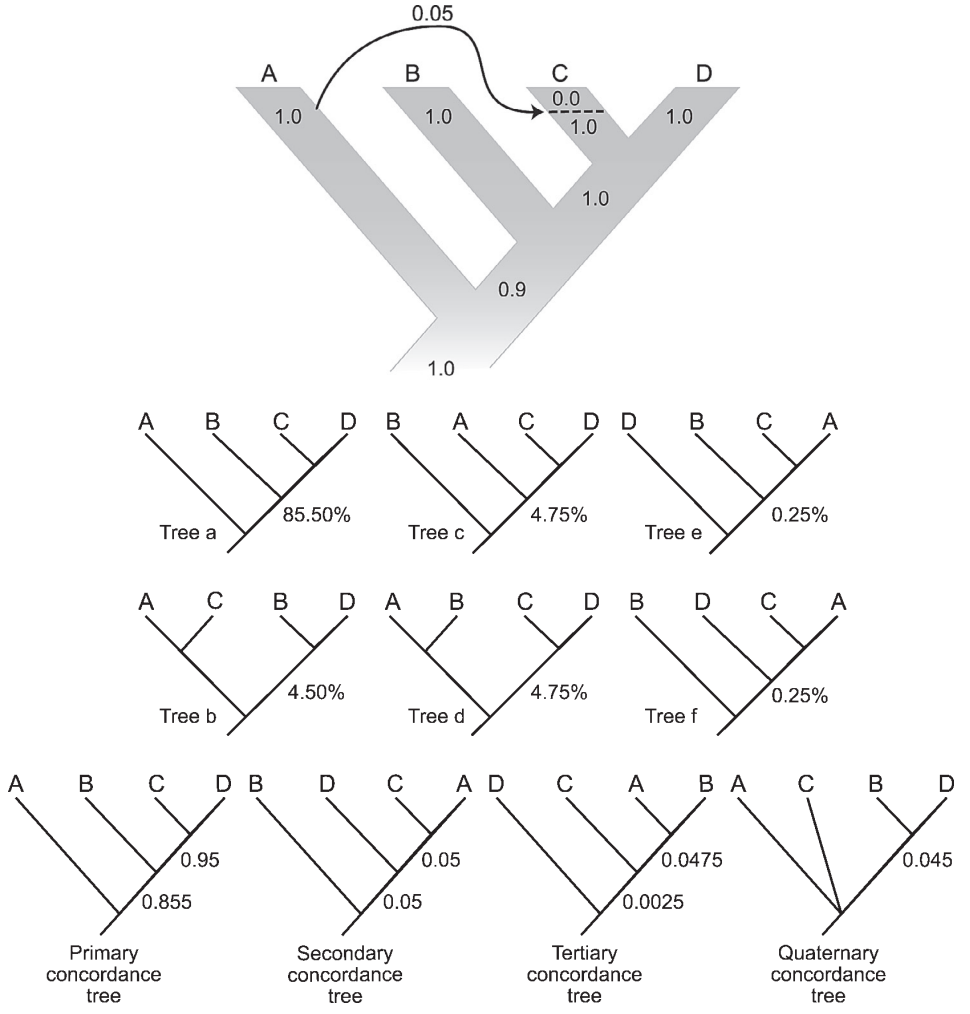

Fig. 1. A simplified example, showing the relationship between a population history, the distribution of gene genealogies, and the corresponding concordance trees. The top panel shows a simple population tree for four populations (A–D). Two internal population lineages and all terminal lineages were long enough (relative to the population size) for 100% (1.0) of genes to follow the population tree, whereas one internal lineage has been subject to incomplete lineage sorting such that only 90% of genes have followed the population history, whereas the other 10% have followed one of the two alternative histories: (A,B)(C,D) or (B(A(C,D). Additionally there has been recent directional introgression of 5% of the genome from A to C. Given this population tree, six distinct genealogies are possible, and the probability of randomly drawing a gene showing each can be calculated (middle panel). This distribution of gene genealogies can be summarized with four concordance trees, as shown in the lower panel. The primary concordance tree provides an effective summary of the primary history.

population tree. On the other hand, this method has the disadvantage that with real data (see subsequent sections) there will be statistical uncertainty as to whether clades have identical CF's. In light of this I will adopt an alternative procedure: when conflicting clades with exactly equal CF are encountered, an arbitrary decision is made as to the relative ranking of those clades. When adopting such a rule it is important to keep track of clades that conflict with clades whose CF's are (or could be) equal.

If one sampled multiple, unlinked genes randomly from within a panmictic population one would expect conflicting clades to occur at the same frequency, in which case the primary concordance tree would be composed of various randomly selected clades, each with a very small concordance factor. At the other extreme, distantly related organisms might yield a primary concordance tree with a CF of 1.0 (= 100%) for all clades, indicating complete unanimity in history across the genome. Between these two extremes, clades with CF's less than 1.0 can be seen as representing a blend of reticulate and divergent signal, quantified by the proportion of the genome conforming to the primary history. Although conventionally we think of reticulation as being more of an issue when studying closely related organisms there is no logical reason to suppose that clades will necessarily have higher CF's than their subclades.

Table 2. The concordance factors of clades found among the gene trees predicted for the example shown in Fig. 1.

Clade	Trees containing the clade	Concordance factor
(CD)	a, c, d	95.00%
(BCD)	a	85.50%
(AC)	b, e, f	5.00%
(ACD)	c, f	5.00%
(AB)	d	4.75%
(BD)	b	4.50%
(ABC)	e	0.25%

A primary concordance tree depicts the primary history, the "central trend in the rich patchwork of evolutionary history" (Wolf & al., 2002), and is arguably the divergent history of most interest to evolutionary biologists. However, it is worth considering how one could extract and represent minor and cominor histories, as would arise for example due to lateral gene transfer, introgression, or incomplete lineage sorting. The procedure to imagine (similar to reverse successive weighting; Trueman, 1998) is to apply the concordance procedure a second time to those clades that were not added to the primary concordance tree. The resulting *secondary* concordance tree (Fig. 1) would be composed of clades with a CF less than 1.0 minus the CF of any contradictory clade on the primary concordance tree (thus no clade on a *secondary concordance tree* can have a CF > 0.5). It is possible to imagine reiterating this procedure, filtering the original distribution to extract genealogical patterns that contradict both primary and secondary concordance trees; finding *tertiary concordance trees,* and so on until all minor histories have been accounted for (Fig. 1).

The preceding thought experiment included primary, minor, and cominor histories, but not coprimary histories, as might arise from hybrid speciation. When coprimary histories are present, one can build a concordance tree using the same approach (Fig. 2). Because multiple conflicting clades are likely to have equal CF's, arbitrary decisions made in ranking clades with identical CF can have a major impact on the resultant trees. This may be disconcerting, but in practice the problem is minor because few exact ties will be found.

CONCORDANCE TREES FROM MULTILOCUS MOLECULAR DATA

Until now I have assumed that we are starting with a known population tree and are using concordance trees to summarize the expected distribution of gene-genealogies

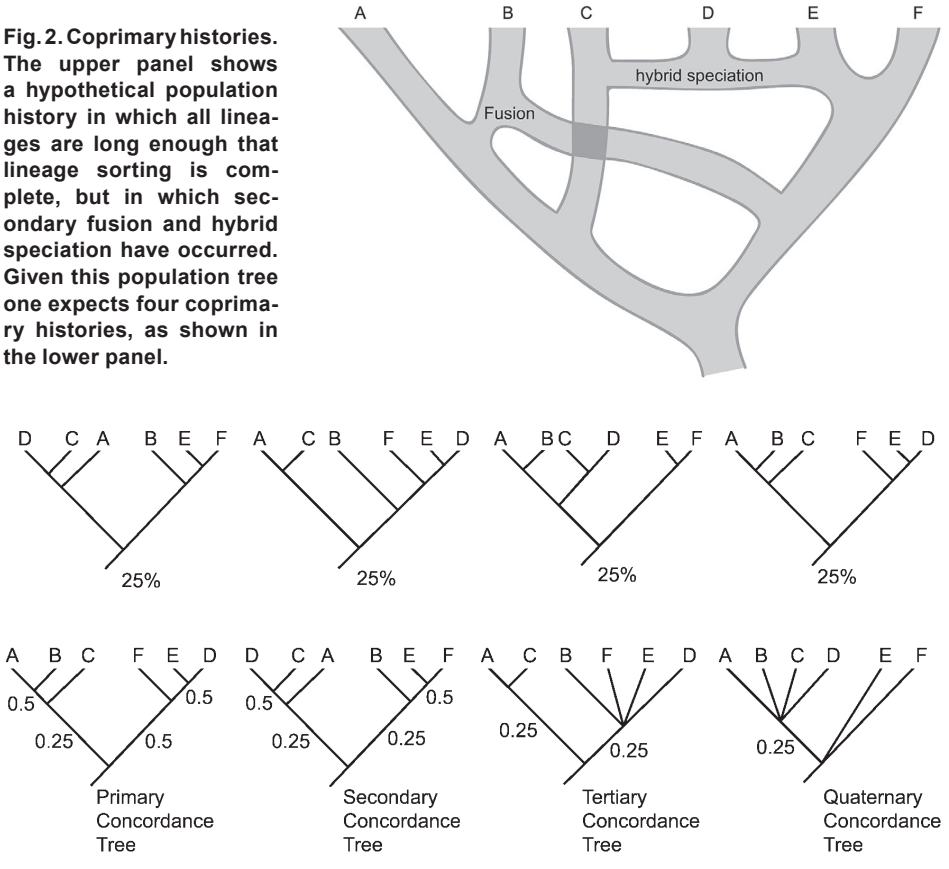

Fig. 2. Coprimary histories. The upper panel shows a hypothetical population history in which all lineages are long enough that lineage sorting is complete, but in which secondary fusion and hybrid speciation have occurred. Given this population tree one expects four coprimary histories, as shown in the lower panel.

in the genome (Fig. 3A). However, in practice one never has a known population history. At best one has molecular data for a number of loci for a set of individuals representative of the taxa under study. If we are prepared to make some assumptions about the nature of populations and genetic processes, these data contain information on the history of the populations from which they are drawn. Therefore, the ideal way to estimate a primary concordance tree would be to first estimate the population history and then calculate the expected frequencies of different genealogies while integrating over uncertainty in the population history (Fig. 3B). Theoretically one could then obtain a point estimate of the CF (with a confidence or credibility interval) for all clades, and could use these estimates to build concordance trees. However, although great progress is being made (e.g., Rosenberg, 2002; Liu & Pearl, 2006), it is not yet practical to reliably infer a population tree from molecular data, especially in cases that involve reticulation at the population level (e.g., introgression, lateral gene transfer).

Instead of estimating a population history from the molecular data it seems possible that one might be able to use the sample of gene genealogies represented in a dataset to directly estimate the proportion of the genome with different histories (Fig. 3C). We could then take the set of estimated gene genealogies, assume that they represent an unbiased sample of the genome (or develop ways to correct for biases such as linkage), and build concordance trees directly from the estimated gene trees. The resulting concordance trees should then provide a useful estimate of the primary history and the degree of reticulation/divergence at various points in that history.

The most straightforward way to estimate the primary concordance tree from multilocus sequence data would be to assign each gene to its optimal tree (e.g., using maximum likelihood analysis) and then build a majority-rule consensus tree. Such an approach resembles the most common implementation of consensus network approaches (e.g., Holland & al., 2004). This basic strategy is problematic because, by using a single optimal tree, one is not taking account of uncertainty in gene tree estimation. Furthermore, even when modified so as to weight gene trees or individual clades on those trees based on measures of statistical confidence (Holland & al., 2006), such approaches do not provide measures of statistical confidence as to the proportion of the genes having a given clade, and do not readily allow information on one gene's genealogy to shape inferences for other genes (Ané & al., 2007).

A more promising approach is provided by Bayesian concordance analysis (Ané & al., 2007), a method for estimating concordance factors and primary concordance trees from multilocus sequence data. The principle of Bayesian concordance analysis is that the amount of overlap in the Bayesian posterior probability distributions of trees for different genes contains information on the genes' degree of concordance. However, in order to extract this information in a Bayesian framework, we need to specify our prior expectations as to the degree of genealogical concordance in the dataset. The basic approach of Bayesian concordance analysis is to start by obtaining a Bayesian posterior distribution over topologies for each gene and to then use the expected degree of concordance among genes to correct each gene's posterior based on the information in the other genes. Then, by looking across the genes, one can determine the mean number of sampled genes having the clade, the *sample-wide* CF, with a credibility interval that takes account of uncertainty in the individual gene-trees. Finally, by specifying a genome size, one can obtain an estimate of the proportion of the genome having the clade, its *genome-wide* CF.

After these steps have been completed one has a list of clades each with a mean and credibility interval for its sample-wide and genome-wide CF's. Concordance trees can then be obtained by creating a tree composed only of clades that are not contradicted by another clade with

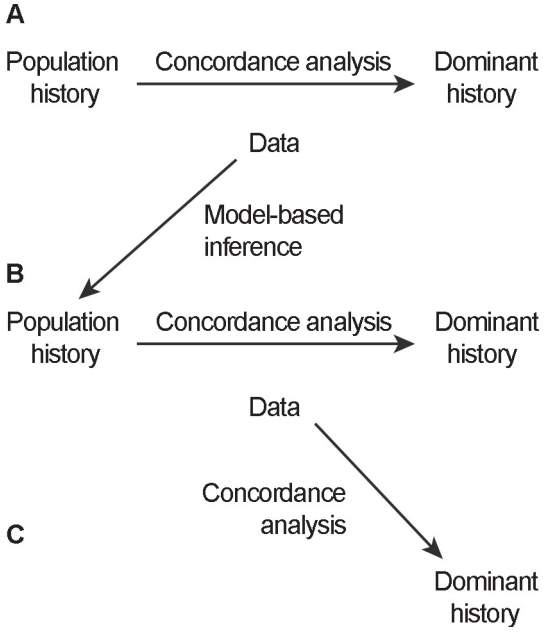

Fig. 3. Relationships between data, population histories and concordance trees. A. In the hypothetical case that one knew the population history one could calculate the expected frequency of different genealogies and then summarize this distribution with a set of concordance trees. B. Given multilocus sequence data for a set of randomly selected genes and a sufficient model of evolution one could theoretically estimate the population history. This in turn could be used to estimate the primary concordance tree for a given inheritance pattern (e.g., diploid-nuclear). C. One can estimate concordance trees directly from multilocus sequence data using a method such as Bayesian concordance analysis.

a higher mean CF. Clades on the resulting tree can be annotated with a CF and an associated credibility interval. It is also helpful to note how certain we can be that the clades on the primary concordance tree are true of more of the genome than any contradictory clade. Two examples of estimated concordance trees are shown in Fig. 4. It should be stressed that the concordance factors are not estimates of support (analogs to bootstrap percentage or Bayesian posterior probabilities), but are estimates of the proportion of the sampled genes or the genome for which the clade is true.

The first case involves a dataset of 106 nuclear genes from eight yeast species, as originally published by Rokas & al. (2003) and reanalyzed by Ané & al. (2007). In this case there is very little discordance in the genome as shown by the fact that all clades on the primary concordance tree have a CF whose 95% credibility interval is fully above 50%. This means that more of the genome has these clades than has tracked any contradictory clades and, thus, we can be confident that the topology shown is indeed the primary history for the taxa studied. At the same time most of the clades have a 95% credibility interval that also excludes 100%, meaning that there is strong evidence of the existence of discordant (minor) histories in these yeast genomes.

The second case involves data of ten nuclear genes for nine cotton species, as originally published by Cronn & al. (2003) and reanalyzed here. Not only are fewer genes sampled from the genome than in the yeast case, but two genes are each missing for one taxon, and several genes have broad single-gene posterior distributions. Thus, it is not entirely surprising that the credibility intervals on

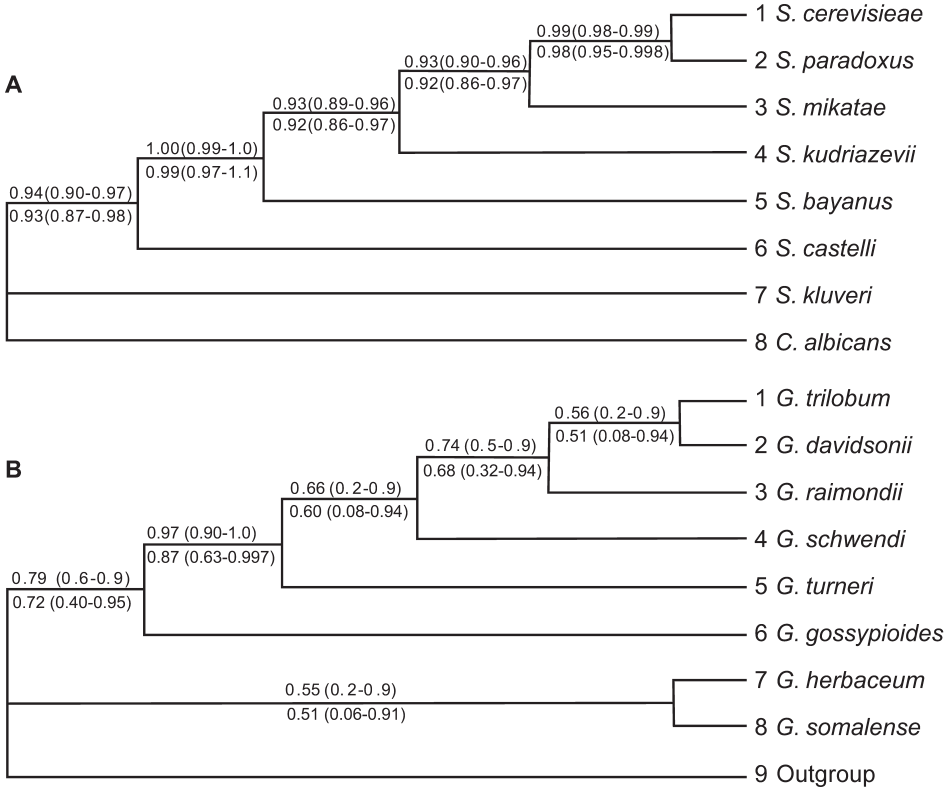

Fig. 4. Examples of the results of Bayesian concordance analysis of real data. A. Results of concordance analysis of eight yeast species analyzed with 106 nuclear genes. Details of this analysis and the priors used can be found in Ané & al. (2007). B. Results of concordance analysis of nine cotton species for ten nuclear genes (S. DeWitt Smith and D. A. Baum, unpubl. data). The original data were published by Cronn & al. (2003). Bayesian concordance analysis was conducted using BUCKy (Ané & al., 2007) with a value of 1.0 for α, the parameter that summarizes prior expectations of the amount of concordance among gene trees. We obtained the genome-wide CF estimates using the method described in Ané & al. (2007), assuming that the yeast genome size includes 6,000 genes comparable to those sampled, and that the cotton genome includes 50,000 genes (the number used had little impact on the conclusions; Ané & al., 2007). For both trees, each clade is annotated with the posterior mean concordance factor and its 95% credibility interval. The numbers above the branches refer to sample-wide concordance factors and those below the branches refer to genome-wide concordance factors. Note that, under these values of α, the prior expectation is that concordance factors will tend to be low. Thus, the posterior mean genome-wide CF's tends to be lower than the corresponding sample-wide CF. This effect is especially pronounced when extrapolating from smaller amounts of data. Likewise, with a small number of genes sampled, credibility intervals on genome-wide CF's can be very wide.

CF's are very broad, especially the genome-wide CF's (Fig. 4). While all clades in the primary concordance tree have a mean posterior CF of 0.5 or above, one clade (*G. raimondii* plus *G. trilobum*) that contradicts the primary concordance tree has a mean sample-wide CF of 0.37, which is within the credibility interval of a contradictory clade that appears on the primary concordance tree (*G. davidsonii* plus *G. trilobum*). Only three clades have a sample-wide CF credibility interval that fully exceeds 0.5 and only one clade has a genome-wide CF credibility interval that fully exceeds 0.5.

DISCUSSION

Number of genes. — If one takes the view that the main aim of systematics is to estimate *the* tree, under the implicit assumption that there is a single tree, then all one needs to do to achieve that goal is to keep adding gene sequence data until a combined analysis yields a well resolved and well supported tree. However, if one subscribes to a concordance worldview in which one allows that different parts of the genome could have tracked different histories, then the challenge becomes amplified. In this case it is not enough to obtain a well resolved total evidence tree, one would also want to estimate the proportion of the genome for which each clade on that tree is true. As illustrated by Fig. 4B, even datasets that are considered large by contemporary systematic standards may yield poor estimates of the concordance tree. For example, even with ten nuclear genes there is only one clade that can be asserted with 95% credibility to be true of a majority of the cotton genome. Does this finding imply that Bayesian concordance analysis is not of general use?

The uncertainty in the cotton dataset need not be taken as indicating that Bayesian concordance analysis is not worth using more widely. First, several of the individual genes yielded rather flat posterior distributions implying either that they lack sufficient variation or that there has been excessive homoplasy, for example due to intragenic recombination. Also, the fact that two genes lacked data for one species added considerable imprecision. More importantly, because there appears to have been a rather rich history of reticulation in the cotton clade (Cronn & al., 2003, and references therein), a sample of ten genes appears insufficient to pick out the primary history. In contrast, among yeasts there appears to have been little enough reticulation that as few as eight genes consistently yields a well supported total evidence tree that is identical to the inferred primary history (Rokas & al., 2003). This shows that the biology of the organisms will be a major determinant of the sampling effort needed to obtain good estimates of the primary history and the concordance factors of the component clades. Luckily, rapid advances in sequencing technology are opening the prospect of genomic scale data for an increasing diversity of organisms.

Diploidy. — In building a primary concordance tree a problem arises with diploid organisms due to the presence of two gene copies per individual. Consensus methods (as used to build a primary concordance tree) require that each tip in an input tree be equated with one and only one tip in each of the other input trees. In the case of diploids, one is faced with the problem of how to decide which allele from organism 1, gene 1 should be equated with which allele for organism 1, gene 2. My proposed solution is to correct the posterior distributions in such a way that each individual organism is represented only once on each tree (Fig. 5). In doing this one can maintain the correct posterior probabilities by dividing the posterior probability of a tree among all possible samplings of alleles represented on that tree. For example, if a tree with an

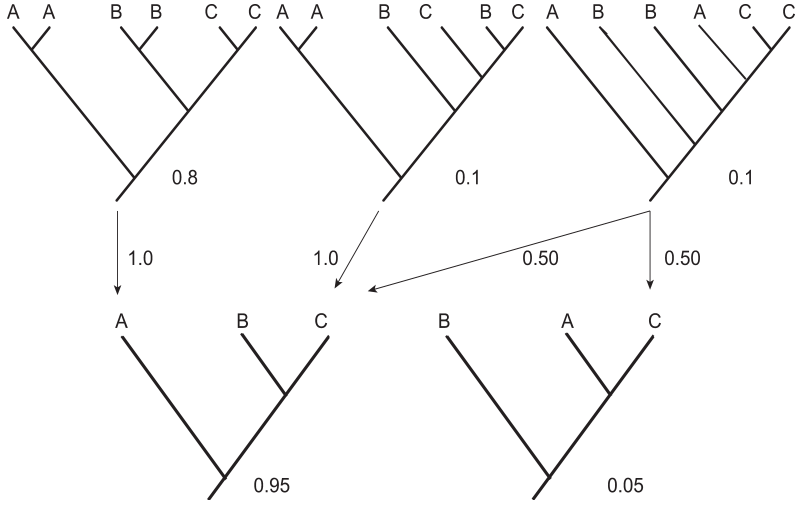

Fig. 5. An illustration of how posterior distributions can be corrected so as to only contain trees with a single tip per individual per gene. In this simple example, the posterior distribution for a single gene (which has two alleles for each of three individuals) is composed of three trees. The left two trees (PP = 0.8 and 0.1) imply the same topology, (A(B,C)), regardless of which allele is sampled per individual. When sampling one allele per individual randomly from the third tree (pruning the other allele, but retaining the basic topology) there is a 50% probability of obtaining the (A(B,C)) topology and a 50% probability of obtaining the (B(A,C)) topology. Therefore, after correcting the posterior distribution one has two trees, one with PP = 0.95 and the other with PP = 0.05.

8% posterior probability contained two separate regions where alleles from a single individual were not sister to each other, one could compress the posterior to four trees, each containing only one tip per individual, each with a 2% posterior probability. This procedure makes sense in that each allele from a heterozygote can be seen as having an equal probability of having been sampled in the case that only one gene (or gamete) was sampled per individual. Once each gene tree's posterior distribution has been corrected to include one tip per sampled individual, it is then possible to input all the individual gene posterior distributions into the second stage of Bayesian concordance analysis or another method (e.g., consensus network analysis) so as to estimate a posterior distribution of concordance factors.

Organellar versus nuclear data. — It is widely appreciated that the entire plastid or mitochondrial genomes of diploid eukaryotes each constitutes a single linkage group (hence gene in the recombinational sense) and, thus, all information obtained from either organellar genome should be combined into a single gene genealogical estimate even if multiple "genes," in the sense of open reading frames, are sampled (Doyle, 1992). A particular problem arises in estimating a primary concordance tree, however, due to the different inheritance pattern of organellar genes (haploid, uniparental) versus most nuclear genes (diploid, biparental). For the same population history one would predict a different distribution of gene genealogies for organellar and nuclear genes. Hence, the true CF of a clade could be different for the nuclear and cytoplasmic genomes. Likewise, genes in the non-recombining regions of sex chromosomes would be expected to show a distinct pattern of discordance from that applying to autosomal genes. How could one combine these into a single concordance analysis?

The ideal solution would be to use all the data to jointly estimate a single population tree, accommodating different inheritance patterns when using data of each kind. Having estimated a population tree (with error estimates) one could obtain a concordance tree for any particular inheritance pattern using analytical methods or simulation. However, methods for such analysis are not yet mature.

If, instead, one directly analyzed multilocus sequence data from a set of gene regions that included different inheritance patterns, a single concordance analysis would be misleading. The concordance factors obtained could not be viewed as valid estimates of the true genome-wide CF's, but rather an average across the sampled inheritance patterns, weighted by the number of genes in each pattern and their degree of phylogenetic signal. As a result the CF's estimated would be abstractions of little biological meaning. Rather, it would be preferable to estimate the primary concordance tree only for autosomal nuclear genes. While one might also aim to directly estimate the primary concordance trees for other inheritance patterns, we will rarely have data from enough potentially recombining genes to make concordance analysis worthwhile.

An emphasis on nuclear genes for concordance analysis should not be taken to imply that plastid and mitochondrial markers do not have a great deal to contribute to the study of reticulate evolution. In cases where there has been little reticulation, there is every reason to assume that organellar data would help infer the primary history of the entire genome. In cases where reticulation has been more pervasive, cytoplasmic markers are still valuable, for example to help determine whether gene flow was through seed or pollen and to evaluate whether incomplete lineage sorting or introgression are more likely explanations for discordance. That being said, if we want concordance trees to serve as a general device for summarizing the biological history of reticulation in a group, it seems wise to restrict our attention to nuclear, autosomal genes.

Combined, consensus, or concordance analysis? — The concordance approach to the analysis of multiple datasets is a modified consensus tree method and might be viewed as being antithetical to combined or "total evidence" analysis (e.g., Barrett & al., 1991; Eernise & Kluge, 1993; de Queiroz & al., 1995). It is therefore important to explore the relation between combined and concordance analysis. Specifically, what does each method tell us and when might they be called for?

Combined analysis of multiple datasets is justified in cases where each dataset has evolved according to the same underlying history, with differences in the estimated trees being due to sampling error or model mis-specification. In that case, combined phylogenetic analysis improves the signal to noise ratio and potentially allows for more accurate estimation of the single, shared genealogy (e.g., Bull & al., 1993). In this circumstance, combined analysis provides more accurate information than can be obtained by analysis of individual datasets followed by consensus generation (Barrett & al., 1991).

In cases where genes have tracked more than one underlying history, some of the differences among datasets would not be due to sampling error but to genealogical discordance. In this case we may wish to estimate the primary history with some information as to the extent to which different genes have followed that history. Is combined analysis a good method for doing this? If there were a similar amount of phylogenetic signal for each sampled gene, there is reason to suppose that a combined analysis would provide a reasonable estimate of the topology of the primary history. Indeed, Rokas & al. (2003) found good support for the probable primary history of their eight yeast species from combined analysis of multiple genes. However, because combined analysis *assumes* that there is a single divergent history and implicitly interprets

discordance as homoplasy, it is not the ideal tool. Indeed, when there has been reticulation, combined analysis may reconcile the discordant signals in the different data partitions in such a way that the inferred tree does not resemble any of the underlying histories (McDade, 1992).

When we suspect that there could have been reticulation, I would submit that one would obtain a much more direct and interpretable summary of the phylogenetic history using a method such as Bayesian concordance analysis, which is not constrained to assume that there was a single tree. Such an approach is not only well-suited to obtaining a statistically defensible estimate of the primary concordance tree but it also provides estimates of the concordance factors for individual clades. Furthermore, Bayesian concordance analysis allows one to make statements of statistical confidence in the estimated concordance factors by taking into account the prior evidence of discordance, uncertainty in gene tree estimates, and the number of genes sampled. Thus, while further methodological development is needed, I would posit that concordance analysis and the conceptual structure of concordance trees provide useful heuristics for analyzing genealogies that are neither fully reticulate nor fully divergent.

Taxonomy and the boundary between reticulation and divergence. — In this paper I have attempted to clarify notions of primary, coprimary, minor, and cominor histories so as to facilitate empirical phylogenetic and phylogeographic research. I have also presented estimated CF's as metrics with utility for the quantitative analysis of reticulate genealogies. Now I will shift from discussing estimates of CF's and concordance trees, as might be obtained from the analysis of real data, to the significance of the concordance concept itself. Given standard evolutionary models, for any homologous base pair in a set of living species, it will either be true or false that a subset of the living species forms a clade. Summing over the genome, the full set of homologous base pairs, it follows that a specified clade is true for an actual proportion of the genome, from zero to one. While we may never be able to estimate a true concordance factor with certainty, it seems valuable to consider the role that true concordance factors could play in systematic theory.

It is generally agreed that the hierarchical relationships among strictly divergent lineages is a real historical pattern worthy of formal taxonomic recognition (Hennig, 1966). In light of this it might be valuable to articulate a clear goal for systematics: the discovery and naming of those and only those clades that are true for more of the genome than they are false for. Primary concordance trees align with such an imperative. While some clades on a primary concordance tree will be too transient or local to warrant formal recognition, it is hard to see why we would wish to name any clade that is not on the true primary concordance tree of life.

Whereas divergent relationships form the foundation of systematics, fully reticulate relationships, such as among individuals within a panmictic population, are not hierarchically structured, and have not been taken as objects for study by systematists. Instead, panmixia provides a domain in which one can ignore actual history and apply equilibrium population genetic models. A question worthy of exploration is: When does one cross the boundary between reticulation and divergence; between systematics and population genetics? Or, to put it another way, which clades on a primary concordance tree are true of too little of the genome to warrant formal taxonomic recognition?

One answer to this question is that only clades with a CF of 1.0 are meaningful, because only for these clades does strict divergence apply (Baum & Shaw, 1995). For all other clades there is some discordant signal and hence a reticulate pattern applies. This argument would be sound if one were interested in pure divergence, but groups can be historically meaningful (i.e., share a significant degree of common history) even if they do not have a CF or 1.0. For example, organisms drawn form a large population that has remained genetically isolated for millions of generations need not become a clade with a CF of 1.0 (Hudson & Coyne, 2002). Thus, while I remain committed to the basic argument that reticulation and divergence can be distinguished by the transition between concordant and discordant gene genealogies, I here retract my previous advocacy of a 100% exclusivity criterion (i.e., a CF of 1.0) as *the* boundary between genealogical reticulation and divergence (Baum & Shaw, 1995).

If one admits that meaningful historical structure persists below a CF of 1.0, how little concordance is significant? A CF of 0.95 might be appealing because of its connection with the traditional type 1 error rate ($\alpha = 0.05$). However, notwithstanding the superficial similarity of CF values and bootstrap percentages or posterior probabilities, the CF is disconnected from notions of statistical uncertainty. The CF is a parameter, the proportion of the genome for which that clade is true, which happens to range from 0.0–1.0. Thus, 0.95 or 0.05 are not readily defensible threshold values.

Another numerical cut-off that might be suggested is 0.50, the logic being that a clade present in 50% of the component gene genealogies must be a majority signal because no contradictory clade with a higher CF can exist. However, although consistent in terms of the mechanics of consensus tree assembly, a 50% cutoff value seems to have operational rather than ontological appeal. What, one should ask, is so different about clades with CF's of 49% and 51%, assuming that in both cases all contradictory clades have a much lower CF?

I would propose that the search for a particular CF threshold that denotes the boundary between reticulation and divergence is doomed. The acquisition of a

divergent structure accrues gradually as a result of gene lineage extinction in reproductively isolated populations or demes (see, for example, Avise & Ball, 1990; Avise & Wollenberg, 1997; Maddison, 1997). From a genealogical perspective the speciation process reaches its ultimate point when a CF of 1.0 is attained. At the other extreme, after populations first become genetically isolated, one expects clades that coincide with the population structure to have CF's that gradually increase beyond background. It follows, therefore, that any clade that has a CF that is greater than all contradictory clades has some genealogical unity indicative of a history of genetic isolation. While one may consider some such a clades too small to be formally named, they are nonetheless meaningful historical entities. The boundary between reticulation and divergence is quantitative not qualitative, and the concordance factor is a useful index of where on that continuum a given group of organisms sits. This leads back to important real world questions such as how taxonomic practice can best mirror a primary concordance tree, how the names of clades can most effectively be anchored to clades, and how one can accommodate the concept of species. These are major questions whose resolution may be facilitated by a terminological structure that incorporates the concepts of concordance factors and concordance trees.

ACKNOWLEDGMENTS

I gratefully acknowledge the conceptual input and constructive criticisms of Cecile Ané, Ivalú Cacho, Stacey Dewitt Smith, Margaret Koopman, Bret Larget, Antonis Rokas, the UW-Madison Phylogeny Working group, and two anonymous reviewers. I thank Richard Cronn for providing the cotton data and Stacey Dewitt Smith for doing much of the reanalysis of these data. Help with artwork was provided by Kandis Elliott. This work was funded by the National Institutes of Health (GM068950-01).

LITERATURE CITED

Ané, C., Larget, B., Baum, D.A., Smith, S.D. & Rokas, A. 2007. Bayesian estimation of concordance among gene trees. *Molec. Biol. Evol.* 24: 412–426.

Avise, J.C. & Ball, R.M. 1990. Principles of genealogical concordance in species concepts and biological taxonomy. *Oxford Surv. Evol. Biol.* 7: 45–67.

Avise, J.C. & Wollenberg, K. 1997. Phylogenetics and the origin of species. *Proc. Natl. Acad. Sci. U.S.A.* 94: 7748–7755.

Bapteste, E., Susko, E., Leigh, J., MacLeod, D., Charlebois, R.L. & Doolittle, W.F. 2005. Do orthologous gene phylogenies really support tree-thinking? *BMC Evol. Biol.* 5: 33.

Barrett, M., Donoghue, M.J. & Sober, E. 1991. Against consensus. *Syst. Zool.* 40: 486–493.

Baum, D.A. & Shaw, K.L. 1995. Genealogical perspectives on the species problem. Pp. 289–303 in: Hoch, P.C. & Stephenson, A.G. (eds.), *Experimental and Molecular Approaches to Plant Biosystematics*. Missouri Botanical Garden, St. Louis.

Bull, J.J., Huelsenbeck, J.P., Cunningham, C.W., Swofford, D.L. & Waddell, P.J. 1993. Partitioning and combining data in phylogenetic analysis. *Syst. Biol.* 42: 384–397.

Cronn, R., Small, R.L., Haselkorn, T. & Wendel, J.F. 2003. Cryptic repeated genomic recombination during speciation in *Gossypium gossypioides*. *Evolution* 57: 2475–2489.

De Queiroz, A., Donoghue, M.J. & Kim, J. 1995. Separate versus combined analysis of phylogenetic evidence. *Ann. Rev. Ecol. Syst.* 26: 657–681.

Doyle, J.J. 1992. Gene trees and species trees—molecular systematics as one-character taxonomy. *Syst. Bot.* 17: 144–163.

Eernisse, D.J. & Kluge, A.G. 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. *Molec. Biol. Evol.* 10: 1170–1195.

Hennig, W. 1966. *Phylogenetic Systematics*. Univ. Illinois Press, Urbana.

Holland, B.R., Huber, K.T., Moulton, V. & Lockhart, P.J. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. *Molec. Biol. Evol.* 21: 1459–1461.

Holland, B.R., Jermiin, L.S. & Moulton, V. 2006. Improved consensus network techniques for genome-scale phylogeny. *Molec. Biol. Evol.* 23: 848–855.

Hudson, R.R. & Coyne, J.A. 2002. Mathematical consequences of the genealogical species concept. *Evolution* 56: 1557–1565.

Huson, D.H. & Bryant, D. 2006. Application of phylogenetic networks in evolutionary studies. *Molec. Biol. Evol.* 23: 254–267.

Liu, L. & Pearl, D.K. 2006. Species trees from gene trees: Reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Mathematical Biosciences Institute Tech. Report* 53. http://mbi.osu.edu/publications/pub2006.html.

Maddison, W.P. 1997. Gene trees in species trees. *Syst. Biol.* 46: 523–536.

McDade, L.A. 1992. Hybrids and phylogenetic systematics. 1. The impact of hybrids on cladistic analysis. *Evolution* 46: 1329–1346.

Pamilo, P. & Nei, M. 1988. Relationships between gene trees and species trees. *Molec. Biol. Evol.* 5: 568–583.

Rivera, M.C. & Lake, J.A. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* 431: 152–155.

Rokas, A., Williams, B.L., King, N. & Carroll, S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425: 798–804.

Rosenberg, N.A. 2002. The probability of topological concordance of gene trees and species trees. *Theor. Pop. Biol.* 61: 225–247.

Trueman, J.W.H. 1998. Reverse successive weighting. *Syst. Biol.* 47: 733–737.

Wolf, Y.I., Rogozin, I.B., Grishin, N.V. & Koonin, E.V. 2002. Genome trees and the tree of life. *Trends Genet.* 18: 472–479.