Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007

*Syst. Biol.* 56(2):302–320, 2007
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701324225

# Distinguishing Terminal Monophyletic Groups from Reticulate Taxa: Performance of Phenetic, Tree-Based, and Network Procedures

PATRICK A. REEVES AND CHRISTOPHER M. RICHARDS

*United States Department of Agriculture, Agricultural Research Service, National Center for Genetic Resources Preservation, 1111 South Mason Street, Fort Collins,Colorado 80521, USA; E-mail: crichard@lamar.colostate.edu (C.M.R.)*

*Abstract.*— Hybridization is a well-documented, natural phenomenon that is common at low taxonomic levels in the higher plants and other groups. In spite of the obvious potential for gene flow via hybridization to cause reticulation in an evolutionary tree, analytical methods based on a strictly bifurcating model of evolution have frequently been applied to data sets containing taxa known to hybridize in nature. Using simulated data, we evaluated the relative performance of phenetic, tree-based, and network approaches for distinguishing between taxa with known reticulate history and taxa that were true terminal monophyletic groups. In all methods examined, type I error (the erroneous rejection of the null hypothesis that a taxon of interest is not monophyletic) was likely during the early stages of introgressive hybridization. We used the gradual erosion of type I error with continued gene flow as a metric for assessing relative performance. Bifurcating tree-based methods performed poorly, with highly supported, incorrect topologies appearing during some phases of the simulation. Based on our model, we estimate that many thousands of gene flow events may be required in natural systems before reticulate taxa will be reliably detected using tree-based methods of phylogeny reconstruction. We conclude that the use of standard bifurcating tree-based methods to identify terminal monophyletic groups for the purposes of defining or delimiting phylogenetic species, or for prioritizing populations for conservation purposes, is difficult to justify when gene flow between sampled taxa is possible. As an alternative, we explored the use of two network methods. Minimum spanning networks performed worse than most tree-based methods and did not yield topologies that were easily interpretable as phylogenies. The performance of NeighborNet was comparable to parsimony bootstrap analysis. NeighborNet and reverse successive weighting were capable of identifying an ephemeral signature of reticulate evolution during the early stages of introgression by revealing conflicting phylogenetic signal. However, when gene flow was topologically complex, the conflicting phylogenetic signal revealed by these methods resulted in a high probability of type II error (inferring that a monophyletic taxon has a reticulate history). Lastly, we present a novel application of an existing nonparametric clustering procedure that, when used against a density landscape derived from principal coordinate data, showed superior performance to the tree-based and network procedures tested. [Hybridization; introgression; parsimony; phylogenetic network; principal coordinate analysis; reticulate evolution.]

Tremendous progress towards understanding higher level seed plant relationships has been made through the application of phylogenetic analysis (Hennig, 1966; Chase et al., 1993; Mathews and Donoghue, 1999; Qiu et al., 1999; Bowe et al., 2000; Chaw et al., 2000; Graham and Olmstead, 2000). In recent years, there has been a push to press the lower bounds of phylogenetic analysis within the seed plants (Schaal and Leverich, 2001; Stuessy et al., 2003; Crawford and Mort, 2004) to test population-level evolutionary hypotheses and to reveal terminal monophyletic groups (i.e., the clade defined by the node that pinpoints the boundary between phylogeny and tokogeny). Accurate identification of terminal monophyletic groups is important because it is a prerequisite for delimiting and testing species boundaries using a variety of analytical methods (reviewed in Sites and Marshall, 2004) and impinges on our ability to define species using some versions of the phylogenetic species concept (de Queiroz and Donoghue, 1988; Baum and Shaw, 1995). Furthermore, the ability to accurately reconstruct historical relationships below the species level using phylogenetic methods is the linchpin of the field of intraspecific phylogeography (Avise et al., 1987) and impacts its increasing application to issues in conservation biology (Goldstein et al., 2000; DeSalle and Amato, 2004; Russello and Amato, 2004)

For the purposes of this study, we define "taxa" as any named assemblage of organisms, regardless of ancestry. We extend the use of the term below the species level to include populations, because "populations are the least inclusive units appropriate for use as terminal taxa" (de Queiroz and Donoghue, 1988). We use the term "monophyletic," as expanded by de Queiroz and Donoghue (1988), to indicate a group of organisms that has descended from a single common ancestor. We believe this has become the common usage, in spite of its purported deviation from Hennig's (1966) original intent (as argued by Wheeler and Nixon, 1990; Goldstein and DeSalle, 2000; and others). "Hybridization" is herein defined as reproduction that involves two individuals from different taxa.

Although its ubiquity has been debated (Ellstrand et al., 1996), it is known that hybridization between many plant species is possible (Grant, 1971; Knobloch, 1972) and that interspecific gene flow occurs in nature (Stebbins, 1959; Arnold, 1997). Below the species level, barriers to gene flow are generally believed to be weak; thus, the potential for movement of alleles between taxa such as populations is obvious. One consequence of gene flow between taxa is the production of reticulate phylogenetic relationships.

The application of methods that produce hierarchical and bifurcating trees ("tree-based methods") to data sets derived from reticulate evolutionary processes has been criticized (Sneath, 1975; Bremer and Wanntorp, 1979; McDade, 1990; Nixon and Wheeler, 1990; Davis and Nixon, 1992; Doyle, 1995; Legendre, 2000, and citations therein; Posada and Crandall, 2001). A variety of alternative

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



procedures (collectively termed "network" phylogenetic methods) have been developed that claim to be appropriate for analyzing such data (Hein, 1990; Templeton et al., 1992; Excoffier and Smouse, 1994; Fitch, 1997; Bandelt et al., 2000; Xu, 2000; Legendre and Makarenkov, 2002). Unlike tree-based methods, network methods can produce graphs that contain cycles (or "loops"). The cycles present in these graphs are a result of the algorithm used to summarize the set of acceptable alternative solutions for a single data set. Cycles within a network can arise as a consequence of conflicting phylogenetic signal in the underlying data set. Conflicting signal may be due to reticulate processes such as hybridization or recombination but may also be caused by routine homoplasy, present in virtually all data sets. It is difficult to distinguish true reticulate signal from routine homoplasy using networks.

In spite of their promise, most network methods have not been evaluated for their ability to reveal correct relationships when applied to data sets with known reticulate history. Cassens et al. (2005) examined the performance of some network methods but considered only nonreticulate simulated data sets. Moreover, tree-based methods continue to be routinely used for the analysis of potentially reticulate data sets despite numerous, clearly articulated concerns. It is therefore important (1) to determine the degree and manner in which an arguably illegitimate application of tree-based analysis may bias conclusions, and (2) to test whether alternative methods of analysis offer a satisfactory solution.

Past studies have explored the consequences of hybridization on tree-based, cladistic analysis (Bremer and Wanntorp, 1979; Nelson and Platnick, 1980; Funk, 1985). Detailed empirical studies by McDade (1990, 1992, 1997) showed that, when considering morphological characters, artificial F1 hybrids were often (65% of the time) resolved at a position in the most-parsimonious tree that was basal to the clade containing the most derived parent. The remainder of the time, the position of the hybrid in the tree and its effect on overall tree topology were less predictable. Although essential to our present understanding of the phenomenon, these studies provide incomplete guidance if the goal is to distinguish terminal monophyletic lineages from taxa with a reticulate history because (1) they relied on morphological characters, which are prone to nonindependence, unpredictable patterns of inheritance in hybrids (Rieseberg and Ellstrand, 1993), and other features that make them undesirable for phylogeny reconstruction (Scotland et al., 2003); and (2) they only consider the fate of F1 hybrid individuals, as opposed to hybrid (or reticulate) taxa (which include backcrossed individuals), in the resulting tree.

In cases where postzygotic isolating mechanisms operate to prevent gene flow between taxa (e.g., in some cases of interspecific hybridization) F1 hybrids may outnumber backcrossed individuals. However, complex backcrossed individuals will predominate over F1 hybrids when biological barriers to reproduction are weak (e.g., in some cases of interspecific hybridization and many cases of intraspecific hybridization). This is because the majority of potential mates for a rare F1 hybrid will be members of the local population and because selection may favor resident, as opposed to migrant, genotypes (Stebbins, 1950). Because F1 hybrids will often be rare relative to back-crossed individuals, understanding the effect of F1 hybrids on phylogeny reconstruction is only one component of verifying the applicability of tree-based methods to reticulate data.

Two studies have considered the effect of including simulated recombinant DNA sequences in phylogenetic analyses (Schierup and Hein, 2000; Posada and Crandall, 2002). These studies are relevant because a single recombinant DNA sequence mimics the chimeric assemblage of character states that would be expected if many unlinked loci were scored in an individual with reticulate ancestry. Schierup and Hein (2000) found that the inclusion of recombinant sequences in distance- and maximum likelihood (ML)-based phylogenetic analyses caused a predictable change in tree shape, with longer terminal branches and shorter internal branches than were found when recombinant sequences were not present. They did not, however, consider the effect of recombinant sequences on the accuracy of the reconstructed phylogeny.

Posada and Crandall (2002) examined the accuracy of tree-based analytical methods when recombinant sequences with characteristics of F1 hybrids, as well as back-crossed individuals, were analyzed. They showed that the impact of including recombinant sequences in an analysis is dependent on the topological relationship between "parent" sequences and the proportion of the recombinant sequence contributed by each parent. Although conceived with a different purpose in mind, the study may be interpreted to suggest that the inclusion of taxa with characteristics of F1 hybrids (a recombinant sequence with 50% of the sequence belonging to each parent) in phylogenetic analyses is likely to result in erroneous tree reconstruction, with frequent topological rearrangements among nonhybrid taxa relative to the true tree. Sequences with characteristics of back-crossed individuals were shown to be less likely to impact the accuracy of phylogeny reconstruction. However, the conclusions of Posada and Crandall (2002) are not easily extended to natural populations because the fate of only a single recombinant sequence within a larger tree of non-recombinant sequences was considered.

The effect of reticulate taxa (as opposed to single individuals or sequences) on phylogeny reconstruction is not well understood. Reticulate taxa are likely to pose special problems for empirical studies at the species level and below because the distribution of character states among individuals of such taxa is not readily predictable (unlike with F1 hybrids). In the simplified case of unidirectional gene flow between two taxa, hybridization may be followed by introgression of character states from a donor taxon into a recipient taxon, resulting in conflicting phylogenetic signal. Depending on random genetic

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



drift and the strength of selection, only a small number of the character states from the donor taxon that are found in the F1 hybrid individual would be expected to become fixed in the recipient taxon. Accordingly, the magnitude of conflicting signal present in individuals of a recipient taxon will be less than that present in an F1 hybrid individual, potentially making taxa with reticulate histories difficult to identify, in spite of significant gene flow. Furthermore, during the intermediate period between the hybridization event and fixation of the introgressed character states, variation in the proportion of character states that are traceable to the donor taxon will be observed among individuals in the recipient taxon. Variation in conflicting signal levels among individuals within a single taxon is potentially problematic for methods that rely on conflicting signal to indicate historical reticulation. The presence of interindividual variation in conflicting signal levels suggests that many individuals per taxon must be included in an analysis to avoid conclusions that are an artifact of inadequate sampling.

In this study, we evaluate the ability of six analytical procedures to distinguish terminal monophyletic groups from reticulate taxa in data sets simulated with a predefined pattern of historical reticulation. We employ a population-level model for simulating multilocus data that allows multiple, complex back-crossed individuals to be sampled during a continuous process of introgressive hybridization. We include an assessment of tree-based methods as well as two phenetic methods that do not assume data fit a tree-like structure. In addition, we evaluate the performance of two network methods and reverse successive weighting (RSW) (Trueman, 1998), a tree-based procedure intended to identify conflicting phylogenetic signal in data sets derived from reticulate historical processes.

## Materials and Methods

### *Generation of Simulated Data Sets*

An unrooted five-taxon tree of the form ((((A,B),C),D),E) was used as the starting point (Fig. 1a). A five-taxon tree was used because it is the simplest tree for which the effect of gene flow on topological relationships between taxa other than those directly involved in gene exchange can be evaluated. Four topologically distinct alternatives for unidirectional gene flow between taxa were considered. Sim1 modeled hybridization between sister taxa and Sim2 and Sim3 explored hybridization between increasingly divergent taxa, both in terms of genetic distance and the topological relationship between them. Sim4, in which gene flow occurred between the same taxa as Sim2, but in the opposite direction, was used to examine whether the direction of gene flow within a tree affected the ability to infer a reticulate history for the recipient taxon.

Five-taxon starting trees were created by simulating four 1000-character data sets using Seq-Gen (Rambaut and Grassly, 1997). The number of substitutions per site was set to 0.1 for all branches. The substitution model used was equivalent to the Jukes-Cantor model (Jukes and Cantor, 1969), but with only two character states allowed so the resulting matrices contained binary data. Although this model was developed to describe DNA sequence evolution, the characters were not subsequently treated as a linked DNA sequence but rather as haploid, unlinked loci, subject to recombination and genetic drift during simulated reproduction. The population size (N) of each taxon was set to 10 by replicating the starting haplotypes derived from Seq-Gen.

Hybridization and genetic drift were simulated using a program written by the authors (available at http://lamar.colostate.edu/~reevesp/Hybridize.html). Reproduction followed a Wright-Fisher model with constant population size, fully random mating (including selfing), and nonoverlapping generations. The simulation proceeded as follows (Fig. 1b):

1. Randomly choose one haplotype from recipient taxon (Parent 1) to be hybridized with the donor taxon (Parent 2).
2. Produce a progeny haplotype by randomly selecting, for each character, the character state present in either Parent 1 or Parent 2.
3. Produce nine additional progeny by choosing both parents at random from within the recipient taxon. Assemble progeny haplotypes as in step 2.
4. Allow 10 generations of random mating within the recipient taxon, then acquire data set for analysis.
5. Repeat steps 1 to 4 an additional 199 times so that 200 data sets (corresponding to 200 total hybridization events during 2200 consecutive generations) are acquired per run.
6. Repeat steps 1 to 5 so that five replicate runs are performed for each of the four starting data sets for all topological models of gene flow.

### *Analysis of Simulated Data Sets*

All data sets contained five taxa and 50 individuals. Two parsimony-based procedures were used to analyze the data. For the "best tree" procedure, a strict consensus of all most parsimonious trees was found subsequent to a PAUP* (Swofford, 1999) search that used the following settings: heuristic search strategy, TBR branch swapping, 100 random taxon addition replicates, MULTREES on. For the "bootstrap" procedure, the "TBR-M" settings of Debry and Olmstead (2000) were used in order to expedite searches without sacrificing accuracy. Five hundred bootstrap replicates were performed. Distance-based analyses were performed using the same procedures as the parsimony analyses except that the neighbor-joining (NJ) algorithm in PAUP* was used for tree reconstruction. Genetic distances were calculated using mean character differences. ML methods were not computationally tractable given the large number of data sets simulated (estimated computation time = 1388 days) and were not used.

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



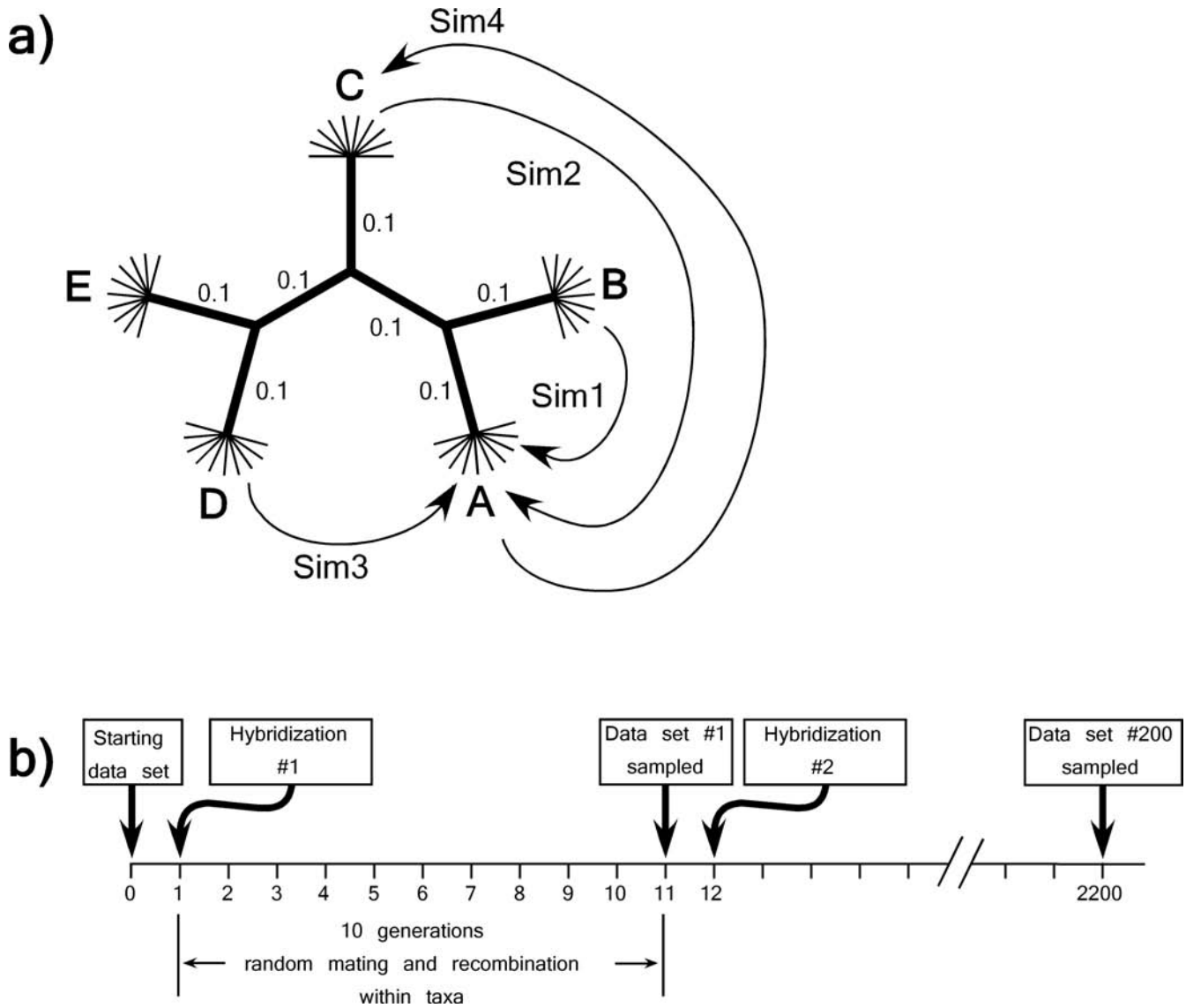



FIGURE 1. (a) Topological models of unidirectional gene flow simulated within five-taxon trees. Each taxon included 10 individuals. Substitution probability for starting data sets is shown along internodes. Arrows indicate direction of gene flow from donor to recipient taxon. (b) Model of simulated reticulate evolution. Hybridization, which produced a single F1 individual in the recipient taxon, was followed by 10 generations of random mating within taxa, after which data sets were sampled.

For parsimony and NJ analyses, support for clades A, B, C, D, E, AB, DE, AC, AD, AE, BC, BD, BE, CD, and CE was extracted from PAUP* log files generated during the analyses. For the "bootstrap" procedure, support was measured as the percent bootstrap support associated with each clade; for the "best tree" procedure, support was measured as the number of consensus trees out of a possible 20 (4 starting data sets × 5 replicate runs per data set) that resolved the clade of interest.

Two network methods were used. Minimum spanning networks (MSN; Excoffier and Smouse, 1994) were calculated using Arlequin 2.0 (Schneider et al., 2000). Split networks were calculated using SplitsTree 4.4b16 (Huson and Bryant, 2006). Genetic distances were calculated using the "UncorrectedP" method. Split networks were constructed with "NeighborNet" distances transformation (Bryant and Moulton, 2004) and "EqualAngle" splits transformation settings. One hundred bootstrap replicates were performed per data set, and support values for the trivial splits A, B, C, D, and E and the nontrivial splits AB, DE, AC, AD, AE, BC, BD, BE, CD, and CE were collected for further analysis. For simplicity, splits are named here using only the smaller of the two possible subsets of taxon names (e.g., split AB|CDE is named AB).

Two nonhierarchical statistical procedures were used: $F_{st}$, and principal coordinate analysis followed by nonparametric modal clustering (PCO-MC). For the $F_{st}$ analysis, only donor and recipient taxa were considered and

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



$\theta_p$ (*sensu* Weir and Cockerham, 1984) was calculated for haploid data using the application GDA (Lewis and Zaykin, 2002). For PCO-MC, Jaccard (1908) distances between haplotypes were calculated using the program module SIMQUAL of NTSYS (Exeter Software). Although the choice of distance coefficient will affect the calculation of principal coordinates, Jaccard distances were used because they have been argued to be appropriate for binary multi-locus data (Landry and LaPointe, 1996) and are commonly used for analysis of dominant marker data sets. The program modules DCENTER and EIGEN were subsequently used to compute the first three principal coordinates. The number of clusters identified within the principal coordinate data was inferred using PROC MODECLUS in SAS (Sarle and Kuo, 1993) and the following options: STANDARD; METHOD=6; CASCADE=1. The value for the fixed-radius smoothing parameter R (which defines the dimensions of the spherical uniform kernel used to estimate cluster density) was approximated as follows:

1. Simulate 200 five-taxon (50 terminal) starting data sets as described above.
2. Obtain the first three principal coordinates for the 200 data sets.
3. Analyze PCO data from step 2 using PROC MODECLUS and a range of R values. For each PCO data set, determine the minimum value for R at which four clusters are present. Values of R that are smaller than this minimum value cause the correct inference of five clusters in a given starting data set.
4. Find the smallest value in the list of 200 R values obtained in step 3. This is the value for R used in this study.

Using this extreme minimum value (in this case, R = 0.0401) for analyses is conservative because the PCO-MC procedure would be expected to predict the correct number of clusters in starting data sets 99.5% of the time (on average, four clusters would be found in only one out of 200 simulated starting data sets).

To examine the level of conflicting phylogenetic signal in the simulated data sets using a tree-based approach, the program RSW1.1 (Trueman and Gibbs, 2002) was used. RSW1.1 is a PERL script that forces PAUP* to implement the RSW procedure of Trueman (1998). RSW identifies conflicting signal as follows:

1. Find an optimal tree for the complete (overall) data set.
2. Assemble a data set of characters that are inconsistent (CI < 1) with tree from step 1.
3. Find an optimal tree for the data set assembled in step 2.
4. Conclude that overall data set contains conflicting signal when trees from step 3 are significantly different from step 1 tree.

For step 4, a "significant" difference is found when different clades, each supported by some user-specified minimum level of bootstrap support (the "cutoff percent"), are present in the bootstrap trees from steps 1 and 3. For this study, the cutoff percent value in RSW1.1 was set to 1% so that RSW1.1 was forced to find "significant" secondary signal in all simulated data sets. In this way, levels of conflicting support could be measured for all clades at all time points. Bootstrap searches conducted by RSW1.1 were performed using the aforementioned "bootstrap" settings, and optimal topologies for each data partition were determined by a single heuristic parsimony search using PAUP* defaults.

### *Evaluation of Performance*

*Erosion of type I error.*—In order to identify regions in which type I error (the erroneous rejection of the null hypothesis, $H_0$: the recipient taxon is not a distinct evolutionary lineage) occurred, a cutoff criterion was established for each analytical procedure, beyond which it was claimed that a reticulate history could be inferred. For the "best tree" searches, reticulate history was defined to be inferred when the recipient taxon was no longer found to be monophyletic in the best tree for a given replicate. In practice, adopting this criterion meant that a reticulate history was inferred for the recipient taxon whenever one or more members was found nested within the donor taxon, or vice versa. For "bootstrap" searches, reticulate history was inferred whenever the support value for the recipient taxon dropped below 95%. For $F_{st}$ analyses, reticulate history was inferred when $F_{st}$ dropped below 0.95 and did not subsequently return to higher values. For PCO-MC, the point at which the number of inferred clusters changed from five to four (with R = 0.0401) and did not return to a higher value, was used. Topological evidence of reticulate history was not apparent in the minimum spanning networks so a nontopological approach was used: Reticulate history was defined to be inferred at a point when the mean length of branches between the donor taxon and recipient taxon individuals ceased to be significantly longer (using a one-tailed *t*-test with $\alpha \leq 0.05$) than the mean length of branches among individuals within the recipient taxon. The effect of this test was to identify the point at which donor and recipient taxa were no longer statistically distinct from one another based on the minimum spanning network. For the NeighborNet method, reticulate history was inferred whenever the bootstrap support value for the trivial split containing the recipient taxon dropped below 95% and did not return to a higher value.

The time to inferred reticulate history (TIRH) was measured as the number of hybridization events required before cutoff criteria were met. TIRH is therefore a measure of the rate of erosion of type I error caused by persistent gene flow. In total, 16,000 data sets (200 data sets per run × 5 replicate runs × 4 starting data sets × 4 topological models of gene flow) were analyzed using each approach (parsimony "best tree," parsimony "bootstrap," NJ "best tree," NJ "bootstrap," $F_{st}$, PCO-MC, MSN, and

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



NeighborNet) and examined to determine whether the relevant cutoff criterion had been met. For a given procedure, differences in TIRH observed between runs could be due to the topological model of gene flow, the starting data set, or from the stochastic processes of recombination and drift built into the model. Nested analysis of variance (ANOVA) was used to apportion the variance and test for significant differences in TIRH caused by these factors.

*Evidence of conflicting signal.*—NeighborNet and RSW can also indicate reticulation by exposing conflicting signal in a data set. However, this evidence of reticulation could not be fairly evaluated using a single cutoff criterion. A different approach, involving sensitivity analysis of type I and type II error across the simulation time course, was used. For NeighborNet, bootstrap support values for the non-trivial splits described previously were collected. Support values for the corresponding clades from the overall and secondary signal partitions of RSW analyses were likewise collected. The relative support values for two contradictory splits (or for contradictory clades in the overall and secondary RSW data partitions) indicate the magnitude of conflicting phylogenetic signal. For example, if split AB had 95% bootstrap support and split AC had 98% bootstrap support, then there would be evidence of substantial conflicting signal for taxon A. The conclusion that taxon A was as closely related to taxon B as it was to taxon C, and hence that taxon A may have a reticulate history, would then be supported. Type I error occurred when such evidence of reticulate history was not present (causing the null hypothesis to be falsely rejected), and type II error occurred when conflicting signal suggested that a taxon known to be monophyletic had a reticulate history.

Conflicting signal was evaluated using 9 different bootstrap support cutoff values (99%, 95%, 90%, 85%, 75%, 70%, 65%, 60%, 50%). When bootstrap support for contradictory clades or splits exceeded these "cutoff percents," "significant" conflicting signal was considered to have been found. For each cutoff percent and sampled time point, the number of correct inferences of reticulate history was determined for the 20 data sets available. Similarly, the number of incorrect inferences (i.e., when a taxon known to be monophyletic was suggested to have a reticulate history based on conflicting signal levels) was determined. The probability of success (correctly identifying the reticulate taxon), the probability of failure (determining that a monophyletic taxon had a reticulate history, i.e., type II error), and the probability of no inference of reticulate history (equivalent to type I error) could then be calculated for each time point (see Appendix 1 for formulas).

## RESULTS

### *Topological Consequences of Reticulate Evolution on Tree-Based Analyses*

The effect of gene flow between taxa on tree-based analyses is summarized using results from the parsimony analysis of Sim3 data. The topological conditions for gene flow examined in Sim1, Sim2, and Sim4 produced a subset of the responses seen in the more complex Sim3. Results from NJ methods were qualitatively similar to parsimony. Differences in the topology of the strict consensus trees were observed between replicate runs at some sampling points. Given this variation, majority-rule consensus trees were used to determine the most frequently recovered clades over the time course and are used as a tool to generalize the topological effects of gene flow between taxa.

Figure 2 shows the topological changes that occurred during parsimony analysis of Sim3 data. The first observed change from the starting condition (Fig. 2a), occurring after six cycles of hybridization and genetic drift ($t = 6$), was a decrease in the frequency of recovery of clade AB (Fig. 2b). The two alternative topologies at this time point either contained polytomy ABC or were like the tree given in Figure 2c. By $t = 11$ (Fig. 2c), the first topological change relative to the starting tree had occurred in the majority of trees. At this point, taxon A had assumed a new position, forcing the creation of clade BC. Alternative topologies at $t = 11$ showed either the starting topology or were like Figure 2d. By $t = 23$ (Fig. 2d), clade A was recovered as sister to clade D in all 20 simulations, completing all internal topological changes found during Sim3. A slightly larger number of hybridization events was required to reach this point for Sim3 ($t = 23$) than for Sim2 ($t = 19$) or Sim4 ($t = 17$). No topological changes occurred in Sim1 because donor and recipient were sister taxa.

Figure 2e shows the last point ($t = 96$) at which the majority-rule tree retained distinct donor and recipient taxa. Alternate topologies included clade A nested within clade D, clade D nested within clade A, and a topology like Figure 2f. After this point, the majority of the strict consensus trees from the 20 runs showed an unresolved polytomy containing all individuals from taxa A and D. Figure 2f shows the point at which the final, stable topology was reached ($t = 163$). At this point, taxa A and D were no longer identified as distinct monophyletic groups in any of the replicate data sets.

Although the exact timing of topological changes varied between replicate runs, all showed an ordered progression through the phases described above. Figure 3 shows trees from a single representative Sim3 run taken at time points described for Figure 2. Characters introduced to recipient taxon A from donor taxon D caused the appearance of false hierarchical structure within taxon A.

Bootstrap support values were calculated for all data sets to determine statistical support for the tree structure during the simulations. Figure 4 shows that support for the monophyly of donor and recipient taxa, on average, eroded gradually, subsequent to internal topological changes. During the period of topological change (which occurred soon after gene flow began), internal branches received high bootstrap support (Fig. 4c, inset).

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



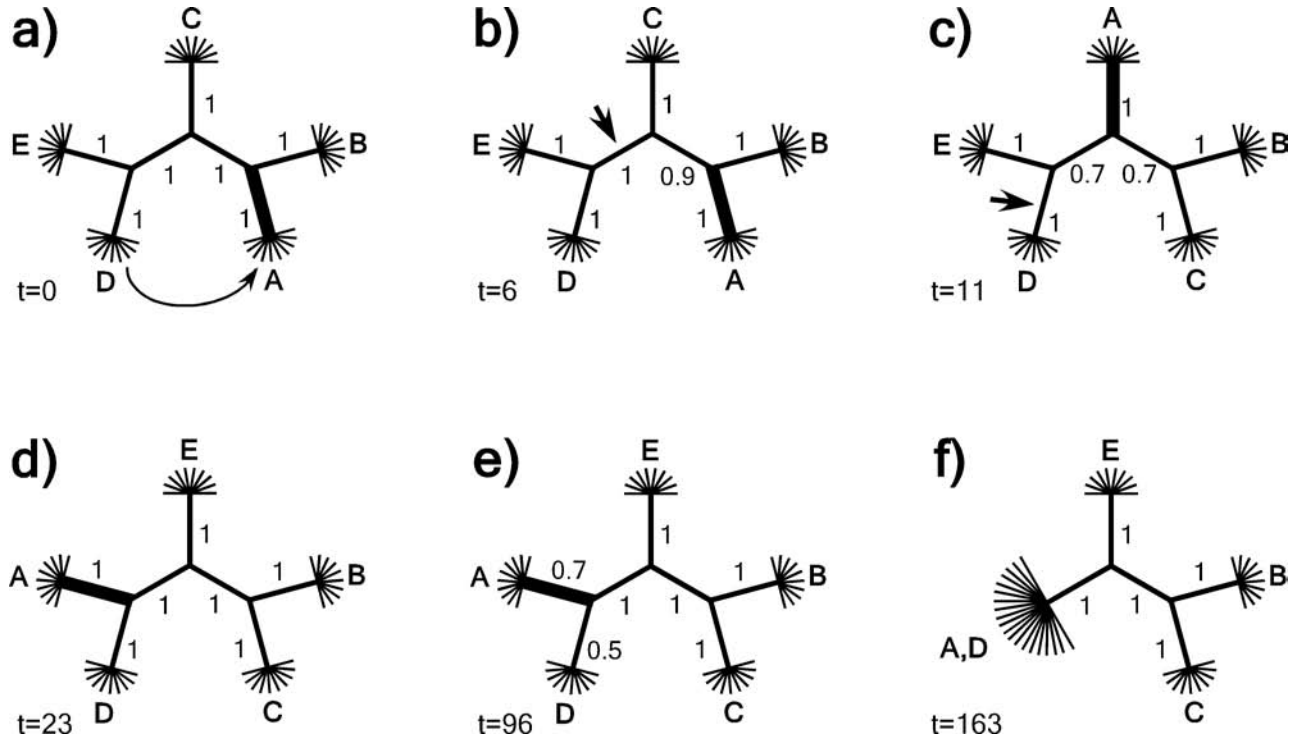



FIGURE 2. Topological effect of ongoing, unidirectional gene flow between taxa on parsimony analysis. The initial topology and direction of gene flow are shown in (a), and the final, stable topology is shown in (f). The recipient taxon (A) is marked by a bold branch. Arrows in (b) and (c) indicate the tree position to which taxon A moved during the next topological change. Time (t), measured in number of hybridization events, is shown at the bottom of each figure. Majority-rule consensus trees are labeled with clade frequency at nodes. Gene flow between taxa separated by two or more internodes caused predictable topological rearrangements as the recipient clade "flowed" within the tree in the opposite direction as genes. Topological rearrangements among internal branches occurred soon after the start of hybridization and were followed by a longer period where support for the donor and recipient taxa as monophyletic groups gradually eroded.

### *Topological Consequences of Reticulate Evolution on Network Analyses*

Four equally likely MSN topologies exist for the starting tree topology used: E-D-C-B-A; E-D-C-A-B; D-E-C-B-A; D-E-C-A-B. Because the four starting data sets used here were chosen at random, only the former three MSNs were represented in this study. Sim3 is used to summarize the response. Topological changes for the four possible starting conditions occurred as in Figure 5, and the MSNs observed during a single Sim3 run are shown in Figure 3. Prior to any topological changes, gene flow induced variation, and therefore structure, within the recipient taxon (Fig. 3; t = 6). The first topological change involved a rearrangement of the network such that donor taxon D and recipient taxon A became adjacent nodes (summarized in Fig. 5, but not visible as a unique event in Fig. 3). With continuing gene flow, as taxon A lost its similarity to taxon B and became more similar to taxon D, a second topological change occurred such that taxon B became adjacent to taxon C (Fig. 3, t = 11). After this final topological transition, the branch between taxon A and taxon D progressively diminished in length (Fig. 3, t = 23, 96) until donor and recipient were no longer distinguishable from one another (t = 163). At no time did the shortest path along the network that included all individuals of taxon A also include an individual from another taxon, thus the recipient taxon remained a distinguishable group for much of the simulation time course, until the cutoff criterion was met. Two possible end conditions were observed: E-A,D-C-B; B-C-E-A,D.

The split networks resulting from application of NeighborNet to a single Sim3 run are shown in Figure 3. Initially, in the starting data sets, the networks were largely tree-like, with only minor evidence of conflicting splits, as measured by split weight (branch length) and bootstrap support. The mean (± SD) bootstrap support for splits that conflicted with the starting tree in 200 simulated starting data sets was 27% ± 26%. However, very high support values for conflicting splits, such as those shown in Figure 3 (t = 0) were occasionally observed in starting data sets (which were simulated under a strictly bifurcating model of evolution). Soon after gene flow began (Fig. 3, t = 6), split BC became more prominent, garnering large bootstrap support values, and a split weight that was comparable to split AB. At this phase, the erroneous interpretation that taxon B had a reticulate history was supported. By t = 11, the split weight for the trivial splits A and D had begun to decrease relative to trivial splits B, C, and E, indicating the growing similarity between donor and recipient taxa. Conflicting splits AD and AB were equally well supported, leading to the

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



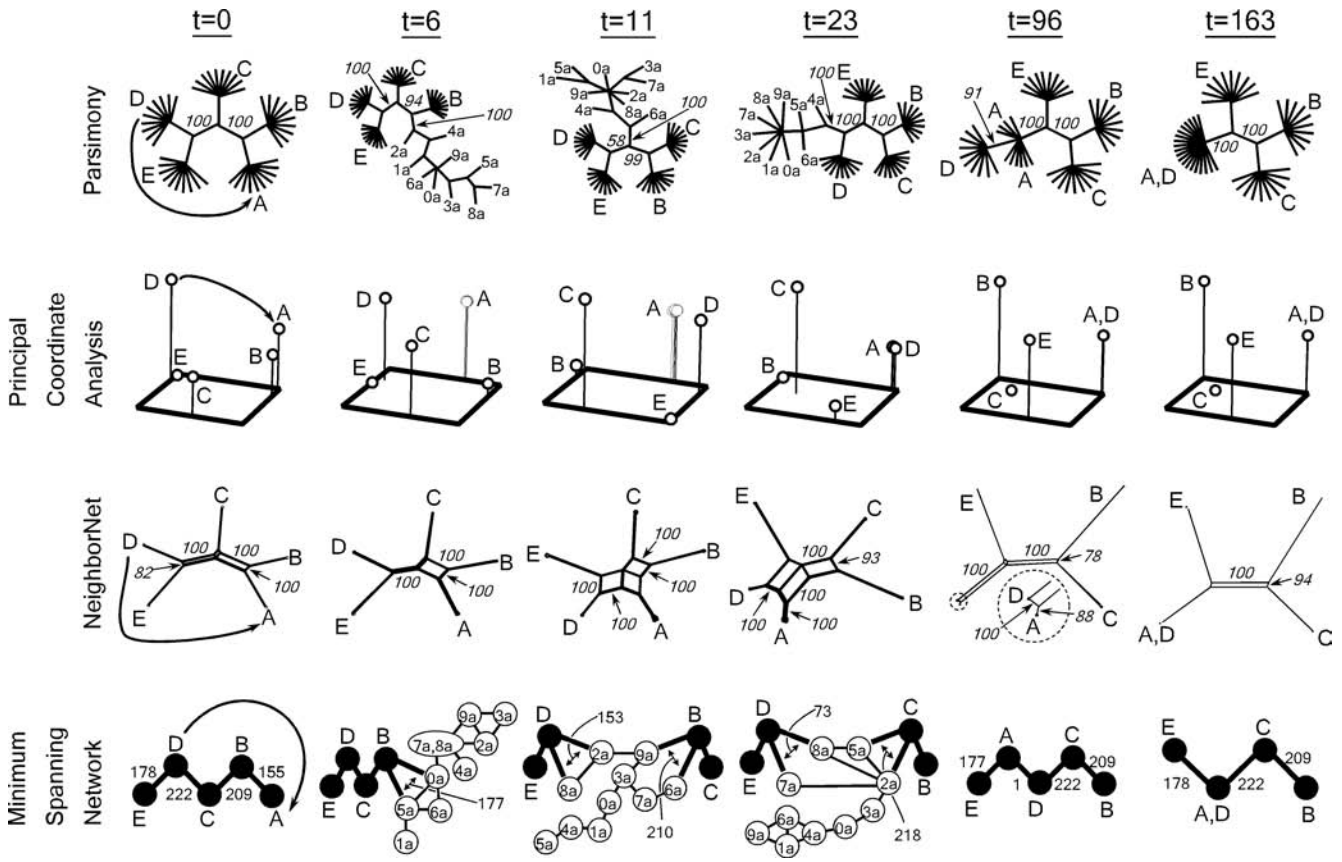



FIGURE 3. Time series showing the signature of reticulate evolution on four analytical methods. Curved arrows indicate direction of gene flow between taxa. Data sets were sampled at time points indicated at top. Taxa (populations) are indicated with capital letters; individuals are in lowercase and are numbered. Relevant bootstrap support values are shown for parsimony and NeighborNet analyses. Relevant branch lengths are shown for MSN analyses. Bold branches in the MSN panels show connections between taxa. These branches were typically 10 to 100 times longer than the branches within recipient taxon A. The first three principal coordinates are plotted for PCO analyses. Minor variation and/or historical structure introduced into the recipient taxon by gene flow from the donor taxon were apparent in all methods.

correct interpretation that taxon A may have a reticulate history. However, taxa B and D were also suggested to be reticulate taxa at this time point. At t = 23, splits AB and DE began to diminish in importance, as measured by split weight, relative to AD and BC. Bootstrap support for split AB also began to decrease. By t = 96, strong evidence of conflicting signal had decreased substantially as the split weight for DE became much smaller than AD. However, bootstrap support did not always similarly decrease: in the Sim3 replicate shown in Figure 3, at t = 96, split AD = 100%, split DE = 100%. Thus split weight and split support values are not always tightly correlated. The final, stable topology was reached by t = 163 as bootstrap support for split DE disappeared and taxa A and D became identical. Replicate simulations followed a similar pattern.

### *Relative Performance of Phenetic, Tree-Based, and Network Procedures*

*Erosion of type I error.*—Nested analysis of variance was used to determine if variation in TIRH among replicate runs was attributable to differences between starting data sets and/or differences between topological models used to define gene flow (Table 1). No significant effect on TIRH was found to be caused by using different starting data sets within a particular topological model of gene flow. Therefore, variation in TIRH found within Sim1, Sim2, Sim3, and Sim4 is largely due to the dynamics of gene flow and genetic drift, rather than the starting data set. Moreover, because no significant effect was found, five replicate runs for each of the four starting data sets were likely sufficient to estimate the TIRH parameter for a given analytical method.

Performance rankings based on TIRH are shown in Figure 6. Averaged across all four topological models, the following pairs of methods were not significantly different from one another: parsimony "bootstrap" and NeighborNet ($P = 0.453$), parsimony "best tree" and MSN ($P = 0.8596$), $F_{st}$ and NJ "best tree" ($P = 0.2284$). Relative performance could also be ranked by comparing the TIRH values of identical simulation runs between methods ($\geq$ indicates better than or equivalent to):

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



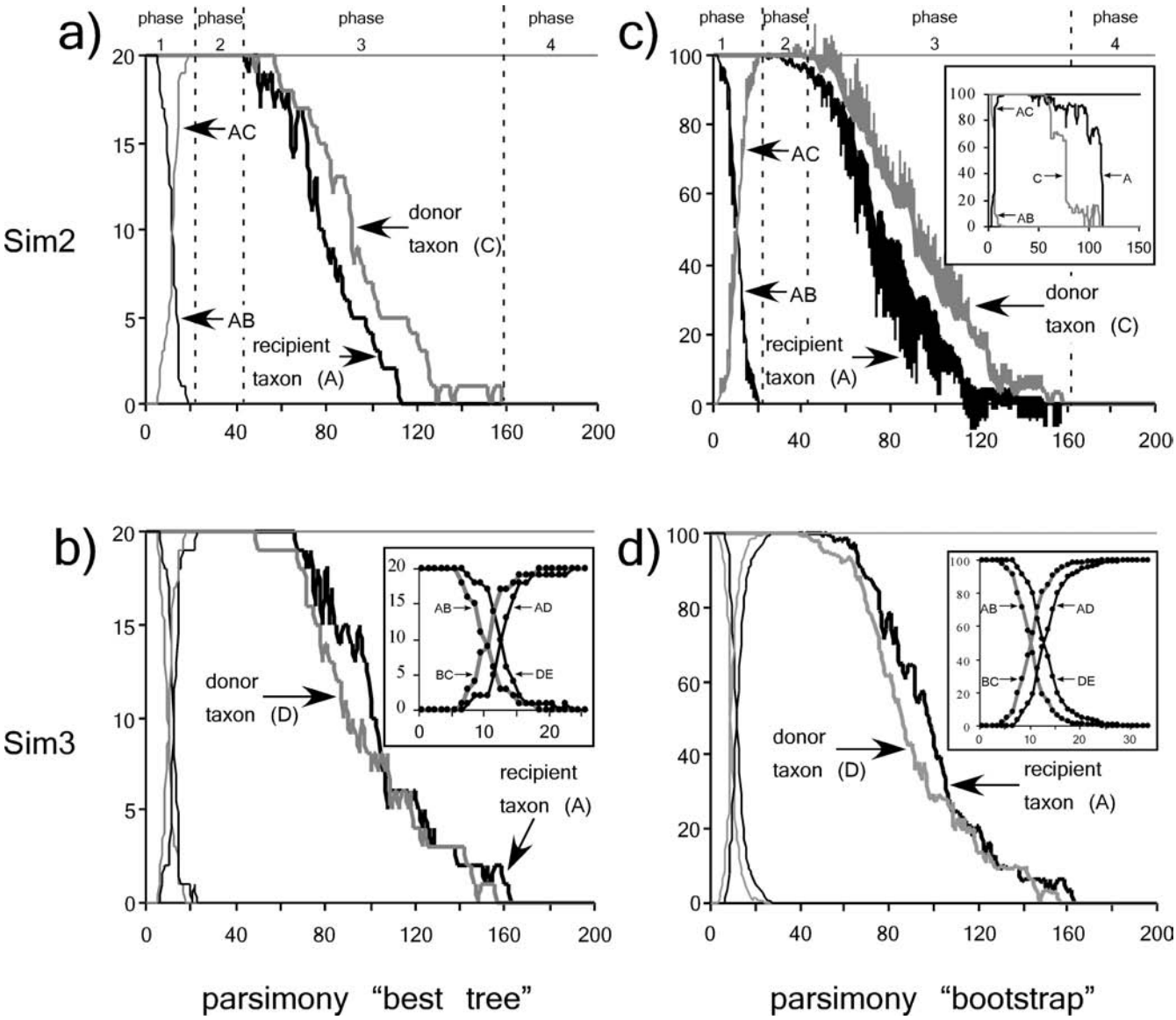



FIGURE 4. Support for selected clades after parsimony analysis of Sim2 and Sim3 data. Criterion used for searches shown at bottom. In all graphs, $x$-axis is number of hybridization events since start of simulation. (a, b) Frequency of selected clades found in parsimony trees from 20 replicate simulations. $y$-Axis is number of trees. (c, d) Mean bootstrap support for selected clades. $y$-Axis is bootstrap support value. For clarity, error bars, shown as one standard deviation from the mean bootstrap value, are indicated in one direction only in (c). Error values for (d) were similar. Sim2 graphs indicate four characteristic phases that are a consequence of gene flow between nonsister taxa: phase 1, topological change; phase 2, erroneous inference that recipient (reticulate) taxon is the monophyletic sister group of donor taxon; phase 3, gradual erosion of support for phase 2 inference; phase 4, donor and recipient genetically identical. Inset (c) shows bootstrap support values from a single representative run. Insets (b) and (d) show summary values but expand the region of topological change (phase 1) for clarity. Sim1 results were similar to Sim2 and Sim3 except that no topological changes occurred, so only donor and recipient clade support values changed during the course of the simulation. Sim4 results were nearly identical to Sim2.

1. PCO-MC $\geq$ NeighborNet (64 of 80 simulations)
2. NeighborNet $\geq$ Parsimony "bootstrap" (62 of 80 simulations)
3. Parsimony "bootstrap" $\geq$ NJ "bootstrap" (77 of 80 simulations)
4. NJ "bootstrap" $\geq$ Parsimony "best tree" (80 of 80 simulations)
5. Parsimony "best tree" $\geq$ MSN (48 of 80 simulations)
6. MSN $\geq$ NJ "best tree" (70 of 80 simulations)
7. NJ "best tree" $\geq$ $F_{st}$ (77 of 80 simulations)

A significant difference in TIRH was found between topological models of gene flow using PCO-MC, parsimony "bootstrap," and NJ "bootstrap" criteria (Table 1). NeighborNet, MSN, parsimony "best tree," NJ "best tree," and $F_{st}$ methods were insensitive ($P > 0.05$) to differences between models of gene flow. No significant differences in TIRH were found between Sim2 and Sim4 for any of the criteria used ($P > 0.05$). Therefore, the direction of gene flow within the tree had no effect on TIRH.

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



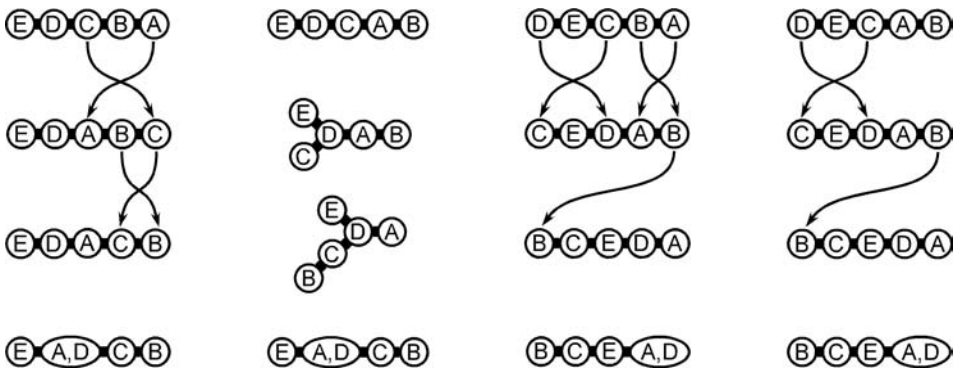



FIGURE 5. Topological effect of ongoing, unidirectional gene flow between taxa on minimum spanning network analysis. One of four possible starting networks is shown at the top of each column. Major topological transitions due to the Sim3 model of gene flow are shown in a downward direction with one of two possible final, stable networks shown at the bottom of each panel. Arrows emphasize rearrangements within the networks.

*Evidence of conflicting signal.*—The magnitude and implications of conflicting signal revealed by the NeighborNet and RSW methods were evaluated by varying the bootstrap support values used to define when conflict was significant. Figure 7 shows the probability of successful and erroneous inferences using nine different bootstrap support values as the cutoff criterion. Conflicting splits or clades are not possible when gene flow occurs between sister taxa. Thus in Sim1, any conflicting signal must be due to routine homoplasy, rather than reticulation, and the probability of success must be zero (as indicated in Fig. 7). The Sim1 results demonstrate that the probability of incorrect inferences increases dramatically as the cutoff percent drops below 90% and 95%, even when it is not possible for correct conflicting signal to be present. Sim2 and 3 show a similar pattern, suggesting that support values below 90% to 95% result in an increased probability of type II error, and should not be used.

In gene flow models where correct conflicting signal could be visualized (Sim2 and 3), the probability of success increased early in the simulation then decreased with continued gene flow. Thus correct conflicting signal appeared soon after the onset of gene flow but was ephemeral in the face of persistent gene flow. For Sim2, the cumulative probability of success (as measured by the

TABLE 1. Nested analysis of variance in TIRH.

| ANOVA | | Parsimony "bootstrap" | | NJ "bootstrap" | |
|---|---|---|---|---|---|
| Source of variation | *df* | SS | *P* | SS | *P* |
| Between gene flow models | 3 | 5,062.7 | < 0.0001 | 3, 129.0 | 0.0222 |
| Between data sets within models | 12 | 1,610.3 | 0.7774 | 1,845.5 | 0.9039 |
| Within data sets (error) | 64 | 12,905.2 | | 19,485.2 | |
| Total | 79 | 19,578.2 | | 24,459.7 | |
| ANOVA cont'd | | Parsimony "best tree" | | NJ "best tree" | |
| Source of variation | *df* | SS | *P* | SS | *P* |
| Between gene flow models | 3 | 5762.7 | 0.0602 | 370.5 | 0.9033 |
| Between data sets within models | 12 | 6894.1 | 0.6729 | 3986.3 | 0.9011 |
| Within data sets (error) | 64 | 47,408.4 | | 41,735.2 | |
| Total | 79 | 60,065.2 | | 46,092.0 | |
| ANOVA cont'd | | $F_{st}$ | | PCO-MC | |
| Source of variation | *df* | SS | *P* | SS | *P* |
| Between gene flow models | 3 | 1901.8 | 0.3956 | 11,989.5 | <.0001 |
| Between data sets within models | 12 | 4657.8 | 0.8205 | 2145.3 | 0.6546 |
| Within data sets (error) | 64 | 40,296.8 | | 14,406.8 | |
| Total | 79 | 46,856.4 | | 28,541.6 | |
| ANOVA cont'd | | MSN | | NeighborNet | |
| Source of variation | *df* | SS | *P* | SS | *P* |
| Between gene flow models | 3 | 2756.1 | 0.0774 | 1248.7 | 0.1222 |
| Between data sets within models | 12 | 2824.7 | 0.8251 | 3263.9 | 0.2354 |
| Within data sets (error) | 64 | 24,657.2 | | 13,292.4 | |
| Total | 79 | 30,238.0 | | 17,805.0 | |

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



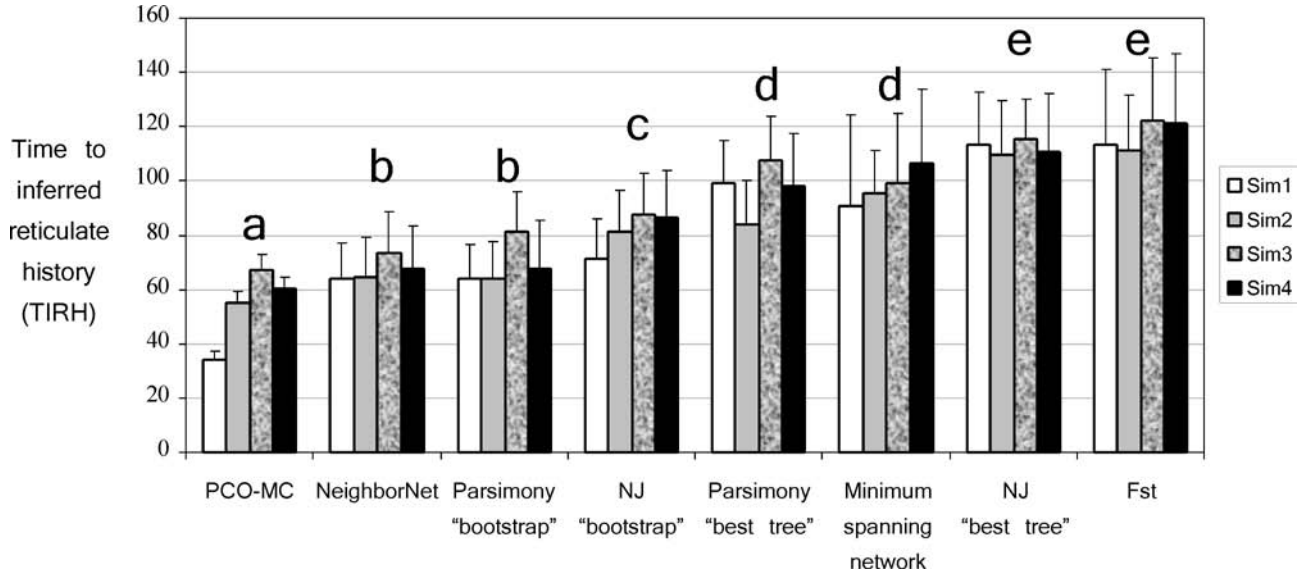



FIGURE 6. Relative performance of analytical procedures used to infer reticulate history under four topological models of gene flow in five-taxon trees. TIRH measured in number of hybridization events since start of simulation. Error bars show one standard deviation above the mean. Letters above columns indicate which groups of analytical procedures were significantly different ($P < 0.001$) from one another. When PCO data were treated as a density landscape and analyzed using a nonparametric modal clustering procedure (PCO-MC), reticulate history was inferred significantly sooner than with the tree-based and network methods tested.

area under the curve) for cutoff percents above 90% was 5- to 57-fold higher for NeighborNet than for RSW. Moreover, for brief periods of time in Sim2, the probability of success for NeighborNet exceeded 0.9, whereas RSW seldom exceeded 0.5. Under the gene flow conditions modeled in Sim2, NeighborNet revealed conflicting signal better than RSW. However, under more complex gene flow conditions (Sim3), both methods performed poorly, both in terms of low cumulative probability of success, and high probability of failure. For the duration of Sim3, the probability of an incorrect inference exceeded the probability of a correct inference for both methods.

## DISCUSSION

A simple model of gene flow was used to generate data sets containing phylogenetic signal (i.e., that generated from a hierarchical, bifurcating process) that was complicated by an ongoing reticulate process (i.e., simulated hybridization between taxa). Two tree-based methods, two network methods, and two nonhierarchical methods were evaluated for their relative performance in revealing correct, reticulate patterns of relationship in the simulated data sets. The analytical methods evaluated were chosen because, in aggregate, their appropriate application spans the continuum between phylogenetic and population genetic processes. For example, parsimony is most appropriately applied to data sets derived from hierarchical, bifurcating processes, whereas $F_{st}$ was designed to describe the partitioning of genetic variation among populations, where relationships are largely tokogenetic in nature. Networks and ordination analysis have been recommended for use when data span both realms (Sneath and Sokal, 1973; Posada and Crandall, 2001).

### *Consequences of Reticulate Evolution on Tree-Based Analyses*

Examination of tree topologies and support values along the course of the simulations revealed four characteristic phases associated with persistent gene flow between previously monophyletic taxa (Fig. 4a, c). First, whenever taxa were separated by two or more nodes in the starting tree (e.g., Sim2, Sim3, Sim4), a rapid shift in internal branching order occurred following the commencement of gene flow. During this period (phase 1), the recipient taxon "flowed" within the tree in the opposite direction as genes, until donor and recipient appeared as monophyletic sister taxa. This rapid restructuring should be expected anytime gene flow begins between non-sister lineages.

As the recipient taxon moved through the underlying tree structure during phase 1 of Sim3, erroneous topologies that placed it with neither its original sister taxon nor the donor taxon were recovered. Bootstrap support for internal branches on the erroneous topology shown in Figure 2c exceeded 85% in 5 of 20, and 9 of 20 replicate runs for parsimony and NJ, respectively. Thus tree-based analysis of real reticulate data sets (i.e., sampled from nature) may result in the inference of highly supported, but incorrect relationships among taxa.

Following the brief period of topological rearrangements, a second phase occurred where the inferred topology remained static, with donor and recipient taxa in a well-supported sister group arrangement (phase 2; Fig. 2d, Fig. 3, t = 23). In real data sets acquired during this and the previous phase, the use of parsimony and NJ will result in the erroneous inference that the recipient taxon is a monophyletic group when in fact it has a reticulate history. This bias has previously been demonstrated

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



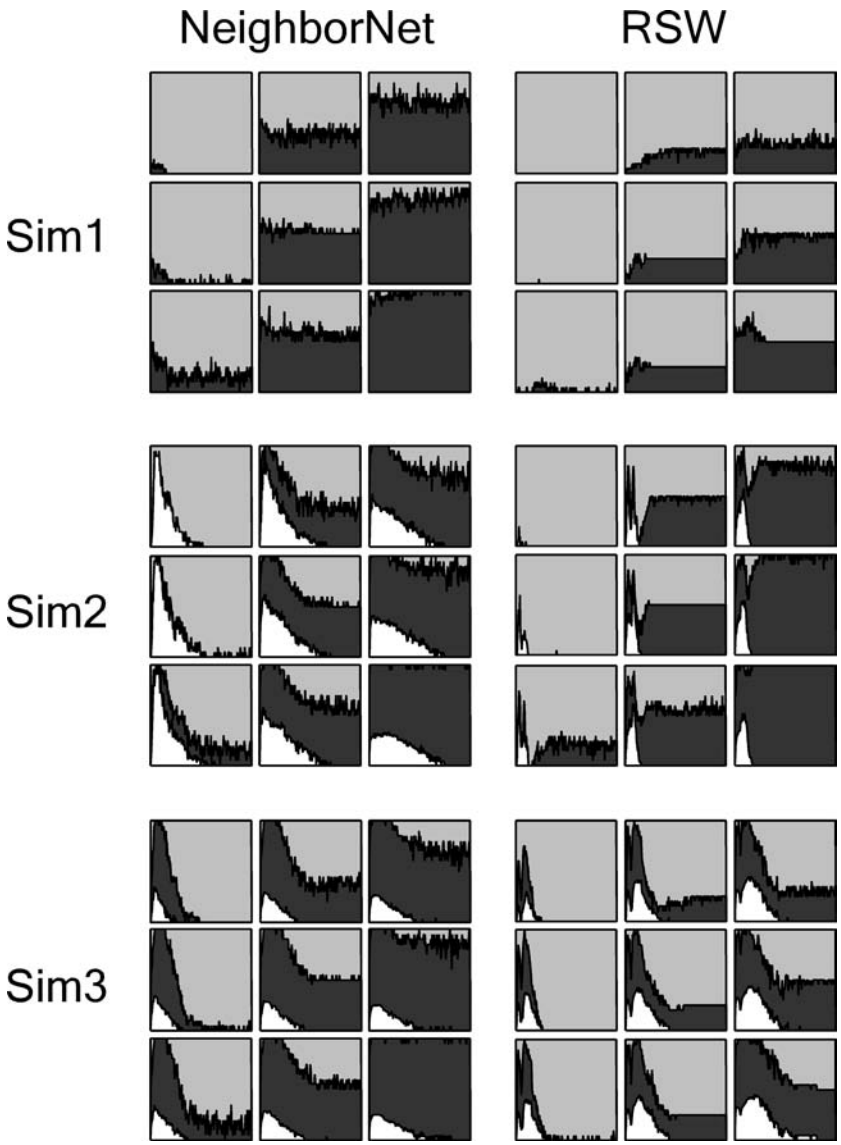



FIGURE 7. Sensitivity analysis of conflicting signal revealed by NeighborNet and RSW. In all graphs, *y*-axis is probability, scaled from 0 to 1. *x*-Axis is time, measured in hybridization events from 0 to 200. Bootstrap values used as the cutoff criterion for inferring significant conflict were, starting at the top of the leftmost column of each group of nine graphs: left column, 99%, 95%, 90%; middle column, 85%, 75%, 70%; right column, 65%, 60%, 50%. White-shaded regions indicate the probability of a successful inference that the recipient taxon has a reticulate history. Black-shaded regions indicate the probability of an erroneous inference that a monophyletic taxon has a reticulate history (type II error). Gray-shaded areas indicate the probability of no inference of reticulate history (type I error).

for the NJ algorithm by Allaby and Brown (2003), who, based on a different simulation approach, concluded that NJ trees derived from multilocus data sets cannot effectively distinguish between single (i.e., monophyletic) and multiple (i.e., reticulate) domestication hypotheses of the evolution of crop species. This bias also presents a great risk for errors in identifying distinct evolutionary lineages for conservation purposes and for defining and delimiting phylogenetic species (*sensu* de Queiroz and Donoghue, 1988). In lower level phylogenetic studies where reticulation is possible, the appearance of a taxon as a monophyletic group in a bifurcating tree is not a reliable indicator that the taxon is a distinct evolutionary lineage.

Third, a lengthy phase occurred during which support for the monophyly of donor and recipient taxa was gradually eroded (phase 3; Fig. 4a, c). This phase delayed the point at which reconstructed trees began to correctly suggest that the reticulate taxon was not monophyletic. Lastly, with persistent introgression, all individuals from the donor and recipient taxa became members of the same clade, and no further change in tree structure occurred (phase 4).

McDade (1992) showed that F1 hybrid individuals may infrequently be found in a tree in places other than with their parents. In our simulation, complex backcrossed individuals were not found outside the recipient clade, except for during the final stages of the

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



collapse into polytomy, when they were occasionally nested within the donor taxon. Humphries (1983), Funk (1985), and Rieseberg and Ellstrand (1993) conclude that the placement of F1 hybrid individuals in cladograms is not readily predicted. Our results indicate that the position of back-crossed individuals, as well as the reticulate taxa to which they belong, should follow a predictable pattern during persistent unidirectional gene flow.

Humphries (1983), Funk (1985), McDade (1992), and Rieseberg and Ellstrand (1993) examined hypothetical character matrices or morphological characters, so their results may not be directly comparable with the present study. On the other hand, their results might be correctly extended to large, multi-locus, molecular data sets containing F1 hybrid individuals. However, they are not necessarily applicable to studies of natural systems where back-crossed individuals predominate over F1 hybrids. In these cases, our results predict that, when recombination mostly occurs within taxa (with infrequent gene flow between taxa), members of a reticulate taxon will generally appear as a monophyletic group, in spite of their reticulate history, in reconstructed phylogenetic trees.

### *Erosion of Type I Error*

Simulated gene flow began with a five-taxon, 50-terminal starting data set that was simulated under a bifurcating model of evolution. Support for the monophyly of each taxon and for the branching relationships among taxa was high in the starting data sets. As soon as a single character state present in the donor taxon was introgressed into the recipient taxon, the recipient taxon became a product of a reticulate evolutionary process and was no longer, strictly speaking, monophyletic. During these very early stages in the simulation, all methods erroneously identified the recipient taxon to be a distinct evolutionary lineage. We wished to determine how much gene flow, as measured by the number of hybridization events, was necessary before the null hypothesis (that the recipient taxon is not a distinct evolutionary lineage) was no longer erroneously rejected by the various analytical methods. The rate of erosion of the type I error apparent early in the simulations is a measure of the efficacy of a method: high-performing methods will cease to erroneously reject the null hypothesis sooner after the commencement of gene flow than poorly performing methods. Put another way, the cumulative probability of type I error over the course of the simulation is lowest for the method that correctly accepts the null hypothesis the soonest after the commencement of gene flow.

*Relative performance of tree-based methods.*—All tree-based methods used were ultimately able to reveal the reticulate history of the recipient taxon by accepting the null hypothesis as true. However, when using tree-based methods, reticulation could only be inferred after extensive gene flow and recombination. Even with the unrealistically small population sizes and high gene flow rates modeled, the number of generations before a reticulate taxon could be identified varied from approximately 600 to greater than 1000.

Although all tree-based methods performed poorly, rankings were unambiguous. Methods that utilized bootstrap support significantly outperformed methods that relied on tree topology alone ($P < 0.0001$). This is not surprising, given that erosion of bootstrap support would be expected to occur prior to loss of resolution of a clade with ever-diminishing character support. Parsimony "bootstrap" performed significantly ($P < 0.0001$) better than NJ "bootstrap." The outperformance of the parsimony "bootstrap" method over NJ "bootstrap" was not just an average effect but occurred in virtually all (77 of 80) simulations. Therefore, of the tree-based criteria tested, there is no reason to recommend any method other than parsimony "bootstrap." If this method were to be used in the analysis of real data with potential reticulate signal, the bootstrap support value required to hypothesize that a terminal taxon is monophyletic should be set very high to decrease the duration of phase 2 and, consequently, the chances that a reticulate taxon will be erroneously asserted to be monophyletic. However, based on the results of this study, even if 100% bootstrap support was required before claiming monophyly, type I error would still occur for data sampled during phases 1 and 2 (i.e., when gene flow has begun recently).

*Relative performance of network methods.*—The criterion devised to test the performance of the MSN method relied on the erosion of significant differences in the mean branch length observed between donor and recipient taxa and the mean branch length observed within the recipient taxon. Using this criterion, MSN performed poorly, similar to parsimony "best tree" (Fig. 6). Moreover, the response of the MSN method to reticulate gene flow was complicated by the possibility of four starting MSNs for the same simulated starting tree topology (Fig. 5). Based upon our results, and those of Cassens et al. (2005), who describe the poor performance of the MSN method for reconstructing relationships in nonreticulate data sets, it is unclear the degree to which the MSN method produces topologies that can be interpreted in a historically meaningful manner. Accordingly, we recommend against the use of MSNs alone for the analysis of phylogenetic data.

Performance of the NeighborNet method was measured by determining the bootstrap support level for the trivial split that contained only members of the recipient taxon. When bootstrap support dropped below 95%, we considered the null hypothesis to be correctly accepted. This criterion performed as well as the parsimony "bootstrap" method. The probability of type I error remains high using this criterion, but this drawback may, in some cases, be mitigated by the possibility of correctly identifying reticulate taxa at earlier time points (phase 1) using evidence of conflicting signal (discussed below).

*Relative performance of nonhierarchical methods.*—Of all methods used, $F_{st}$ performed the poorest, even when

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



using the generous criterion of inferring reticulation when $F_{st}$ dropped below 0.95 and including in the calculation only those taxa known to be hybridizing. When all taxa were included, $F_{st}$ did not deviate from 1.0 due to fixed differences between nonhybridizing taxa. Thus, $F_{st}$ does not appear to be an appropriate metric for identifying reticulation in data sets containing phylogenetic signal. This is not a criticism of a legitimate use of $F_{st}$ for describing genetic variation among tokogenetically related populations.

PCO-MC significantly outperformed all other methods, even when using a conservative smoothing parameter value ($R = 0.0401$) that biased results toward large TIRH values. Plots over time show clusters representing the hybridizing taxa to rapidly acquire similar principal coordinate values following the commencement of gene flow. For example, in Sim3 at $t = 23$ (Fig. 3), the first three principal coordinates for taxa A and D were nearly identical at a point when parsimony retained 100% bootstrap support for taxon A as the monophyletic sister taxon of taxon D, the distance along the minimum spanning network between taxa C or D and A remained large, and the NeighborNet split that distinguished taxon A from all other taxa received 100% bootstrap support.

The outperformance of PCO-MC over other methods may be due to collusion of two factors peculiar to data sampled across the phylogeny/tokogeny boundary. First, in the presence of recombination, an overall genetic distance measure, calculated as an average of distances contributed by many independent coalescent genes, may be a better indicator of historical relationships among individuals than synapomorphy. Second, the rigid, low-dimensional structure of bifurcating trees (and, to a lesser extent, networks) may prevent some associations among individuals from becoming apparent. When principal coordinates are calculated, such associations may more freely emerge in a large number of possible dimensions (or eigenvectors). Subsequent application of modal clustering permits many dimensions to be simultaneously scrutinized for statistical associations.

Further improvement of the PCO-MC method might be achieved by more accurate estimation of the smoothing parameter (R) used to evaluate cluster densities and membership. Although difficult to define precisely for any data set, R is a useful parameter because it permits control over the inevitable trade-off between type I and type II errors. Large R values promote type II error, whereas small R values promote type I error. Improvements in the PCO-MC method might also be realized when using real data sets, where natural variation within taxa should lead to a smoother "density landscape" than with the simulated data (where at least four of five clusters exhibited no internal variation). A smoother density landscape would permit the legitimate use of significance tests for cluster number available in PROC MODECLUS. When analyzing real data, a range of R values should be explored to determine the most stable cluster number and membership for a given data set.

Nonparametric modal clustering has a number of advantages over other nonhierarchical clustering methods. First, no assumptions about the underlying distribution of data are required; thus, nonuniform or nonnormal data (such as principal coordinates) may be used. Secondly, the expected number of clusters need not be defined a priori. This is essential when the goal is to identify the number of evolutionarily distinct groups indicated by a data set. Based on the promising results of this simple study further examination of the ability of ordination-based methods to correctly identify reticulate taxa in complex real and simulated data sets is warranted.

### *Evidence of Conflicting Signal*

NeighborNet and RSW have been proposed as methods for identifying conflicting signal in phylogenetic data sets caused by reticulate processes such as hybridization or intergenic recombination (Trueman, 1998; Bryant and Moulton, 2004). Network methods such as NeighborNet are now commonly used for data sets in which reticulation may be important, whereas RSW, although intuitively attractive, has been infrequently implemented (Nicholas and Trueman, 2002; Raxworthy et al., 2002; Abbott and Double, 2003). To our knowledge, neither method has yet been formally tested using appropriate simulated data with known reticulate history.

We have shown that, under particular topological models of gene flow (e.g., Sim2) during the early stages (i.e., corresponding roughly to phase 1) of an ongoing reticulate process, both RSW and NeighborNet were capable of producing correct indications of reticulate history by revealing conflicting phylogenetic signal (Fig. 7). When the possibility of correctly identifying reticulate taxa based on conflicting signal is coupled with its second-place performance ranking for TIRH, NeighborNet might be expected to have a lower cumulative probability of type I error than parsimony "bootstrap" for certain highly specific reticulate processes (e.g., those modeled by Sim2). Nevertheless, the regions of the simulation time course where conflicting signal criteria result in a correct inference do not overlap with the regions where correct inferences could be obtained using "erosion of type I error" criteria. During a period corresponding roughly to phase 2, NeighborNet analysis would still be expected to result in type I error.

When gene flow occurred between two distantly related taxa (e.g., Sim3), both NeighborNet and RSW were prone to type II error: identifying a taxon to have a reticulate history when in fact it was monophyletic. These results imply that, when methods that identify conflicting phylogenetic signal (such as NeighborNet and RSW) are applied to real reticulate data sets sampled from nature, the probability of type II error will be high whenever recent gene flow has occurred between taxa separated by two or more internodes in the underlying phylogenetic tree. Although a correct inference of reticulate history remained possible under topologically complex gene flow conditions, incorrect inferences were even more likely. Lastly, both methods were incapable of revealing correct conflicting signal when gene flow occurred between sister taxa (e.g., Sim1). Thus, conflicting signal criteria

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



were clearly affected by the topological complexity of gene flow, whereas "erosion of type I error" criteria were not.

Given that (1) gene flow between sister taxa may be common, (2) evidence of significant conflicting signal may be observed even when the history is not reticulate, and (3) a simple reticulate process (such as unidirectional gene flow between two distantly related taxa) may induce the indication of more generalized reticulation in some data sets, the utility of the conflicting signal revealed by network analyses is not entirely clear. Our findings support prior suggestions that split network methods may be primarily useful for preliminary examination of the degree to which a data set, taken in its entirety, supports a tree-like topology (Huson and Bryant, 2006). In other words, while network methods such as NeighborNet permit the visualization of the extent to which a data set contains conflicting signal (the hallmark of reticulate evolution), networks may not be universally useful for testing the hypothesis that a specific taxon within a data set has a reticulate history.

### *Generality of Findings*

As with all simulations, a number of limitations apply to the present study. The results are most applicable to cases of intraspecific gene flow between taxa with weak or nonexistent barriers to reproduction. The study is concerned with the simultaneous analysis of data from many unlinked loci as opposed to separate analysis of independent coalescent genes. The model of introgressive hybridization was highly simplified: it was haploid and considered only unidirectional gene flow between two taxa within five-taxon trees.

The reason for using this simplified model, however, was not to precisely mimic any particular process that might occur in nature but rather to generate reticulate data sets with a wide range of conflicting signal levels, representative of the range of conflicting signal that might be encountered in the analysis of real data (irrespective of the specifics of the process that generated it). The model was based in biological reality, producing data akin to what might be observed in a study of secondary contact between conjugating bacterial lineages using binary molecular markers such as RAPD, ISSR, AFLP, or SNP. We believe that this simplified model provides a foundation for predicting the consequences of unmodeled variables that may be important when analyzing more complex data sets from natural systems.

Several variables important in natural systems were not modeled in this simulation. These include population size, ploidy level, mutation rate, variation in gene flow rates, and the effect of bidirectional or multidirectional gene flow. An increase in population size is expected to greatly increase TIRH (Appendix 2). For realistic population sizes (e.g., $N > 1000$) and levels of starting genetic distance and gene flow such as those modeled here, TIRH may be so large (>10,000 hybrid events or >100,000 generations) that true reticulate history would seldom be revealed using bifurcating tree-based procedures that are inherently biased towards finding distinct (e.g., monophyletic) groups. In nature, fluctuation in population size over time may ameliorate the consequences of this bias. However, this finding is important because it suggests that reticulate history cannot be ruled out for many lower level plant taxa now presumed to be monophyletic based on strongly supported, tree-based phylogenetic analyses. The effect of analyzing data from diploid or polyploid organisms should be similar to the effect of increasing population size.

The effect of mutation is similar to that of gene flow in that it introduces novel character states into a population. However, as mutations occur within taxa and drift to fixation, new characters that distinguish (rather than homogenize) donor and recipient taxa arise. Thus the effect of mutation is to increase TIRH. As the rate at which mutations are fixed in a population approaches the rate at which introgressed characters are fixed, TIRH approaches infinity. Mutation rates are generally thought to be low in eukaryotes, on the order of $10^{-10}$ per nucleotide per replication (Drake et al., 1998), so the impact of mutation on TIRH for an ongoing hybrid process may be minimal. An increase in gene flow is expected to decrease TIRH. Neither variation in mutation rate nor gene flow rate seems likely to affect the rank order of relative performance for the methods examined.

In natural systems with structured populations, bidirectional or multidirectional gene flow is possible. With bidirectional gene flow, the time to homogeneity between the two hybridizing taxa is expected to decrease, so TIRH should also decrease. However, in cases where more than one internode separates the hybridizing taxa, the effect of bidirectional gene flow on tree topology is less predictable than in the unidirectional case. Bidirectional flow of character states would pull both of the taxa undergoing gene exchange away from their starting position in the tree simultaneously, forcing them towards one another by eroding character support for internal branches. A novel assemblage of apparent synapomorphies that support the inclusion of the two taxa exchanging genes in a single clade would emerge. Note that these apparent synapomorphies would appear due to gene flow, via the stochastic processes of drift and fixation, rather than the usual mechanism by which synapomorphies accumulate: mutation and common ancestry. Thus, in contrast to the unidirectional case, the position of this new clade in the tree is difficult to predict.

### *Recommendations*

When analyzing real data, it will often be difficult to know, a priori, if the data were derived from a reticulate process. Furthermore, because past hybridization events will not usually have been observed, it is not realistic to use real data to evaluate the performance of various analytical methods. As a surrogate, we undertook a blind reanalysis of our simulated data. We randomly sampled 200 data sets from the totality of non-trivial data sets (i.e., where donor and recipient taxa were not yet homogeneous). We then used PCO-MC, NeighborNet,

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



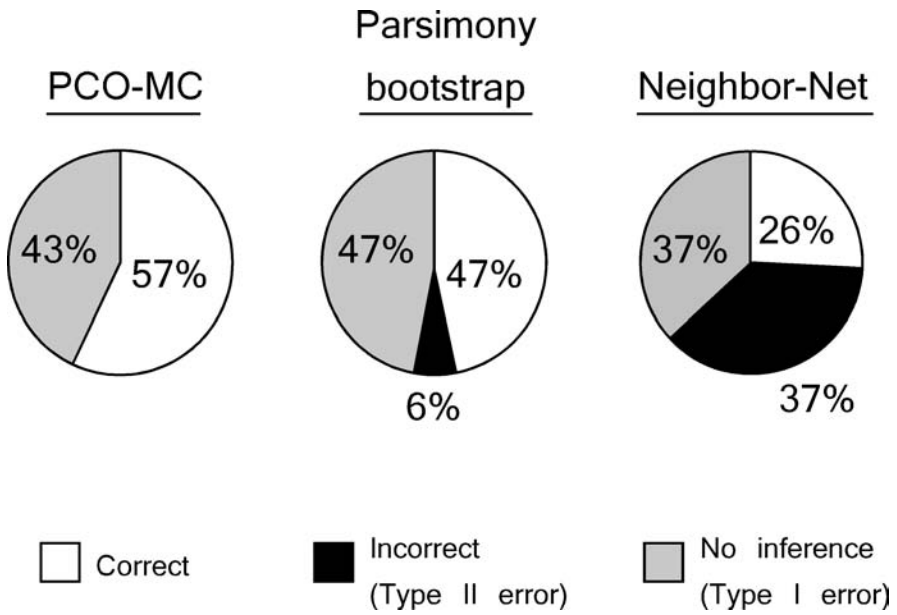



FIGURE 8. Relative performance of PCO-MC, parsimony "bootstrap," and NeighborNet methods in data sets sampled at random from all four models of simulated gene flow. A "correct" answer was obtained when the null hypothesis was correctly accepted for the reticulate taxon, and correctly rejected for all monophyletic taxa. An "incorrect" answer resulted from a failure to make a correct inference for all taxa included in the analysis. "No inference" of reticulate history was made when all taxa were inferred to be monophyletic.

and parsimony "bootstrap" methods to determine which taxa had a reticulate history and which were monophyletic using the "erosion of type I error" criteria described previously. Finally, we compared our inferences to the correct answer, which was known but hidden until after the inferences had been made.

Figure 8 shows that PCO-MC outperformed both parsimony "bootstrap" and NeighborNet. NeighborNet underperformed the other two methods primarily because of a high frequency of type II error. When cutoff criteria were altered to maximize the performance of each method (R = 2.0001 for PCO-MC, bootstrap cutoff = 100% for parsimony and NeighborNet), the rank order was the same. For PCO-MC, the frequency of correct answers, type II errors, and type I errors was 96%, 3.5%, and 0.5%, respectively. For parsimony bootstrap, these frequencies were 64.5%, 3%, and 32.5%, and for NeighborNet, 37.5%, 34.5%, and 28%. Furthermore, no advantage could be attained by combining methods over that achieved using PCO-MC alone. When data sets that resulted in no inference using PCO-MC were subjected to parsimony analysis, a correct result was obtained 7% of the time, but an incorrect result was obtained 6% of the time. When NeighborNet was likewise applied, a correct result was obtained 6% of the time, and an incorrect result 20% of the time. Therefore, provided that an appropriate R value can be identified, this study provides no evidence to recommend any of the tested methods other than PCO-MC for identifying reticulate taxa in phylogenetic data sets.

## CONCLUSIONS

This study demonstrates some of the pitfalls of applying bifurcating tree-based methods, appropriate for reconstructing higher level relationships among monophyletic terminal taxa, to lower level studies, where past hybridization may have lead to reticulate relationships. We show that taxa with reticulate histories may not be readily identifiable using tree-based methods because they are prone to type I error—biased towards finding a taxon to be monophyletic even when it has a substantial reticulate history. Moreover, we show that tree-based analysis of reticulate data may result in the inference of highly supported, erroneous relationships among taxa. These results challenge the conclusions of some studies that purport to reveal strongly supported phylogenetic relationships among taxa known to hybridize in nature. Our results suggest that the use of tree-based procedures alone to identify terminal monophyletic groups for the purpose of defining or delimiting species, or for prioritizing populations for conservation purposes, is questionable.

This study suggests that network methods such as NeighborNet, or tree-based methods such as RSW, which are designed to reveal conflicting phylogenetic signal in data sets, may (under certain, very restrictive conditions) be successfully used to identify reticulate histories. However, such methods are highly prone to type II error—biased towards finding that clearly monophyletic groups are a product of reticulate evolution. Nonhierarchical ordination procedures (such as PCO-MC) may prove to be more generally useful than tree-based or network methods for distinguishing terminal monophyletic groups from reticulate taxa in data sets that span the boundary between phylogeny and tokogeny.

## ACKNOWLEDGEMENTS

We thank Melissa Islam, Mark Simmons, David Tank, and Libing Zhang for helpful comments on a draft version of the manuscript.

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



## REFERENCES


Abbott, C. L., and M. C. Double. 2003. Phylogeography of shy and white-capped albatrosses inferred from mitochondrial DNA sequences: Implications for population history and taxonomy. Molecular Ecology 12:2747–2758.

Allaby, R. G., and T. A. Brown. 2003. AFLP data and the origins of domesticated crops. Genome 46:448–453.

Arnold, M. L. 1997. Natural hybridization and evolution. Oxford University Press, New York.

Avise, J. C., J. Arnold, R. M. Ball, E. Bermingham, T. Lamb, J. E. Neigel, C. A. Reeb, and N. C. Saunders. 1987. Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. Annu. Rev. Ecol. Syst. 18:489–522.

Bandelt, H.-J., V. Macaulay, and M. Richards. 2000. Median networks: Speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. Mol. Phylogenet. Evol. 16:8–28.

Baum, D. A., and K. L. Shaw. 1995. Genealogical perspectives on the species problem. Pages 289–303 *in* Experimental and molecular approaches to plant biosystematics (P. C. Hoch and A. G. Stephenson, eds.). Missouri Botanical Garden, St. Louis.

Bowe, L. M., G. Coat, and C. W. DePamphilis. 2000. Phylogeny of seed plants based on all three genomic compartments: Extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. Proc. Natl. Acad. Sci. USA 97:4092–4097.

Bremer, K., and H.-E. Wanntorp. 1979. Hierarchy and reticulation in systematics. Syst. Zool. 28:624–627.

Bryant, D., and Moulton, V. 2004. NeighborNet: An agglomerative algorithm for the construction of phylogenetic networks. Mol. Biol. Evol. 21:255–265.

Cassens, I., P. Mardulyn, and M. C. Milinkovitch. 2005. Evaluating intraspecific "network" construction methods using simulated sequence data: Do existing algorithms outperform the global maximum parsimony approach? Syst. Biol. 54:363–372.

Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, R. A. Price, H. G. Hills, Y. -L. Qiu, K. A. Kron, J. H. Rettig, E. Conti, J. D. Palmer, J. R. Manhart, K. J. Sytsma, H. J. Michaels, W. J. Kress, K. G. Karol, W. D. Clark, M. Hedren, B. S. Gaut, R. K. Jansen, K.-J. Kim, C. F. Wimpee, J. F. Smith, G. R. Furnier, S. H. Strauss, Q.-Y. Xiang, G. M. Plunkett, P. S. Soltis, S. M. Swensen, S. E. Williams, P. A. Gadek, C. J. Quinn, L. E. Eguiarte, E. Golenberg, G. H. Learn, S. Graham, S. C. H. Barrett, S. Dayanandan, and V. A. Albert. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbc*L. Ann. Missouri Bot. Garden 80:528–580.

Chaw, S.-M., C. L. Parkinson, Y. Cheng, T. M. Vincent, and J. D. Palmer. 2000. Seed plant phylogeny inferred from all three genomes: Monophyly of extant gymnosperms and origin of Gnetales and conifers. Proc. Natl. Acad. Sci. USA 97:4086–4091.

Crawford, D., and M. J. E. Mort. 2004. Single-locus molecular markers for inferring relationships at lower taxonomic levels: Observations and comments. Taxon 53:631–635.

Davis, J. I., and K. C. Nixon. 1992. Populations, genetic variation, and the delimitation of phylogenetic species. Syst. Biol. 41:421–435.

De Queiroz, K., and M. J. Donoghue. 1988. Phylogenetic systematics and the species problem. Cladistics 4:317–338.

Debry, R. W., and R. G. Olmstead. 2000. A simulation study of reduced tree-search effort in bootstrap resampling analysis. Syst. Biol. 49:171–179.

DeSalle, R., and G. Amato. 2004. The expansion of conservation genetics. Nat. Rev. Genet. 5:702–712.

Doyle, J. J. 1995. The irrelevance of allele tree topologies for species delimitation, and a non-topological alternative. Syst. Bot. 20:574–588.

Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow. 1998. Rates of spontaneous mutation. Genetics 148:1667–1686.

Ellstrand, N. C., R. Whitkus, and L. H. Rieseberg. 1996. Distribution of spontaneous plant hybrids. Proc. Natl. Acad. Sci. USA 93:5090–5093.

Excoffier, L., and P. E. Smouse. 1994. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: Molecular variance parsimony. Genetics 136:343–359.

Fitch, W. M. 1997. Networks and viral evolution. J. Mol. Evol. 44(Suppl.): S65–S75.

Funk, V. A. 1985. Phylogenetic patterns and hybridization. Ann. Missouri Bot. Garden 72:681–715.

Graham, S. W., and R. G. Olmstead. 2000. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. Am. J. Bot. 87:1712–1730.

Grant, V. E. 1971. Plant speciation. Columbia University Press, New York.

Goldstein, P. Z., and R. DeSalle. 2000. Phylogenetic species, nested hierarchies, and character fixation. Cladistics 16:364–384.

Goldstein, P. Z., R. DeSalle, G. Amato, and A. P. Vogler. 2000. Conservation genetics at the species boundary. Conserv. Biol. 14:120–131.

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

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

Humphries, C. J. 1983. Primary data in hybrid analysis. Pages 89–103 *in* Advances in cladistics: Proceedings of the Second Meeting of the Willi Hennig Society (N. I. Platnick and V. A. Funk, eds.). Columbia University Press, New York.

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

Jaccard, P. 1908. Nouvelles rescherches sur la distribution florale. Bulletin de la Société Vaudoise des Sciences Naturelles 44:223–270.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism, III (H. N. Munro, ed.). Academic Press, New York.

Knobloch, I. W. 1972. Intergeneric hybridization in flowering plants. Taxon 21:97–103.

Landry, P. A., and F.-J. LaPointe. 1996. RAPD problems in phylogenetics. Zool. Scripta 25:283–290.

Legendre, P. (ed.) 2000. Special section on reticulate evolution. J. Classif. 17:153–195.

Legendre, P., and V. Makarenkov. 2002. Reconstruction of biogeographic and evolutionary networks using reticulograms. Syst. Biol. 51:199–216.

Lewis, P. O., and D. Zaykin. 2002. GDA: Genetic data analysis. Software distributed by the authors.

Mathews, S., and M. J. Donoghue. 1999. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. Science 282:947–950.

McDade, L. 1990. Hybrids and phylogenetic systematics. I. Patterns of character expression in hybrids and their implications for cladistic analysis. Evolution 44:1685–1700.

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

McDade, L. 1997. Hybrids and phylogenetic systematics. III. Comparison with distance methods. Syst. Bot. 22:669–683.

Nelson, G., and N. I. Platnick. 1980. Multiple branching in cladograms: Two interpretations. Syst. Zool. 29:86–91.

Nicholas, W. L., and J. W. H. Trueman. 2002. The taxonomy of the family Xyalidae Chitwood, 1951 (Monhysterida: Nematoda): A cladistic analysis. Nematology 4:453–470.

Nixon, K. C., and Q. D. Wheeler. 1990. An amplification of the phylogenetic species concept. Cladistics 6:211–223.

Qiu, Y.-L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, and M. W. Chase. 1999. The earliest angiosperms: Evidence from mitochondrial, plastid and nuclear genomes. Nature 402:404–407.

Posada, D., and K. A. Crandall. 2001. Intraspecific gene genealogies: Trees grafting into networks. Trends Ecol. Evol. 16:37–45.

Posada, D., and K. A. Crandall. 2002. The effect of recombination on the accuracy of phylogeny estimation. J. Mol. Evol. 54:396–402.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Raxworthy, C. J., M. R. J. Forstner, and R. A. Nussbaum. 2002. Chameleon radiation by oceanic dispersal. Nature 415:784–787.

Rieseberg, L. H., and N. C. Ellstrand. 1993. What can molecular and morphological markers tell us about plant hybridization? Crit. Rev. Plant Sci. 12:213–241.

Russello, M. A., and G. Amato. 2004. A molecular phylogeny of *Amazona*: Implications for Neotropical parrot biogeography, taxonomy, and conservation. Mol. Phylogenet. Evol. 30:421–437.

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



Sarle, W. S., and A.-H. Kuo. 1993. The MODECLUS procedure. SAS Technical Report P-256. SAS Institute, Cary, North Carolina.
Schaal, B. A., and W. J. Leverich. 2001. Plant population biology and systematics. Taxon 50:679–695.
Schierup, M. H., and J. Hein. 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics 156:879–891.
Schneider, S., D. Roessli, and L. Excoffier. 2000. Arlequin: A software for population genetics data analysis, version 2.0.1.1. Genetics and Biometry Lab. Department of Anthropology, University of Geneva.
Scotland, R. W., R. G. Olmstead, and J. R. Bennett. 2003. Phylogeny reconstruction: The role of morphology. Syst. Biol. 52:539–548.
Sites, J. W., and J. C. Marshall. 2004. Operational criteria for delimiting species. Annu. Rev. Ecol. Evol. Syst. 35:199–227.
Sneath, P. H. A. 1975. Cladistic representation of reticulate evolution. Syst. Zool. 24:360–368.
Sneath, P. H. A., and R. R. Sokal. 1973. Numerical taxonomy; the principles and practice of numerical classification. W. H. Freeman, San Francisco.
Stebbins, G. L. 1950. Variation and evolution in plants. Columbia University Press, New York.
Stebbins, G. L. 1959. The role of hybridization in evolution. Proceedings of the American Philosophical Society 103:231–251.
Stuessy, T. F., K. Tremetsberger, A. N. Mullner, J. Jankowics, Y. P. Guo, C. M. Baeza, and R. M. Samuel. 2003. The melding of systematics and biogeography through investigations at the populational level: examples from the genus *Hypochaeris* (Asteraceae). Basic Appl. Ecol. 4:287–296.
Swofford, D. L. 1999. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.
Templeton, A. R., K. A. Crandall, and C. F. Sing. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics 132:619–633.
Trueman, J. W. H. 1998. Reverse successive weighting. Syst. Biol. 47:733–737.
Trueman, J. W. H., and M. Gibbs. 2002. RSW1.1. Software distributed by the authors.
Weir, B. S., and C. C. Cockerham. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370.
Wheeler, Q. D., and K. C. Nixon. 1990. Another way of looking at the species problem: A reply to de Queiroz and Donoghue. Cladistics 6:77–81.
Xu, S. 2000. Phylogenetic analysis under reticulate evolution. Mol. Biol. Evol. 17:897–907.

*First submitted 6 June 2006; reviews returned 12 September 2006; final acceptance 27 November 2006*
*Associate Editor: Adrian Paterson*

## APPENDIX 1

Formulas for calculating the probability of successful and erroneous inferences of reticulate history based on conflicting signal levels for NeighborNet and RSW analyses are presented. Three outcomes were possible: a correct inference of reticulate history, an incorrect inference of reticulate history, and no inference of reticulate history. The former two outcomes were nonexclusive, that is, both correct and incorrect inferences could be observed for any one data set. Therefore, at any particular point during the time course of the simulation,

$$\Pr(\text{success}) + \Pr(\text{failure}) + \Pr(\text{no inference}) = 1 \quad (1)$$

where "Pr" indicates "probability."

Let $N$ = the total number of data sets in which no inference of reticulate history was obtained for a particular time point. Given that 20 different data sets were analyzed for each time point:

$$\Pr(\text{no inference}) = N/20. \quad (2)$$

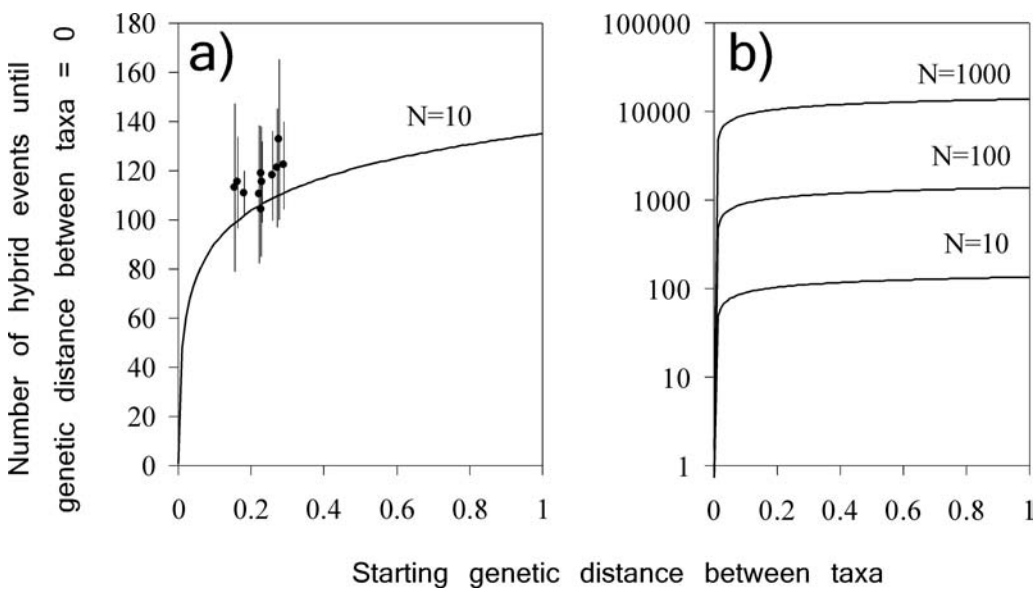



FIGURE A1. (a) Effect of starting genetic distance on the time required for two hybridizing taxa to become genetically indistinguishable. Curve indicates theoretical time to homogeneity for population size N = 10. Points show mean values (±SD) observed in this study. (b) Effect of population size on theoretical time to homogeneity. Axes are the same as (a), except that the $y$-axis is in $\log_{10}$ scale.

Downloaded By: [USYB - Systematic Biology] At: 21:43 30 April 2007



Accordingly, the probability that a run yields an inference, Pr(inference) $= 1 - (N/20)$.

Let C = the total number of correct inferences and I = the total number of incorrect inferences summed across all 20 data sets analyzed at a particular time point. The maximum value for C = 20 because there is only one correct inference possible per data set. The maximum value for I > 20 because multiple incorrect inferences are possible. The Pr(success) = Pr(inference) multiplied by the probability of sampling a correct inference from the totality of inferences, or:

$$\Pr(\text{success}) = [1 - (N/20)] \times [C/(C + I)]. \quad (3)$$

Likewise:

$$\Pr(\text{failure}) = [1 - (N/20)] \times [I/(C + I)]. \quad (4)$$

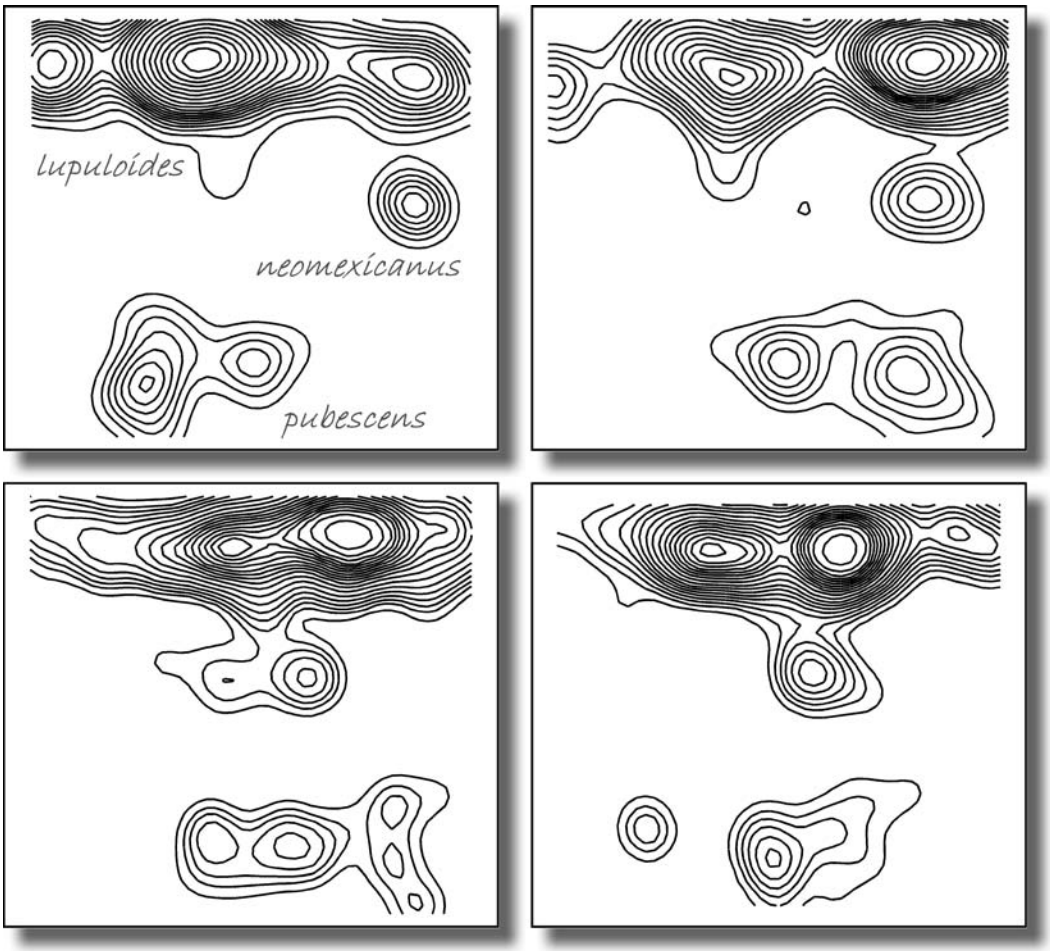
Therefore, the expansion of equation (1) is:

$$[1 - (N/20)] \times [C/(C + I)] + [1 - (N/20)] \times [I/(C + I)] + N/20 = 1. \quad (5)$$

By factoring $[1 - (N/20)]$ from the first two terms, Equation (5) is easily shown to reduce to $1 = 1$.

## APPENDIX 2

In the present haploid model, the time required for two hybridizing taxa to become genetically identical can be approximated by the following equation: $d[D,R]_t \approx d[D,R]_{t-1} - (0.5)(d[D,R]_{t-1})(1/N)$, where $d[D,R]_t$ and $d[D,R]_{t-1}$ are the genetic distances between the donor and recipient taxa at time t and t-1 (measured as the number of hybridization events since gene flow began), the proportion of donor parental character states present in F1 hybrid progeny is 0.5, and the proportion of introgressed alleles expected to become fixed via genetic drift is 1/N. This equation assumes that a single F1 hybrid individual is produced per gene flow event, that all character states are neutral, and that genetic drift until fixation has occurred in the recipient taxon prior to data set sampling. The time required for hybridizing taxa to become identical increases logarithmically with starting genetic distance (Fig. A1a). Because the simulation used a fixed number of generations (10) of random mating, most (but not all) of the introgressed characters either were fixed or lost prior to data set sampling. Therefore, as shown in Figure A1a, values observed in this study were slightly larger than the values predicted by the formula. Figure A1b shows the relationship between population size and the number of hybridization events necessary for two populations to become genetically identical under this model. For realistic population sizes, many thousands of hybridization events will be required before two interbreeding taxa will no longer be recognized as distinct lineages using most of the analytical methods considered in this study.




The reticulate evolutionary history of native North American hops varieties (*Humulus lupulus* vars. *lupuloides*, *neomexicanus*, and *pubescens*) visualized as a density landscape of principal coordinate data. Each pair of plotted principal coordinates provides a slightly different perspective on relationships among potentially interbreeding populations, which appear as "peaks" in the images.