Syst. Biol. 55(1):21–30, 2006
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150500354928

Inferring Phylogeny Despite Incomplete Lineage Sorting

WAYNE P. MADDISON[1] AND L. LACEY KNOWLES[2]

[1]*Departments of Zoology and Botany and Biodiversity Research Centre, University of British Columbia, Vancouver, B.C., Canada, V6T 1Z4; E-mail: wmaddisn@interchange.ubc.ca*
[2]*Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109-1079, USA; E-mail: knowlesl@umich.edu*

Abstract.—It is now well known that incomplete lineage sorting can cause serious difficulties for phylogenetic inference, but little attention has been paid to methods that attempt to overcome these difficulties by explicitly considering the processes that produce them. Here we explore approaches to phylogenetic inference designed to consider retention and sorting of ancestral polymorphism. We examine how the reconstructability of a species (or population) phylogeny is affected by (a) the number of loci used to estimate the phylogeny and (b) the number of individuals sampled per species. Even in difficult cases with considerable incomplete lineage sorting (times between divergences less than 1 N_e generations), we found the reconstructed species trees matched the "true" species trees in at least three out of five partitions, as long as a reasonable number of individuals per species were sampled. We also studied the tradeoff between sampling more loci versus more individuals. Although increasing the number of loci gives more accurate trees for a given sampling effort with deeper species trees (e.g., total depth of 10 N_e generations), sampling more individuals often gives better results than sampling more loci with shallower species trees (e.g., depth = 1 N_e). Taken together, these results demonstrate that gene sequences retain enough signal to achieve an accurate estimate of phylogeny despite widespread incomplete lineage sorting. Continued improvement in our methods to reconstruct phylogeny near the species level will require a shift to a compound model that considers not only nucleotide or character state substitutions, but also the population genetics processes of lineage sorting. [Coalescence; divergence; population; speciation.]

The challenges associated with inferring evolutionary relationships of recently diverged species or populations differ significantly from those for deep phylogenetic divergence. At shallow time depths, hazards such as incomplete lineage sorting predominate, whereas accurate molecular phylogenetic inference at greater time depths may be hampered by saturation, misalignment, unrecognized paralogy, and interlineage inhomogeneity of models (Sanderson and Schafer, 2002; Felsenstein, 2004). The difficulties posed by incomplete lineage sorting have been well described (Avise et al., 1983; Pamilo and Nei, 1988; Takahata, 1989; Doyle, 1992; Maddison, 1997; Rosenberg, 2002, 2003): the genealogical histories of individual gene loci may appear misleading or uninformative about the relationships among species or populations because of retention and stochastic sorting of ancestral polymorphisms. This is especially likely if the widths of lineages (i.e., the effective population sizes, N_e) are large relative to their lengths (i.e., the time between divergences). In this case, genetic drift is unlikely to have time to bring loci to fixation before subsequent divergences (Pamilo and Nei, 1988). Although phylogenetic patterns generated by incomplete lineage sorting have been discussed for many years (e.g., Throckmorton, 1965; Farris, 1978; Felsenstein, 1979; Arnold, 1981), considerable work remains to develop and assess methods that consider these issues during phylogenetic reconstruction.

To illustrate the problem, we show in Figure 1 a gene tree simulated within a species tree whose time depth (t) is 100,000 generations and whose lineages have haploid effective population sizes (N_e) of 100,000. The gene tree is highly discordant with the species tree—the gene copies sampled from a monophyletic group on the species tree do not correspond to monophyletic groups in the gene tree. Our concern, therefore, is that the gene tree may not accurately reflect the species tree. However, inspecting the figure shows that the gene tree bears some (although noisy) relation to the species tree. This raises the hope that information from this confusing gene tree can be extracted to obtain an inference of the species tree. The discord between inferred gene trees and species trees has been used to extract information about species histories, and in particular estimates of species divergence times and ancestral population sizes (e.g., Edwards and Beerli, 2000; Hey and Nielsen, 2004; Rannala and Yang, 2003; Takahata and Satta, 2002; Wall, 2003). In principle, these approaches could be used to obtain a series of divergence time estimations that could then be compiled into an estimate of a larger species tree, but such a combined analysis to bridge population genetics and phylogenetic estimation has yet to be developed. These approaches nonetheless highlight the information content inherent in patterns of deep coalescence at the population genetic-phylogenetic boundary.

Sophisticated methods are available to infer gene trees, but quantitative methods to infer species trees containing those gene trees are little developed and little explored, although the groundwork for such methods is being laid (e.g., Edwards and Beerli, 2000; Takahata and Satta, 2002; Rannala and Yang, 2003; Wall, 2003; Hey and Nielsen, 2004; Degnan and Salter, 2004). To reconstruct gene trees we must consider the process of nucleotide substitution; to reconstruct species trees we must consider in addition the process of sorting of gene lineages within populations. Sorting within populations is considered by gene frequency-based methods (e.g., Edwards and Cavalli-Sforza, 1964), but these methods fail to consider fully the process of nucleotide substitution (and our ability thereby to discern genealogical relationships among alleles). Needed, therefore, are methods that consider explicitly both the processes of substitution (within gene lineages) and sorting (among gene lineages). Just as the incorporation of explicit models of evolutionary

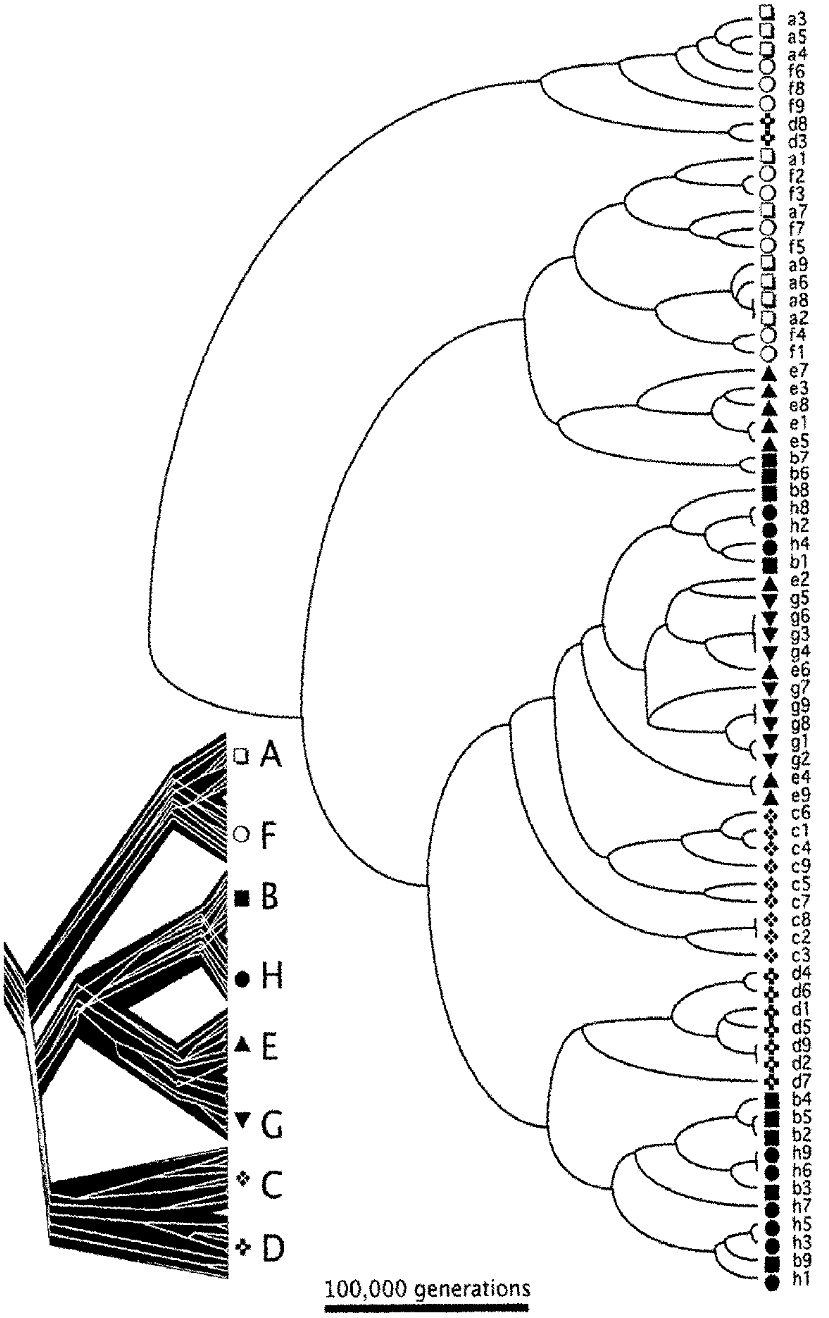

FIGURE 1. Gene tree (right) simulated by neutral coalescence within simulated species tree (lower left). Discord between species tree and gene tree shows the considerable amount of incomplete lineage sorting expected for a recent divergence (total depth of species tree from root to tips =1 N_e, where N_e = 100, 000). Nine individuals were sampled within each species. The two trees are drawn to the same horizontal scale in generations. Note that the gene tree has coalescences extending back to 300,000 generations even though the species tree is only 100,000 generations deep.

character change, whether stochastic (e.g., Felsenstein, 1981) or not (e.g, Hennig, 1966), was vital to the development of phylogenetic methods, incorporation of explicit models of lineage sorting will be needed for continued development of phylogenetic inference near the species level.

To explore whether there is sufficient information for reconstructing species trees, even when gene trees are highly discordant with the actual species tree (e.g., Fig. 1), we use two simple approaches for reconstructing species trees that explicitly consider the process of incomplete lineage sorting. The first uses reconstructed gene trees to seek the species tree that minimizes the number of deep coalescences (Maddison, 1997). The second clusters species directly by their most similar contained sequences (shallowest coalescences), which is based on a suggestion by Takahata (1989). We recognize that these methods are not ideal. Although they consider the process of incomplete lineage sorting, under neither are the actual probabilities of incomplete lineage sorting quantified using a stochastic model. We expect that likelihood and Bayesian methods that incorporate stochastic models of both nucleotide substitution and lineage sorting processes (Maddison, 1997; Rosenberg, 2002; Degnan and Salter, 2004) would perform better at reconstructing species trees than the simple methods we study here, but such probabilistic methods are not yet developed. Our goal here is not to determine the best method for reconstructing species trees, but rather to explore what information can be extracted by even crude methods. And, by studying two rather different methods, it is our hope that any common results would be general to most methods that consider lineage sorting.

In addition to exploring whether accurate species trees are recoverable in the face of considerable incomplete lineage sorting, we will also investigate the tradeoff in allocating effort to sequencing more loci versus more individuals (e.g., see Pamilo and Nei, 1988, versus Takahata, 1989). We examine this tradeoff for recent as well as older divergences, corresponding to higher and lower degrees of incomplete lineage sorting, respectively. Gains in accuracy associated with increasing the numbers of both loci and individuals sequenced are also studied. Although most of the results are presented in the context of reconstructing species phylogeny, they also apply to estimating population relationships, except that with populations there is a higher risk that shared haplotypes might reflect gene flow rather than relationship or incomplete lineage sorting. Our exploration assumes there is no gene flow; i.e., the only process that yields confusing gene trees is failure to sort ancestral polymorphisms.

METHODS

Our approach was to simulate the processes of lineage sorting and nucleotide substitution in an evolving species tree, then use the resulting sequence data to attempt to infer the species tree. The parameters used in the simulations were selected such that the extent of incomplete lineage sorting varied from very high to low (i.e., corresponding to situations of recent and older divergences, respectively). We also use simulated data that would be comparable to what a biologist might have available, with respect to the length of sequence per locus, the total number of individuals and loci sequenced, as well as the model of sequence evolution.

The simulation protocol is outlined in Figure 2. We first used speciation simulations to generate species trees. Within each species tree, we performed coalescent simulations to generate gene trees (Kingman, 1982; Hudson, 1990). We then simulated nucleotide sequence evolution along the lineages of those gene trees to generate a set of observed gene sequences. The simulated gene sequences were then used to infer a species tree using two different methods (discussed in detail below): the gene sequences were used either directly (for the Shallowest Divergence method) or via inferred gene trees (for the Minimize Deep Coalescences method). To evaluate accuracy of inference, each reconstructed species tree was compared to the original tree used in the simulation.

All simulations and inferences were done with a version of Mesquite (Maddison and Maddison, 2004c) functionally equivalent to version 1.01 and PAUP* version 4.0b10 (Swofford, 2002). Species trees were simulated by a pure birth process using Mesquite's Uniform Speciation (Yule) module. Gene trees were simulated using Mesquite's Neutral Coalescence module, which uses an exponential approximation to avoid fully explicit modeling of individuals. An effective population size (N_e) of 100,000 was used for all simulations, and the organisms (or organelles) are assumed haploid. This population size was chosen because it would be reasonable for many organisms, but we expect that the results should be robust to changes in population size as long as mutation rates are adjusted to keep average sequence divergences unchanged. Coalescence of genes was modeled backward in time within each species lineage until a divergence event (node in the species tree). At that point, any remaining (uncoalesced) gene copies were combined with those from the other lineages descendant from the node and coalescence was continued down into the ancestral lineage. Population size of the ancestral species lineage was not the sum of the two descendants, but rather was set to the common size of 100,000. For each simulated gene tree, sequence evolution was simulated along it using Mesquite's Genesis package (Maddison and Maddison, 2004a). A 1000-basepair sequence was simulated for each individual and locus according to an HKY85 model with transition-transversion ratio of 3.0 and a discrete gamma distribution with four categories and shape parameter 0.8. An ancestral sequence was assigned states randomly with probabilities 0.3 A, 0.2 C, 0.2 G, 0.3 T; this 3:2 AT:GC bias was maintained as the equilibrium frequency distribution throughout the tree. These substitution parameters were chosen to introduce some complexity to the model, but we expect that varying them would have little effect because with such small sequence divergences most mutations would be unique, nonhomoplasious, and not needing saturation corrections. Varying the parameters would therefore be equivalent to varying the overall mutation rate. Regarding mutation rate, the gene tree's branch lengths are measured in units of generations, but the rates in the model of evolution implicitly assume a different scaling (e.g., the model assumes trees will have branch lengths of a few units long, not thousands of units as when measured in generations). Because of this, the rates of the model of evolution had to be scaled down considerably. A scaling factor of 3×10^{-8} was chosen because it yielded sequence divergences comparable to those found in empirical studies (see Results, Table 1). Although we could have explored the parameter space more fully, for instance to explore other scaling factors (mutation rates), we decided that it was more important to choose a few parameters to yield empirically reasonable data, to have larger sample size, and to attempt different methods of reconstruction, than to do a massive survey of parameter

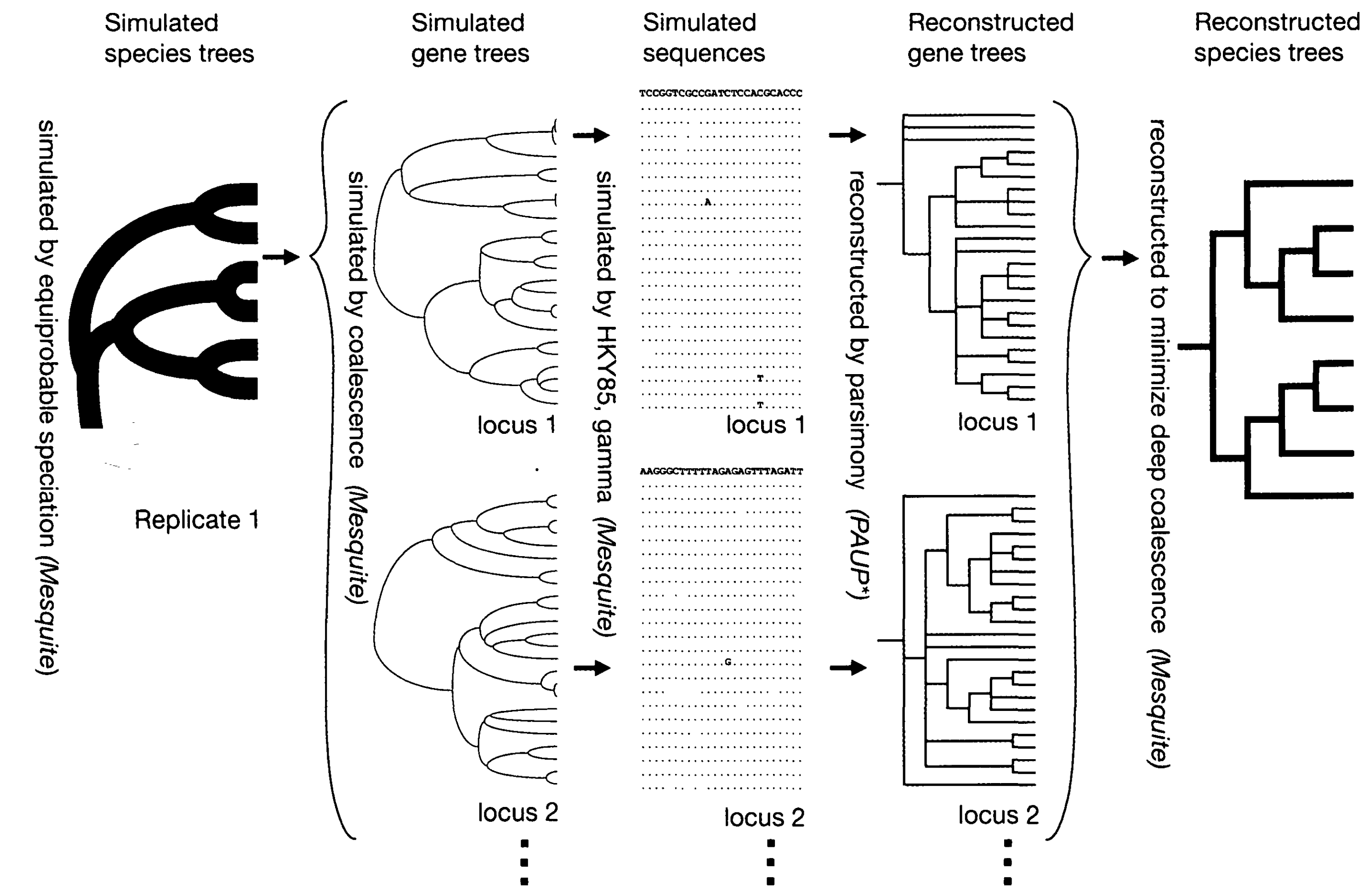
Simulated species trees
Simulated gene trees
Simulated sequences
Reconstructed gene trees
Reconstructed species trees
simulated by equiprobable speciation (Mesquite)
Replicate 1
simulated by coalescence (Mesquite)
locus 1
locus 2
simulated by HKY85, gamma (Mesquite)
TCCGGTCGCCGATCTCCACGCACCC
AAGGGCTTTTTAGAGAGTTTAGATT
reconstructed by parsimony (PAUP*)
reconstructed to minimize deep coalescence (Mesquite)

FIGURE 2. Steps involved in the simulation used to assess accuracy of species tree reconstruction using the Minimize Deep Coalescences method. Shown is one of the 500 replicates. In each replicate a species tree was simulated, within which gene trees were simulated by coalescence, then sequence evolution was simulated on the gene trees. Those gene sequences were used to reconstruct gene trees, which were then used to reconstruct species trees; the reconstructed species trees were compared to the originals. The simulations using Shallowest Divergence method followed a similar procedure except that species trees were reconstructed directly from the simulated DNA sequences.

space. Our primary goal was not an exhaustive guide to parameter space, but rather to ask whether reconstruction is possible despite incomplete lineage sorting.

The implementation of this protocol in Mesquite used a script that controlled the simulation of species trees, gene trees within them, and sequence evolution of the gene trees. In Mesquite, a tree window was set up to show a simulated species tree. The script then requested that the Batch Architect package (Maddison and Maddison, 2004b) save one or more sequence matrices obtained by simulating sequence evolution on a series of gene trees (each corresponding to one locus). As it was doing this, Batch Architect was saving PAUP* command files for the subsequent inference of gene trees during the reconstruction step. The script then asked the tree window to simulate the next species tree; the process was repeated 500 times. Other scripts were then run to use the inferred gene trees (in the case of reconstruction by deep coalescences) or the sequence matrices directly (in the case of reconstruction by shallowest divergences) to infer the species trees. Scripting files used in this study are available at http://www.systematicbiology.org.

Simulation of Species Trees

In order to understand reconstructability over a reasonably natural spectrum of topologies and branch length distributions we simulated 500 trees of eight

TABLE 1. Average amount of sequence divergence and incomplete lineage sorting observed for shallow (recent divergences) and deep (older divergences) species trees. Calculated for 10 replicates of each case; standard errors are shown in parentheses. Sequence divergences are average raw uncorrected percent pairwise differences, presented to confirm simulations generated divergences typical in empirical studies (values for 9 individuals not calculated, presumed to be bracketed by results from 1, 3, and 27). Incomplete lineage sorting is measured as the minimal number of deep coalescences required (Maddison, 1997).

	Average percent sequence divergence		Average amount of incomplete lineage sorting	
Individuals per locus and species	Recent divergences (1 N_e)	Older divergences (10 N_e)	Recent divergences (1 N_e)	Older divergences (10 N_e)
1	0.9 (±0.04)	3.9 (±0.23)	7.6 (±15.2)	1.8 (±6.3)
3	1.1 (±0.02)	4.2 (±0.07)	28.7 (±24.3)	6.9 (±9.0)
9			63.2 (±21.3)	14.7 (±7.7)
27	1.0 (±0.002)	3.5 (±0.006)	114.4 (±15.4)	25.7 (±5.0)

species each, rather than choosing a single species tree and assessing how well it can be reconstructed in many simulation replicates. To examine how the extent of incomplete lineage sorting affects the ability to reconstruct a species history, species trees were simulated to have a total time depth (summed length of branches from any terminal down to the root) of either 100,000 or 1,000,000 generations. With an N_e of 100,000, a total time depth of 100,000 (i.e., 1 N_e) generations leads to considerably more incomplete lineage sorting than a total time depth of 1,000,000 (i.e., 10 N_e) generations. Thus, there were a total of 1000 simulated trees: 500 trees at 1 N_e total depth and 500 at 10 N_e total depth. However, the topologies of the two sets of trees are identical; the latter set of trees is equivalent to the former set with branch lengths multiplied by 10.

Number of Individuals and Loci Used in Simulations

Within each species tree, 1, 3, 9, or 27 gene trees representing unlinked loci were simulated independently with either 1, 3, 9, or 27 gene sequences simulated for each locus per species. Thus, the smallest gene tree simulated had 8 sequences, with one individual sequenced in each of the 8 species; the largest gene tree simulated had 216 sequences, 27 individuals in each of 8 species. The maximum number of sequences in any replicate was limited to 216, achieved by 27 individuals × 1 locus ×8 species, 9 × 3 × 8, 3 × 9 × 8, or 1 × 27 × 8. The triangular matrix of each possible combination of number of loci and individuals sequenced per locus in each species was then used to examine how the accuracy of the phylogenetic reconstruction was affected by (a) increasing the total sampling effort per species (i.e., either 1, 3, 9, or 27 individuals sequenced, or either 1, 3, 9, or 27 loci sequenced) and (b) increasing the number of individuals per locus versus the number of loci per species for a given sampling effort.

Minimize Deep Coalescences Method

The Minimize Deep Coalescences method of reconstructing the species tree is based on searching for species trees that minimize the implied number of deep coalescences in the contained gene trees (Maddison, 1997). First gene trees were inferred from the simulated sequence data by a simple parsimony search using PAUP* (factory default heuristic search; MAXTREES = 100), which is reasonably efficient for the low levels of sequence divergence analyzed. If multiple most parsimonious trees resulted, their strict consensus was used. Species trees were inferred using Mesquite's tree search facility to find trees minimizing the total number of deep coalescences summed over the loci considered. The number of deep coalescences was counted assuming the reconstructed gene trees were unrooted. Thus, Mesquite counted deep coalescences for each possible rerooting of the gene tree, and used the smallest count from any rooting as the deep coalescence cost of that gene tree within the proposed species tree. The search used an As Is taxon addition sequence, followed by SPR branch swapping, saving only a single tree at any stage (MAXTREES = 1). This method of course is not guaranteed to find the tree minimizing deep coalescences, although the results suggest it may be a reasonable approximation.

Shallowest Divergence Clustering Method

The Shallowest Divergence method is based on Takahata's (1989) observations that the order of interspecific coalescences provides a high probability of consistency with the actual species history. Under the expectation that there is a correspondence between the number of nucleotide differences between sequences and the order of interspecific coalescences, the most similar sequences between species will represent the shallowest coalescence (Takahata and Nei, 1985). Therefore to infer the species tree, we used a cluster algorithm that first grouped species together that contain gene sequences with the fewest differences (i.e., the two species with the most similar sequences). Because multiple hits are unlikely for times as shallow and rates as low as these, we used as our measure of sequence divergence a simple uncorrected distance. The distance between two clades is similarly defined, and thus the method is equivalent to a single linkage cluster analysis (Sneath and Sokal, 1973:216). For multiple loci, the distance between two clusters is the average of the distances based on individual loci. Mesquite's Cluster Analysis facility was used for this, ignoring ties (i.e. resulting in only a single tree per cluster analysis).

Accuracy Assessment

The inferred species trees were compared to their respective original trees by calculating accuracy in both an unrooted and a rooted sense. The unrooted measure was the number of partitions of the species in common between the inferred and original trees. A partition represents a branch, which implicitly divides the taxa into two groups (one of which may be paraphyletic). The maximum conceivable number of shared partitions for trees of 8 species, which occurs when the trees are identical, is 5. The rooted measure was the number of clades in common between the inferred and original trees. The maximum conceivable number of shared clades is 6. So that both shared clades and shared partitions can be compared on the same scale, the results are reported as the proportion of clades or partitions shared out of the totals of 6 or 5, respectively.

Results and Discussion

In our simulations, average sequence divergences and incompleteness of lineage sorting varied as expected depending on the divergence times in the species tree. Percent sequence divergences averaged about 1% and 4% for the total tree depths of 1 N_e and 10 N_e, respectively (Table 1). These divergences are similar to those encountered in species-level studies (Arbogast, et al., 2002), especially those involving Pleistocene divergences (e.g., Knowles, 2000, 2001; Masta and Maddison, 2002; Hewitt, 2004). Discordance between gene trees and species trees ranged from mild to severe, with up to an average of 114 deep coalescences in the worst case (Table 1). The shallow species trees exhibited about four times the average amount of incomplete lineage sorting compared to the deeper trees for any given number of individuals sequenced per locus (Table 1). Figure 1 is a representative instance from the case of 9 individuals per species, with a shallow species tree. The gene tree is about 300,000 generations deep, which is within the range expected, because the gene tree cannot coalesce shallower than 100,000 generation, the root of the species tree. Such discordances are expected in typical studies, both between any single gene tree and the species history, and between different loci (Takahata, 1989; Rosenberg, 2003; Knowles and Maddison, 2002). Given such discordance, our simulated data should well reflect the challenges faced in reconstructing phylogeny near the species level.

Evidence of Phylogenetic Signal

When species are recently diverged as in Figure 1, many coalescences extend deep below divergence points. Although in such cases we might have expected little chance of correctly inferring the species tree, our results suggest that considerable signal remains in the data. The average accuracy of phylogeny reconstruction across the 500 replicate species trees ranged from about 0.26 to 0.89 unrooted or 0.17 to 0.88 rooted (Table 2), with accuracies under most conditions between 0.4 and 0.8.

Accuracy was generally very similar regardless of whether measured in an unrooted way (via shared partitions) or a rooted way (via shared clades) for both methods of reconstruction (Table 2; Fig. 3), with two notable exceptions. First, although the Shallowest Divergences method performed about as well in recovering either the rooted or unrooted tree, the Deep Coalescence method appeared to have considerable trouble choosing the root of the species tree. Its unrooted accuracy was about the same as that of the shallowest divergences method, but its rooted accuracy was especially poor when only a single individual was sampled per species. This suggests it was obtaining the correct unrooted form of the tree, but was rooting incorrectly. That it had difficulty rooting is perhaps not surprising as no outgroup was given to the reconstruction method. In empirical studies we might not even be tempted to choose a species tree root without an outgroup, and so perhaps the difficulties of deep coalescence should not be held against it. From that viewpoint, however, we can be pleased at the success the shallowest divergences method in determining the root.

Second, the Shallowest Divergences method showed a peculiar decline in accuracy from 9 individuals per species sampled to 27 individuals with a single locus. We have confirmed this result by repeating the simulations. We lack a firm explanation for this pattern, although T. Collins (personal communication) has suggested that the large number of individuals may give more opportunity for some to be seriously misplaced given rate variation among sites and low divergence.

TABLE 2. Accuracy of phylogenetic inference over the 500 simulated species trees with a total tree depth of (a) 100,000 (1 N_e) and (b) 1,000,000 (10 N_e) generations for varying numbers of individuals and loci sampled per species. Accuracy is shown as two numbers, "unrooted accuracy/rooted accuracy" for which unrooted accuracy is the average proportion of partitions correct (those in the inferred tree matching the true tree) out of five total partitions and rooted accuracy is the average proportion of clades correct out of six total clades.

	1 locus	3 loci	9 loci	27 loci
a. Total tree depth of 1 N_e				
1 individual				
Deep Coalescences	0.26/0.17	0.34/0.21	0.42/0.26	0.63/0.39
Shallowest Divergence	0.27/0.24	0.33/0.31	0.43/0.43	0.61/0.62
3 individuals				
Deep Coalescences	0.47/0.40	0.58/0.52	0.65/0.63	
Shallowest Divergence	0.53/0.50	0.64/0.62	0.73/0.72	
9 individuals				
Deep Coalescences	0.59/0.53	0.65/0.63		
Shallowest Divergence	0.60/0.58	0.74/0.72		
27 individuals				
Deep Coalescences	0.64/0.58			
Shallowest Divergence	0.56/0.55			
b. Total tree depth of 10 N_e				
1 individual				
Deep Coalescences	0.76/0.49	0.79/0.51	0.86/0.53	0.89/0.54[a]
Shallowest Divergence	0.73/0.73	0.79/0.75	0.85/0.86	0.89/0.88[b]
3 individuals				
Deep Coalescences	0.79/0.57	0.82/0.62	0.87/0.65	
Shallowest Divergence	0.78/0.76	0.84/0.82	0.88/0.88	
9 individuals				
Deep Coalescences	0.80/0.60	0.85/0.65		
Shallowest Divergence	0.79/0.77	0.86/0.85		
27 individuals				
Deep Coalescences	0.82/0.61			
Shallowest Divergence	0.84/0.82			

[a] Average is based on 499 simulated species trees due to computing error during simulations.

[b] Average is based on 390 simulated species trees due to computing error during simulations.

With the above exceptions, results from the two species-tree reconstruction methods and accuracy measures are in close agreement throughout (Table 2). Therefore Figures 4 and 5 show only the results from the Minimize Deep Coalescences method for clarity of graphics. Figure 4 illustrates the distributions of shared partitions among the 500 replicates for different numbers of loci and individuals per species. Figure 5 illustrates average accuracies for both the 1 N_e and 10 N_e species trees.

Average accuracies of 0.6 or more can be considered reasonably successful, given that the shared partition and shared clade measures are sensitive to minor changes in tree structure. The misplacement of a single species can reduce shared partitions to 0. An average accuracy of 0.6 is approximately equivalent to a single terminal taxon being out of place: moving a single terminal taxon randomly to a different branch of the tree results in an average reduction from 5 to 3 shared partitions (as determined by the "Random Branch Moves" utility of Mesquite operating on Yule process simulated trees).

Species trees were reasonably well recovered even with only a single locus. For instance, with only 9 individuals sequenced per species, even in the difficult case of depth = 1 N_e, 3 partitions on average were correctly reconstructed out of a total of 5 (Table 2).

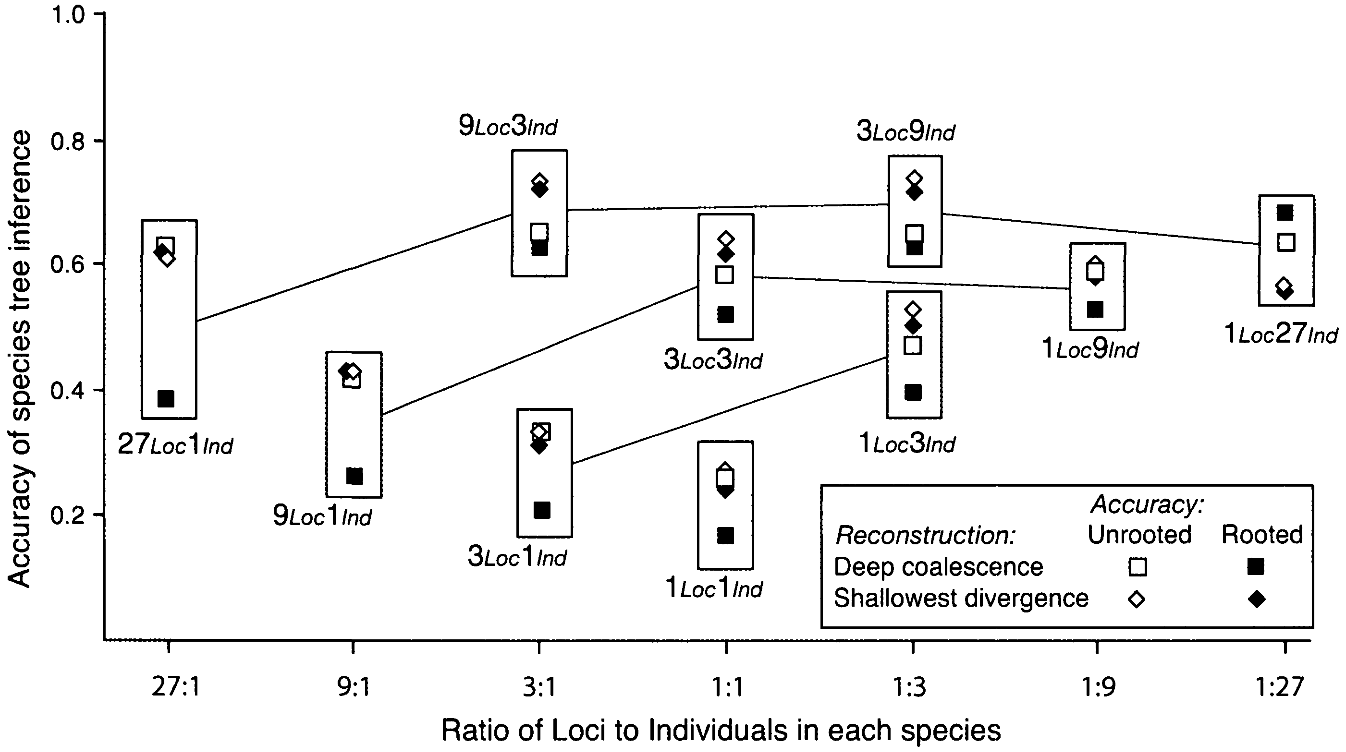

FIGURE 3. Species tree accuracy with different methods of reconstruction and different accuracy measures, for species trees of depth 1 N_e. Lines join points with equal numbers of total sequences, with numbers of loci (*Loc*) and sampled individuals (*Ind*) indicated. Accuracy measured as average accuracy between true and inferred tree over the 500 simulated species trees. Reconstruction methods and accuracy measures yield similar results except for the Minimize Deep Coalescences for rooted accuracy, which is notably lower.

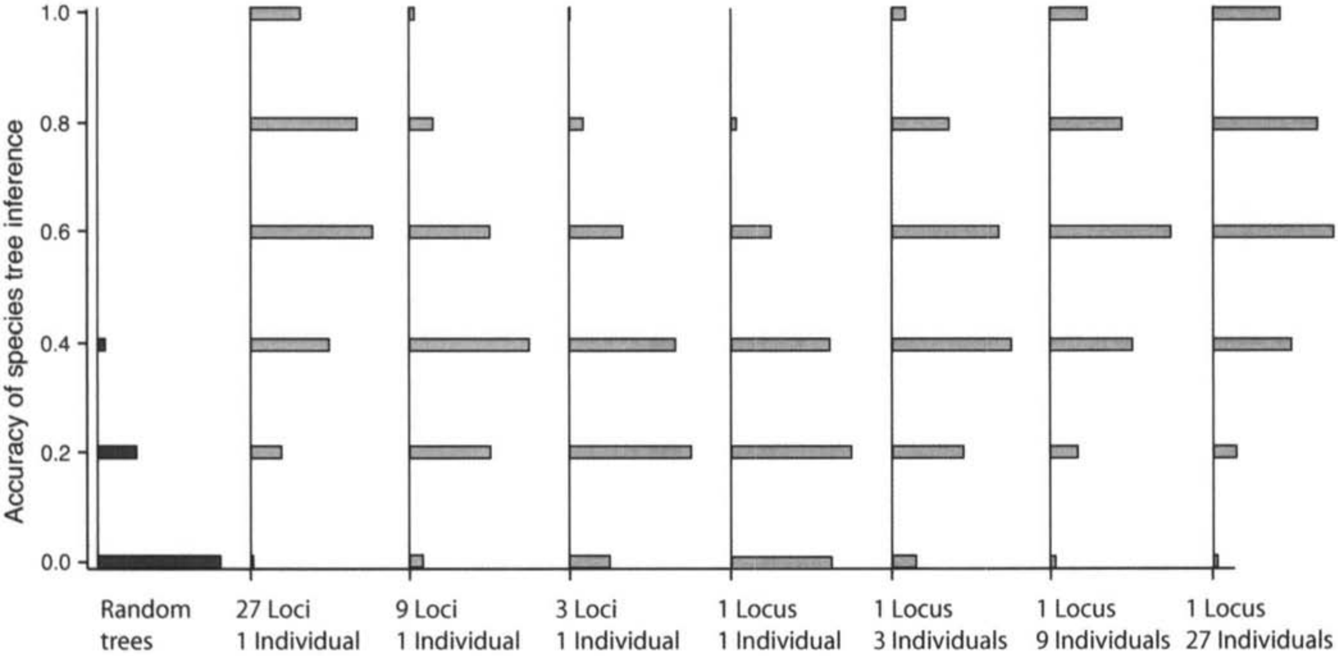

FIGURE 4. Accuracy of phylogeny reconstruction among 500 replicate species trees with depth 1 N_e. Except for the leftmost chart, each chart represents a different sampling intensity. At center is the lowest intensity, one locus and one individual per species. From the center toward the right the charts represent increasing numbers of individuals sampled. From center toward the left, the charts represent increasing numbers of loci sampled. The reconstruction method is Minimize Deep Coalescences; the accuracy measure is shared partitions. The leftmost chart shows for comparison the accuracy of a randomly chosen tree; the randomly chosen trees are generated by the Yule Process and each is compared against a different one of the 500 simulated species trees used in the study. This figure can be directly compared to Figures 3 and 5, e.g., the chart for 27 loci 1 individual represents the full distribution of results for the 27 Loc 1 Ind point on Figure 5.

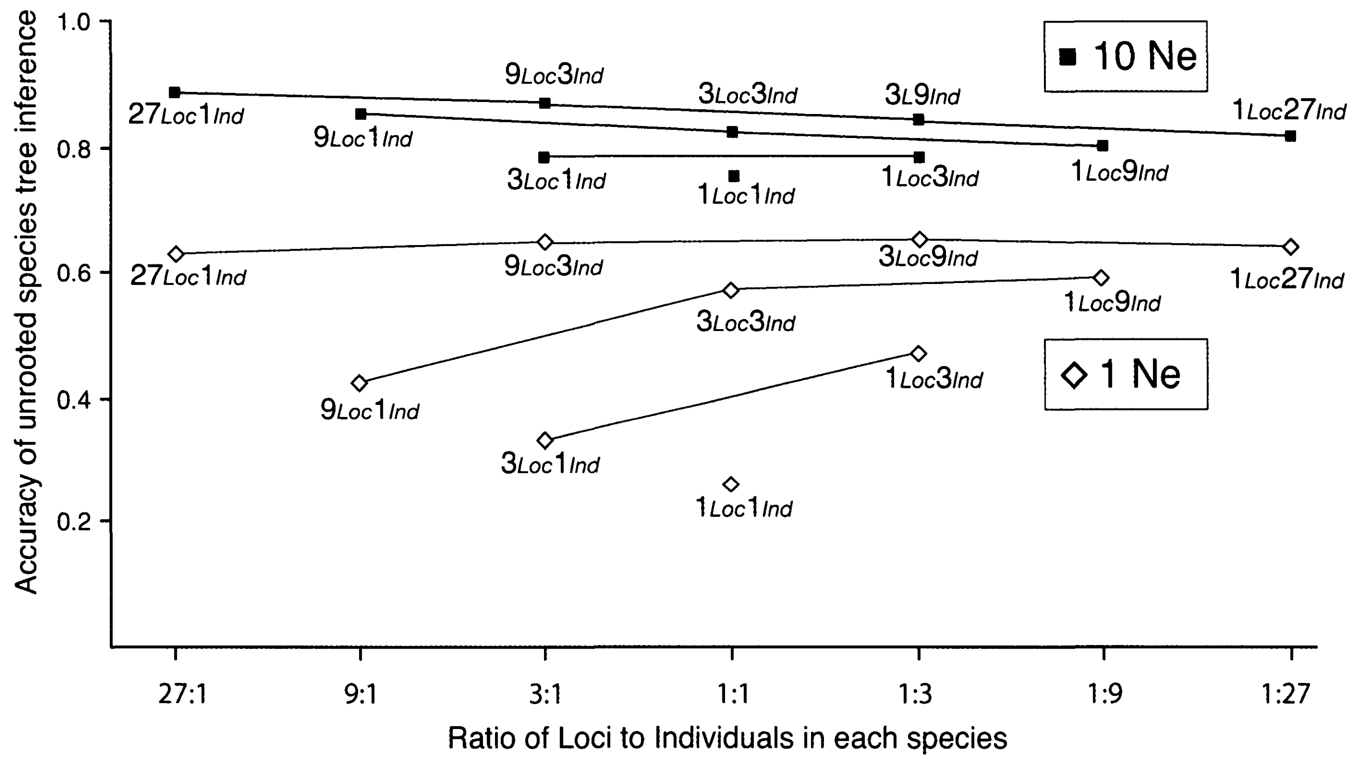

FIGURE 5. Tradeoff in species tree accuracy between sampling more loci versus individuals for a fixed sequencing effort. Each line joins points with equal numbers of total sequences, with numbers of loci (*Loc*) and sampled individuals (*Ind*) indicated. Accuracy measured as average accuracy between true and inferred tree over the 500 simulated species trees. Trees inferred by Minimize Deep Coalescences; accuracy measured by shared partitions. Note that with trees of depth 1 N_e (high levels of incomplete lineage sorting) emphasizing individuals over loci is advantageous with nine or three sequences per species, while with trees of depth 10 N_e, emphasizing loci is slightly better in general.

Phylogenetic signal is apparent (although weak) even in the most difficult condition of a species tree depth = 1 N_e (i.e., considerable incomplete lineage sorting, Table 1) and the least amount of data (1 locus and 1 sequence per species). Under these conditions, the most common outcomes are 0, 1, or 2 shared partitions with the "true" tree (Fig. 4). Although this may appear low, the distribution is much more favorable than expected by chance. The curve marked "random trees" in Figure 4 shows the distribution for the accuracy measure (shared partitions) in comparing two randomly simulated trees. Most random trees have no shared partitions with a target tree.

Accuracy varied as expected with differing depths of species divergences and amounts of data (Table 2). The lowest accuracy occurred in the most challenging scenario of a recent divergence (i.e., species tree depth = 1 N_e) with the least amount of data (i.e., only 1 locus and 1 individual sequenced per species); the highest accuracy with the least challenging case of a deeper divergence with the largest amount of data (i.e., 10 N_e and 27 sequences per species).

Tradeoff of More Loci versus Individuals

To examine more closely the effects of sampling design on phylogenetic accuracy, we compared the average phylogenetic accuracy among different allocations of sampling effort between increasing loci versus increasing number of individuals. The allocation of sampling effort and its impact on the accuracy of the phylogenetic estimates differed between the recent and deeper species divergence (Fig. 5). With deeper species trees (depth = 10 N_e) shifts between sampling loci as opposed to individuals had little effect for a given sampling effort. Thus, for example, with a total of 9 sequences, similar accuracies are obtained whether they are obtained as 9 loci × 1 individual, 3 loci × 3 individuals, or 1 locus × 9 individuals. All sampling regimes give accurate trees with little effect of sampling strategy. There appears to be a slight advantage to sampling more loci instead of more individuals, but the major factor influencing phylogenetic accuracy is the total sampling effort (Table 2).

With shallower species trees (depth = 1 N_e), sampling more individuals gives better results than sampling more loci (Fig. 5) for most measurements. Thus, sampling 9 loci × 1 individual gives an average accuracy of about 0.43, whereas sampling 1 locus × 9 individuals gives an accuracy of about 0.59. This result is perhaps not surprising. With shallow species trees, the many gene lineages that independently reach as deep as the species divergence (because of failure to sort lineages) can each provide independent clues to species relationships: each coalescence with a sister species' genes provides extra evidence that the species are sisters. By sampling more individuals per species, there can be more such interspecific coalescence between sister species to provide evidence for their relationship (see also Takahata, 1989). This advantage appears to outweigh the advantage of extra loci. On the other hand, with deep species trees, adding extra sampled individuals within a species does little good, because the ancestors of most of those gene copies coalesce to a single gene lineage before the species divergence, and thus they do not give independent evidence of species relationships. Extra loci can supply independent evidence of deeper relationships regardless of coalescence within species lineages. (Note that the shape of the curves suggests that N loci × 1 individual may be better than 1 locus × N individuals for N's larger than we examined, Figure 5; however, the recommendation of increasing the number of individuals sampled, as opposed to loci, appears to apply in general given a reasonable range of loci based on practical limitations.)

The pattern of greater accuracy by increasing individuals rather than loci for shallow species trees may not be a general pattern, however. The curve of increase of accuracy by adding loci is rising rapidly at 27 loci (Fig. 5), whereas the curve for adding individuals appears to be levelling off at 27 individuals. Thus, the accuracy for, say, 54 loci, may be greater than that for 54 individuals.

Implications for Accurate Phylogenetic Inference

Our results lead us to two principal conclusions: that sufficient signal to reconstruct species trees remains even in the face of considerable incomplete lineage sorting, and that optimal sampling strategy (multiple loci versus multiple individuals) depends on recency of divergences. These conclusions were reached with both of our species-tree reconstruction methods, which suggests that they might be expected to hold even with more sophisticated methods that could be imagined, such as likelihood methods that model both nucleotide substitution and lineage sorting as stochastic processes.

A central position of this article is that progress in developing better methods of reconstructing recently diverged species trees will depend on considering explicitly the population genetics processes within species lineages. Although crudely, our reconstruction methods did consider these processes. Can we claim that we achieved better reconstructions than we would have achieved had we used methods that did not consider the process of lineage sorting? The difficulty in making this comparison directly is that it is not clear how one would reconstruct a species tree like that of Figure 1 *without* considering lineage sorting. The 9 individuals' gene copies within each of the species of Figure 1 do not form monophyletic groups on the gene tree. If one's data were sequences from each of these gene copies, how would one derive a species tree unless one implicitly allowed and accounted for this lack of monophyly? We are not necessarily recommending our particular species tree reconstruction methods, but unless methods like them are developed and become widely used, it would seem that biologists facing a case like Figure 1 will be able to do no more than take a nonquantitative guess at the species tree.

A single locus's genealogical history is subject to many stochastic effects, and hence data from multiple independent loci are important to offer independent information contributing to inference of a species tree. Nonetheless, our results suggest it is not universally the case that increased sequencing effort should be allocated to more loci instead of more individuals. For recent divergences, sequencing more individuals yielded greater improvement than sequencing more loci, at least up to 27 loci or individuals. However, three caveats should be given to any recommendation to sequence more individuals. First, the advantage of sampling individuals over loci may not persist in samples of more than 27 total sequences per species, the largest sampling effort we studied. Second, the actual cost of sequencing is not proportional simply to the total sequence length; it depends also on the relative costs of specimen collection, DNA extraction, and primer design. Third, our simulations assume that all loci are evolving under the same neutral model of evolution. If some loci are under selection or linked to loci that are, then their behavior may not be clearly informative about the species tree. Obtaining several loci will give a better sample of varying models of evolution, and hence may give a clearer picture of the species tree.

Our simulations considered a fairly simple case of a species tree evolving by a simple Yule process. Two more complex cases might be expected to pose greater difficulties for recovering the species tree and should be investigated as new population-aware species tree reconstruction methods are developed. The first involves gene flow among species or populations. Gene flow may seriously degrade the accuracy of some inference methods, even when levels of gene flow are low enough that the species phylogeny can still be considered fundamentally a branching process. We would expect, for example, that the shallowest divergence method would be particularly sensitive to misinterpreting a recently introgressed gene copy as solid evidence for species relationship. The second difficult case is that of an old but rapid set of divergences, in which a rapid burst of speciation long in the past yielded many short branches deep in the species tree. To recover these ancient but short branches we cannot rely on sequencing many individuals, because their gene copies will have coalesced to a single ancestral copy well before reaching those short branches. Thus, adding more loci will be the only route to increasing resolution. Even that will have limited success, however, given that genes with a high mutation rate would be needed to resolve these short branches, but those very markers would suffer from saturation when species divergence is old.

A worthwhile goal of future studies could be to explore thoroughly the parameter space. One factor likely to affect the results strongly is the overall mutation rate, here represented as the scaling factor. Perhaps the most common problem encountered in empirical studies at this level would be insufficient variation (e.g., mutation rates too low) degrading accuracy. As more sophisticated methods are developed to reconstruct species trees by considering incomplete lineage sorting, it would be valuable to assess their accuracy when variation is low.

Incomplete lineage sorting poses a daunting challenge to our efforts to reconstruct the phylogenetic relationships of recently derived species. Yet, even in cases with widespread incomplete lineage sorting, a significant historical signal persists. By taking into account the genetic process generating incomplete lineage sorting, phylogenies can be accurately inferred from gene trees. Our hope is that phylogenetic methodologists take up the challenge to develop reconstruction techniques that consider both nucleotide substitution and the processes of population genetics.

ACKNOWLEDGMENTS

Roderic Page, Tim Collins, Peter Beerli, Noah Rosenberg, and an anonymous reviewer provided helpful comments and suggestions. Development of the Mesquite evolutionary analysis system used in this work was supported in part by a David and Lucile Packard Fellowship to WPM. Simulations were run with equipment and software purchased on grants from CFI Canada and NSERC Canada to WPM.

REFERENCES

Arbogast, B., S. V. Edwards, J. Wakeley, P. Beerli, and J. B. Slowinski. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic time scales. Annu. Rev. Ecol. Syst. 33:707–740.

Arnold, E. N. 1981. Estimating phylogenies at low taxonomic levels. Z. Zool. Syst. Evol. Forsch. 19:1–35.

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.

Degnan, J. H., and L. A. Salter. 2005. Gene tree distributions under the coalescent process. Evolution 59:24–37.

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

Edwards, A. W. F., and L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. Pages 67–76 *in* Phenetic and phylogenetic classification (V. H. Heywood and J. McNeill, eds.). Systematics Association Publ. No. 6, London.

Edwards, S. V., and P. Beerli. 2000. Perspective: Gene divergence, population divergence, and the variance in coalescence time in phylogeography studies. Evolution 54:1839–1854.

Farris, J. S. 1978. Inferring phylogenetic trees from chromosome inversion data. Syst. Zool. 27:275–284.

Felsenstein, J. 1979. Alternative methods of phylogenetic inference and their interrelationship. Syst. Zool. 28:49–62.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Hennig, W. 1966. Phylogenetic systematics. University of Illinois Press, Urbana.

Hewitt, G. M. 2004. Genetic consequences of climatic oscillations in the Quaternary. Phil. Trans. R. Soc. Lond. B 359:183–195.

Hey, J., and Nielsen, R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence times, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167:747–760.

Hoelzer, G. A. 1997. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees revisited. Evolution 51:622–626.

Hudson, R. 1990. Gene genealogies and the coalescent process. Oxford surveys in Evol. Biol. 7:1–44.

Kingman, J. F. C. 1982. The coalescent. Stochastic Process. Appl. 13:235–248.

Knowles, L. L. 2000. Tests of Pleistocene speciation in montane grasshoppers from the sky islands of western North America (genus *Melanoplus*). Evolution 54:1337–1348.

Knowles, L. L. 2001. Did the Pleistocene glaciations promote divergence? Tests of explicit refugial models in montane grasshoppers. Mol. Ecol. 10:691–701.

Knowles, L. L. and W. P. Maddison. 2002. Statistical phylogeography. Mol. Ecol. 11:2623–2635.

Lake, J. A., and M. C. Rivera. 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: Conditioned reconstruction. Mol. Biol. Evol. 21:681–690.

Machado, C. A., R. M. Kliman, J. A. Markert, and J. Hey. 2002. Inferring the history of speciation from multilocus DNA sequence data: The case of *Drosophila pseudoobscura* and close relatives. Mol. Biol. Evol. 19:472–488.

Maddison, D. R., and W. P. Maddison. 2004a. Genesis: Models of character evolution. A package of modules for Mesquite. Version 1.01. http://mesquiteproject.org

Maddison, D. R., and W. P. Maddison. 2004b. Batch Architect. A package of modules for Mesquite. Version 1.01. http://mesquiteproject.org

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

Maddison, W. P., and D.R. Maddison. 2004c. Mesquite: A modular system for evolutionary analysis. Version 1.01. http://mesquiteproject.org

Masta, S. E., and W. P. Maddison. 2002. Sexual selection driving diversification in jumping spiders. Proc. Nat. Acad. Sci. USA 99:4442–4447.

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

Pluzhnikov, A., and P. Donnelly. 1996. Optimal sequencing strategies for sureying molecular genetic diversity. Genetics 144:1247–1262.

Rannala, B., and Y. Yang. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.

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

Rosenberg, N. A. 2003. The shapes of neutral gene genealogies in two species: Probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. Evolution 61:225–247.

Sanderson, M., and H. B. Shafer. 2002. Troubleshooting molecular phylogenetic analyses. Annu. Rev. Ecol. Syst. 33:49–72.

Sneath, P. H. A., and R. R. Sokal. 1973. Numerical taxonomy. W. H. Freeman, San Francisco.

Sota, T., and A. P. Vogler. 2003. Reconstructing species phylogeny of the carabid beetles *Ohomopterus* using multiple nuclear DNA sequences: Heterogeneous information content and the performance of simultaneous analyses. Mol. Phyl. Evol. 26:139–154.

Swofford, D. L. 2002. PAUP* Phylogenetic analysis using parsimony (*and other methods). Version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

Takahata, N. 1989. Gene genealogy in 3 related populations—consistency probability between gene and population trees. Genetics 122:957–966.

Takahata, N., and M. Nei. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325–344.

Takahata, N., and Y. Satta. 2002. Pre-speciation coalescence and the effective size of ancestral populations. Pages 52–71, *in* Modern developments in theoretical population genetics (M. Slatkin and M. Veuille, eds.). Oxford University Press, Oxford, UK.

Throckmorton, L. H. 1965. Similarity versus relationship in *Drosophila*. Syst. Zool. 14:221–236.

Wakeley, J. 2003. Inferences about the structure and history of populations: Coalescents and intraspecific phylogeography. Pages 193–215 *in* The evolution of population biology (R. Singh and M. Uyenoyama, eds.). Cambridge University Press, Cambridge, UK.

Wall, J. D. 2003. Estimating ancestral population sizes and divergence times. Genetics 163:395–404.

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

First submitted 25 October 2004; reviews returned 21 January 2005; final acceptance 25 May 2005
Associate Editor: Tim Collins