vol. 174, no. 2 the american naturalist august 2009

E-Article

A Statistical Approach for Distinguishing Hybridization and Incomplete Lineage Sorting

Simon Joly,* Patricia A. McLenachan, and Peter J. Lockhart

Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

Submitted October 10, 2008; Accepted February 12, 2009; Electronically published June 11, 2009

abstract: The extent and evolutionary significance of hybridization is difficult to evaluate because of the difficulty in distinguishing hybridization from incomplete lineage sorting. Here we present a novel parametric approach for statistically distinguishing hybridization from incomplete lineage sorting based on minimum genetic distances of a nonrecombining locus. It is based on the idea that the expected minimum genetic distance between sequences from two species is smaller for some hybridization events than for incomplete lineage sorting scenarios. When applied to empirical data sets, distributions can be generated for the minimum interspecies distances expected under incomplete lineage sorting using coalescent simulations. If the observed distance between sequences from two species is smaller than its predicted distribution, incomplete lineage sorting can be rejected and hybridization inferred. We demonstrate the power of the method using simulations and illustrate its application on New Zealand alpine buttercups (*Ranunculus*). The method is robust and complements existing approaches. Thus it should allow biologists to assess with greater accuracy the importance of hybridization in evolution.

Keywords: coalescence theory, hybridization, incomplete lineage sorting, nonmonophyletic species, predictive posterior distribution, *Ranunculus*.

Introduction

Hybridization is a feature of plant and animal evolution (Anderson 1949; Stebbins 1959; Grant 1981; Arnold 1997; Rieseberg 1997; Barton 2001). While the extent of its evolutionary significance remains controversial (Seehausen 2004), it is increasingly seen as an important process for generating biotic diversity (Arnold 1997; Ferguson and Sang 2001; Rieseberg et al. 2003) and for rapid adaptation (Grant and Grant 1996; Ellstrand and Schierenbeck 2000; Arnold 2004). Hybridization is often inferred from incongruence among gene trees from independent loci (Linder and Rieseberg 2004) or from trees in which species are not monophyletic (Funk and Omland 2003). However, because incomplete lineage sorting and gene duplication also produce these gene tree features, elimination of these potential causes is necessary before hybridization can be accepted as a reasonable explanation for the evolution of the data.

Where comparison of paralogs (resulting from gene duplication) rather than orthologs is the cause of incongruence among gene trees or nonmonophyly of species in gene trees, gene duplication can in general be readily detected from phylogenetic analyses with adequate sampling (Small et al. 2004). More difficult to distinguish is hybridization from incomplete lineage sorting, and this problem has attracted much recent interest (Wang et al. 1997; Sang and Zhong 2000; Holder et al. 2001; Machado et al. 2002; Huson et al. 2005; Buckley et al. 2006; Holland et al. 2008). Yet no effective and widely applicable approach exists for distinguishing these processes.

In this article, we introduce a parametric method for distinguishing hybridization from incomplete lineage sorting. On the basis of the observations that hybridization and incomplete lineage sorting make different predictions for gene phylogenies (Holder et al. 2001), we propose a test statistic based on minimum sequence distances between species and assess the power of the method using simulations. The method is then applied to examine instances of species nonmonophyly in gene trees of the New Zealand alpine buttercups (*Ranunculus* L.).

A Statistical Approach for Identifying Hybridization Events

Hybridization and incomplete lineage sorting make different predictions regarding the topologies and branch lengths of gene trees that evolve in accordance with the underlying species phylogeny (Holder et al. 2001). Consider an example of nonmonophyly, where a sequence from one species is more similar to sequences of another species than to those of its own. If explained by lineage

* Corresponding author. Present address: Department of Biology, McGill University, 1205 Docteur Penfield, Montréal, Québec H3A 1B1, Canada; e-mail: simon.joly@mail.mcgill.ca.

Am. Nat. 2009. Vol. 174, pp. E54–E70.
DOI: 10.1086/600082

sorting, the similar sequences will have coalesced before the divergence of the two species (looking forward in time). Therefore, under incomplete lineage sorting, the time elapsed since the speciation event represents a lower bound for the minimum divergence time between these sequences (fig. 1). If explained by hybridization, the similar sequences from different species could coalesce either before or after the species divergence (fig. 1). Where coalescence occurs before species divergence, there will be similar expectations for the minimum divergence time between sequences under hybridization and lineage sorting scenarios. However, in the case of hybridization and where coalescence of the sequences occurs subsequent to species divergence, the expectation for the minimum divergence time between the sequences will be smaller than that under lineage sorting (Joly et al. 2006; fig. 1*C*).

This suggests that the genetic distance between two sequences from different species can be used as a test statistic to distinguish hybridization from lineage sorting. The distribution of this statistic, under the null hypothesis that incomplete lineage sorting is a sufficient explanation of nonmonophyly, can be obtained through simulation using the coalescent theory with no migration (Kingman 1982*a*, 1982*b*). The observed distance can then be compared with the null distribution to determine whether we could reject the null hypothesis of incomplete lineage sorting. If the observed distance is smaller than $100(1 - \alpha)\%$ of the distances derived from coalescent simulations, where α is the predetermined Type I error, the null hypothesis can be rejected and hybridization inferred.

The distribution of minimum sequence distances expected under incomplete lineage sorting can be obtained in the following way with empirical data sets: (1) estimate population sizes and divergence times for all branches of the species phylogeny (which could be inferred or assumed); (2) simulate gene trees using the coalescent with no migration on the species tree; (3) for these trees, assume an optimal nucleotide substitution model and simulate character matrices that have the same characteristics as the original data set (sequence length, number of sequences per species); and (4) calculate the minimum distance between any sequences for every pair of species in the simulated data sets to calculate null distributions for the test statistic.

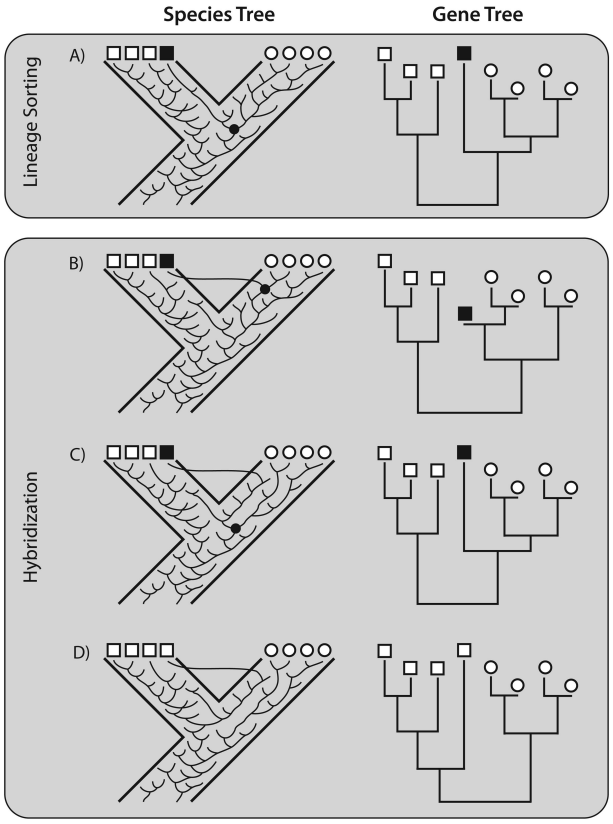

Figure 1: Examples of possible gene tree scenarios expected under lineage sorting (*A*) and hybridization (*B–D*) for two species. Sequences sampled from the two species are distinguished by squares and circles. The solid square represents a sequence that is more closely related to the sequences of the other species, causing the species to be nonmonophyletic. In the lineage sorting scenario (*A*), the incongruent sequence will always coalesce (at the position of the solid circles) with sequences of the other species before the speciation event (looking forward in time). In the hybridization scenario, it could coalesce after (*B*) or before (*C*) the speciation event. Hybridization need not always lead to nonmonophyletic gene trees (*D*), in which case it will go undetected.

Simulations

The power of the method for detecting hybridization events was assessed using simulations. These simulations assumed known population sizes and divergence times and do not incorporate the uncertainty associated with their estimation, a topic outside the scope of this article. Instead, we evaluated the performance of the method for identifying hybridization events across a wide range of parameter values.

A simple framework with no hybridization was first considered, in which two species, A and B, diverged $t_1 + t_2$ generations ago (fig. 2). The minimum distance between eight sequences of species A and eight sequences of species B was calculated for 1,000 sequence data sets simulated using the coalescent with no migration. This constituted the null distribution of minimum distances expected under an incomplete lineage sorting scenario.

A second set of sequences was simulated under a hybridization scenario wherein a hybridization event occurred t_1 generations in the past and t_2 generations after the divergence of the two species (fig. 2). In effect, the

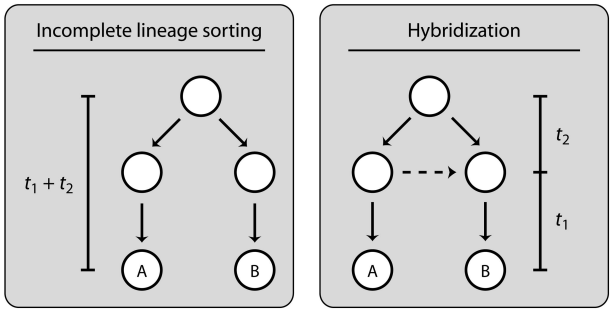

Figure 2: Simulation settings for testing the power of the method. Null distributions for minimum distances between two species were obtained by simulating sequences using the coalescent under an incomplete lineage sorting scenario. To simulate hybridization, a sequence is transferred from one population to another t_2 generations after the speciation event and then allowed to survive for t_1 generations until the present.

hybridization event transferred one sequence from species A to species B, and the sequence was allowed to persist until the present. Because we were interested only in determining whether a hybrid sequence could be identified as such, we did not consider the probability that the hybrid sequence becomes extinct in species B as a consequence of genetic drift. In this case, the simulation scheme can be simplified by simulating one sequence from species B and eight from species A, assuming that the species have diverged t_1 generations ago. One thousand sequence data sets were simulated, and the minimum distance between sequences of species A and B was calculated for each. The minimum distance for each replicate was compared with the null distribution obtained under an incomplete lineage sorting scenario, and a hybridization event was assumed to be correctly identified when the value obtained was smaller than $100(1 - \alpha)\%$ of the values obtained under incomplete lineage sorting, where α is the predetermined Type I error. A Type I error level of 5% was assumed for all the simulations.

Simulations were performed in MCMCcoal (ver. 1.2; Rannala and Yang 2003) with a Jukes and Cantor (1969) DNA substitution model. The population size parameter used in the simulations was $\theta = 4N_E\mu$, whereas time was scaled in values of $\tau = T\mu$, where T is time in number of generations. Simulations were performed for τ values that range from 0 to 0.05 for t_1 and t_2, which, according to a hypothetical mutation rate (μ) of 1×10^{-9} mutations per site per generation, would correspond to a range of 0–50 million generations. Simulations were also performed for θ values of 0.00001, 0.001, 0.01, and 0.1, which were kept constant across the phylogeny. Given $\mu = 1 \times 10^{-9}$, this would correspond to effective population sizes of 2,500, 250,000, 2,500,000, and 25,000,000 individuals, respectively. Finally, simulations were performed for three different sequence lengths (500, 1,000, and 5,000 base pairs [bp]).

Preliminary results showed that simulating 1,000 data sets represented a good compromise between length of computing time and variance of the estimates among independent simulations (data not shown). The same simulations were also performed with 16 and four sequences sampled per species. Results obtained with these different sampling sizes were highly correlated (Pearson's $r > 0.95$) and did not differ significantly (Wilcoxon signed rank test; $P > .05$) from the simulations presented here with eight sequences sampled per species.

Although the simulation framework is simple, it does account for more complex scenarios possible with larger species phylogenies (i.e., where more lineages diverge before or after the hybridization event). Effectively, adding splits before or after the hybridization event does not alter the results because it does not affect the probability of coalescence of two alleles within a species or between the two species involved in a hybridization event (data not shown). The only parameters of importance are the time since the divergence of the two species, the ancestral population size, and the population sizes of the two species following the speciation event.

Alpine *Ranunculus* of New Zealand

As an application for the method, we tested whether hybridization could explain the nonmonophyly of species in gene trees (see Lockhart et al. 2001) of a morphologically and ecologically diverse group of New Zealand alpine buttercups (*Ranunculus* L.). For the sake of brevity, we report most methods and detailed results that pertain to the *Ranunculus* example in the appendix. Here we describe only the main features of our analysis, and below we describe the major findings. Fourteen individuals from six well-defined species were sequenced for five chloroplast regions (*trn*C-*trn*D, *trn*L-*trn*F, *psb*A-*trn*H, *trn*D-*trn*T, *rp*L16), which were concatenated in further analyses. Sequences from the internal transcribed spacer (nrITS) region (127 individuals sequenced) and the J_{SA} chloroplast region (122 individuals) were also used to estimate the species tree and the population size and branch length parameters used in the simulations.

A species tree for these species was reconstructed by gene tree parsimony using parsimony consensus trees from all three data sets. Divergence times ($\tau = T \times \mu$) and population sizes ($\theta = 4N_E\mu$) for branches of the species tree were then estimated with the Bayesian method implemented in MCMCcoal (ver. 1.2; Rannala and Yang 2003; Burgess and Yang 2008) from the DNA sequences of the three data sets. The analyses were performed with various

prior distributions for each parameter to investigate the impact of prior choice on the results. Posterior predictive distributions for the test statistic under a null hypothesis of lineage sorting (no hybridization) were obtained through coalescent simulations. For each data set (*i*), 10,000 gene trees were simulated on the species tree in MCMCcoal, selecting different parameters τ_i, θ_i, r_i (the locus relative mutation rate), and h_i (heredity scalar) for each gene tree according to their marginal posterior distributions. Character matrices of the same length as the originals were simulated independently for each data set using the nucleotide substitution model that best fitted the original data. The shortest Hamming distance between any two sequences for all pairs of species was calculated for all replicates for the three data sets; these collections of values constituted the posterior predictive distributions.

Results

Simulations

The simulations showed that the method does not have an inflated Type I error, which is the probability of rejecting the null hypothesis when it is true. Over all simulations, the null hypothesis was rejected 2.8% ± 1.6% of the time when there was no hybridization event (i.e., when $t_2 = 0$; fig. 3, *bottom row*). The power of the test varied for different values of t_1 and t_2 but also according to sequence length and population sizes (fig. 3). For all simulation scenarios, larger t_2 and smaller t_1 values resulted in greater power to detect hybridization events. When the hybridization event occurred rapidly after the speciation event (small t_2), hybridization events were more easily detected when the time since the hybridization event was short (small t_1). However, when the time between the speciation event and the hybridization event (t_2) was large, the time elapsed since the hybridization event (t_1) did not have a strong effect on the power to detect hybridization events.

Simulations also showed that the shorter the sequence length, the harder it was to detect hybridization. This was expected since the variance associated with estimates of substitution number is more important for shorter sequences (Edwards and Beerli 2000). Consequently, the difference in minimum distances between species simulated under incomplete lineage sorting and hybridization scenarios is not as distinct when short sequences are considered. The power of the test was effectively best when sequences had a length of 5,000 bp, at which value the observed distances approach expected values that would be obtained with sequences of infinite length. The population size ($\theta = 4N_E\mu$) also affected the power of the test where larger population sizes resulted in reduced power. Indeed, a greater number of ancestral polymorphisms is expected to be maintained in larger populations, which means that a hybrid sequence is more likely to coalesce with the other sequences before the speciation event, making the hybridization event undetectable (i.e., fig. 1*C*, 1*D*). Yet the power of the test was strongly affected only for very large population sizes ($\theta = 0.1$). Such a population size parameter is unrealistic for most plant and animal species because it means that the expected proportion of different sites between sequences drawn at random from the species is 0.1, an unrealistic value (if θ were of this size, a realistic mutation rate [μ] of 1×10^{-9} would imply an effective population size of 25 million diploid individuals).

In general, the simulations suggest that the method is relatively powerful unless sequences are very short or the hybridization event has occurred very rapidly after the speciation event. Also, because the method does not have an inflated Type I error, one can be confident that hybridization events detected by the method are correct, with a probability α of making a wrong decision.

Hybridization in New Zealand Alpine Ranunculus

The concatenated chloroplast data set (4,135 bp) shows evidence of nonmonophyly for some species (fig. 4), a pattern also observed in the J_{SA} (480 bp) and the nrITS (603 bp) gene trees (appendix). A single species tree was obtained by gene tree parsimony (fig. 5) with a score of 27 deep coalescences when rooted with *Ranunculus acaulis*. The mean estimates for divergence time (τ) and population size (θ) parameters are summarized on the species tree (fig. 5; for more details, see the appendix). Although population size estimates were somewhat affected by the choice of prior (appendix), the general conclusions were unaffected.

The posterior predictive distributions generated for the concatenated chloroplast data set under a model of lineage sorting showed that six empirical distances were smaller than 95% of the predicted values (table 1), a result that holds irrespective of the chosen priors (data not shown). Moreover, all tests remained significant when accounting for simultaneous statistical testing by fixing a false discovery rate (Benjamini and Hochberg 1995) such that at most one of the six significant tests was falsely rejected. All significant distances involved individuals from the same chloroplast lineage from breeding group 2 (containing *Ranunculus crithmifolius* from Mount Lyndon and a *Ranunculus insignis* from Mount Hutt) and species from breeding group 1 (*Ranunculus haastii*, *Ranunculus lyallii*, and *Ranunculus sericophyllus*). Hybridization is thus a likely hypothesis for the chloroplast lineage present in *R. crithmifolius* from Mount Lyndon and in *R. insignis* from

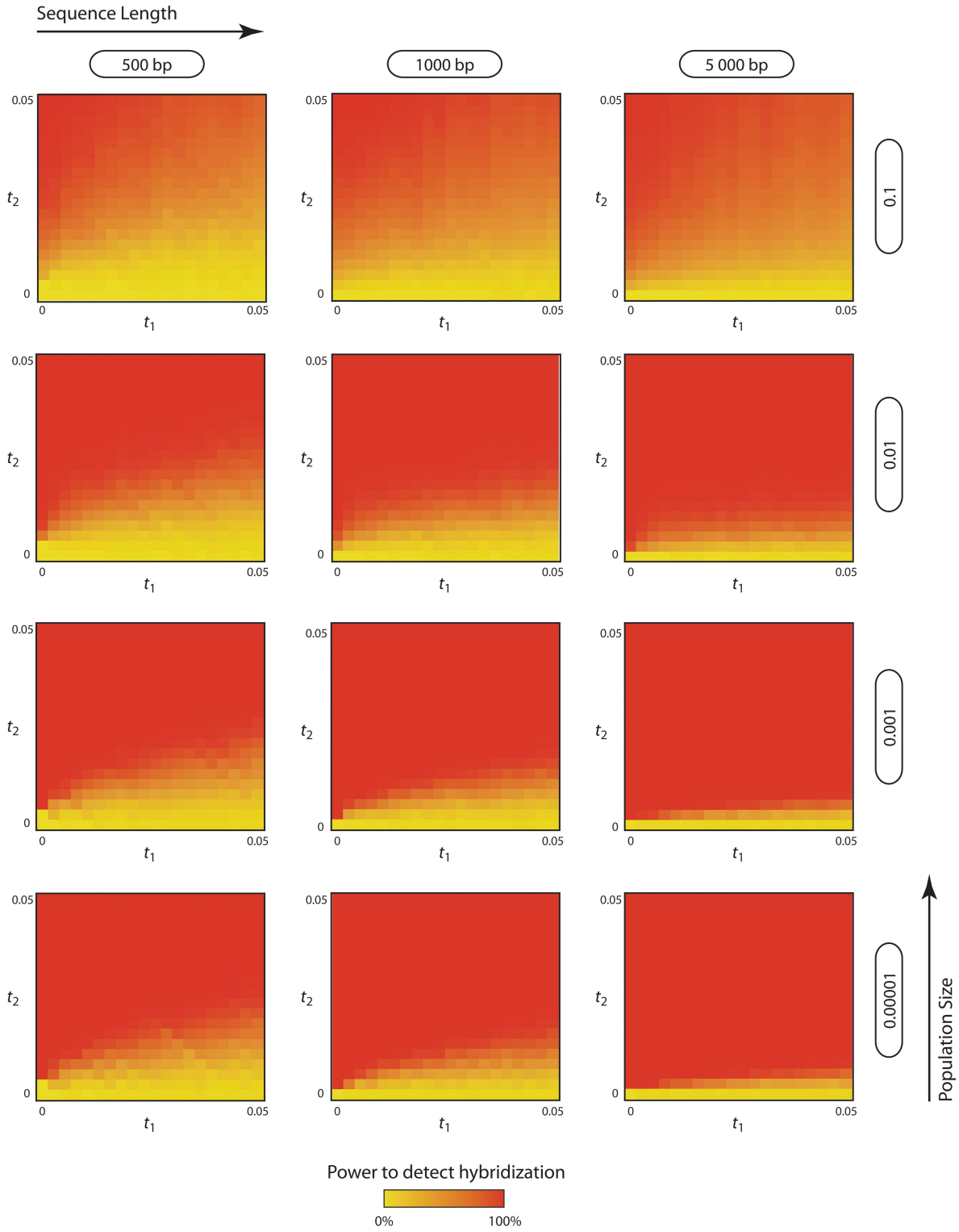

Figure 3: Power to detect hybridization under different parameters as determined by simulations. The population size parameter is in units of $\theta = 4N_E\mu$, where N_E is the effective population size and μ is the mutation rate; t_1 and t_2 are in units of $\tau = T\mu$, where T is the number of generations.

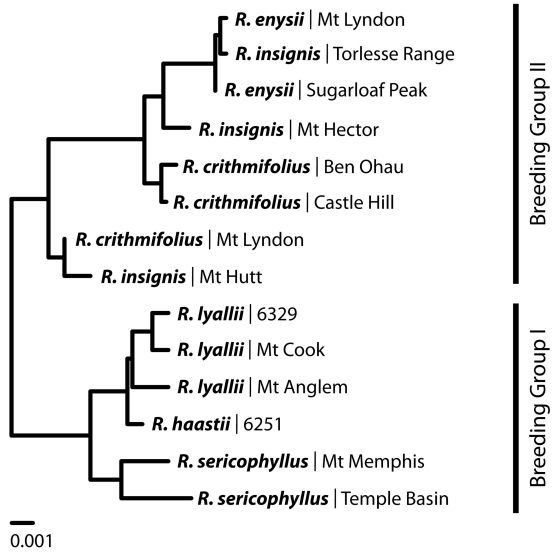

Figure 4: Chloroplast phylogeny obtained by maximum likelihood (ML) from five noncoding regions ($\ln L = 6{,}505.459$). The maximum parsimony search found two optimal trees (length = 133, consistency index = 0.917, retention index = 0.878), one of which was identical to the ML tree and the other differing only by the relative positions of *Ranunculus enysii* from Sugarloaf Peak and *Ranunculus insignis* from Torlesse Range, which are interchanged.

Mount Hutt. Moreover, because the significant distances associated with individuals from this lineage involved all species from breeding group 1, the inferred hybridization event has probably occurred between the common ancestor of *R. crithmifolius* and *R. insignis* and a common ancestor of species from breeding group 1 (*R. haastii*, *R. lyallii*, and *R. sericophyllus*), or between 3 and 4 million years ago according to the timescale of figure 5.

Further cases of nonmonophyly involving *Ranunculus enysii* and *R. insignis* were evident in the concatenated chloroplast phylogeny (fig. 4), but incomplete lineage sorting could not be rejected as a sufficient explanation ($P > .132$). Similarly, incomplete lineage sorting could not be rejected as the cause of the nonmonophyly observed in the nrITS and J_{SA} gene trees.

Discussion

Without a statistical framework to distinguish the effects of incomplete lineage sorting from hybridization, observations of phylogenetic incongruence and nonmonophyly are difficult to interpret (Holland et al. 2008). This point has particular relevance for efforts in reconstructing the phylogeny of hybrid species (Legendre and Makarenkov 2002; Huson et al. 2005; Jin et al. 2006) or in evaluating the extent (Bordewich et al. 2007) and consequences (Seehausen 2004) of hybridization in evolution. Several methods have been proposed to distinguish hybridization from incomplete lineage sorting. Some of these are based on a test statistic that is calculated directly from a phylogenetic tree. One such method was proposed by Sang and Zhong (2000). They tried to identify hybridization events from incongruent gene trees on the basis of the idea that the divergence time from the most recent common ancestor to the tips of two parental lineages should be the same across genes under a hybridization hypothesis and different under a lineage sorting scenario. But their test did not account for variance in coalescence time between lineages in the ancestral population (Edwards and Beerli 2000). When such variance is incorporated, their test becomes less powerful (Holder et al. 2001). Huson et al. (2005) have proposed another statistical test based on the topology of several gene trees. It makes use of the idea that if $(ab)c$ is the species tree, then topologies $(ac)b$ and $(bc)a$ are expected to be found in equal frequencies in a set of gene trees under a strict lineage sorting scenario. In the case of hybridization, one of these alternative topologies

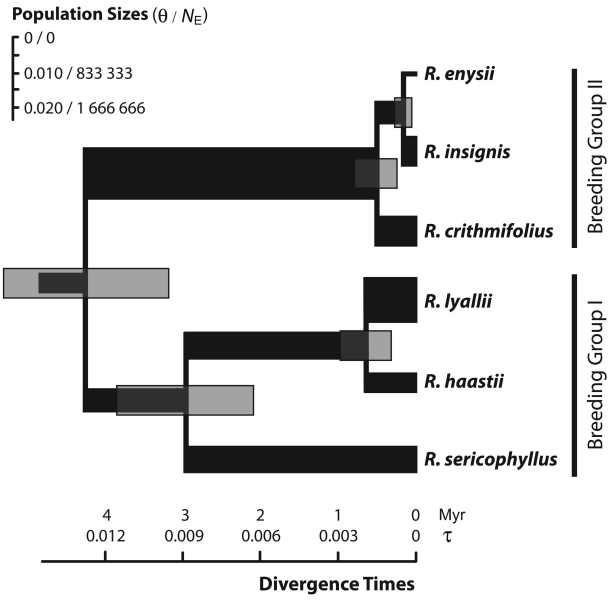

Figure 5: Species chronogram for the alpine *Ranunculus* species. The branch lengths of the species phylogeny are proportional to the posterior mean for divergence times, and the boxes represent the 95% credible sets. The timescale is given in terms of τ (years $\times$ μ) and in millions of years (*Myr*), given $\mu = 3 \times 10^{-9}$ substitutions per site per year. The width of the branches is proportional to the posterior mean for population sizes in units of θ ($=4N_E\mu$) and N_E for $\mu = 3 \times 10^{-9}$. The branch lengths and population size parameters were obtained using the program MCMCcoal (Rannala and Yang 2003) on the species tree obtained by gene tree parsimony. *Ranunculus acaulis* was used to root the phylogeny but is not shown.

Table 1: Interindividual Hamming distances that provide evidence for hybridization with the concatenated chloroplast data set

Individuals	Distance	*P*
Ranunculus crithmifolius Mt. Lyndon—*Ranunculus haastii* 6251	.008128	.0025
Ranunculus insignis Mt. Hutt—*R. haastii* 6251	.009361	.0151
R. crithmifolius Mt. Lyndon—*Ranunculus sericophyllus* Mt. Memphis	.009356	.0197
R. crithmifolius Mt. Lyndon—*Ranunculus lyallii* Mt. Anglem	.009373	.0199
R. crithmifolius Mt. Lyndon—*R. lyallii* Mt. Cook	.009639	.0276
R. crithmifolius Mt. Lyndon—*R. lyallii* 6329	.009639	.0276

Note: Probabilities obtained with the original priors are shown.

should occur more frequently. Their test evaluates whether the observed frequencies of gene trees deviate significantly from what is expected under a model of lineage sorting. One limitation of this approach is that many gene trees (>30) are needed to be able to detect hybridization. Further, when there are more than three species to consider, the probabilities for alternative gene trees associated with each of these species trees are not equal (Pamilo and Nei 1988; Rosenberg and Nordborg 2002; Degnan and Rosenberg 2006), and so expectations need to be specified for each individual case. More recently, Buckley et al. (2006) have proposed an approach for testing a hybridization hypothesis using coalescent simulations. They calculated the probability that a particular species is sister to one species in two gene trees and to another in two others. However, this approach is difficult to generalize because hypotheses need to be formulated anew for each species comparison.

Other methods use test statistics that are not calculated directly from a tree, although all use coalescent simulations for making statistical decisions. A common approach of this type involves estimating gene flow between populations using population genetics approaches (MIGRATE: Beerli 2006; IM: Hey and Nielsen 2004). Estimates of migration rate that can be statistically differentiated from 0 (e.g., not included in the 95% credible sets in Bayesian analyses) indicate that hybridization is very likely to occur. However, these methods are presently limited to comparing populations that have been evolving separately for a long time; they are not taking into account population histories (Beerli 2006). Plus, if they do consider population histories, they are limited to the study of two populations (Hey and Nielsen 2004).

Wang et al. (1997) have proposed a test based on the idea that gene flow should not affect all regions of the genome equally because some genes will be under strong selection for maintaining species cohesion. Since gene flow is expected to increase shared polymorphisms and reduce fixed differences at neutral loci in hybridizing taxa, these authors formulated a statistic based on whether the discrepancies in fixed differences and shared polymorphisms among loci are greater than expected under a strict divergence (lineage sorting) scenario, and they assess the significance using coalescent simulations. The potential of the method for detecting hybridization depends on an appropriate sampling of loci. In cases where only a few genes are under strong selection to maintain species cohesion (e.g., Noor et al. 2001), the likelihood of sampling genes or loci linked to genes under selection will be small, and this will limit the effectiveness of the approach. Machado et al. (2002) have suggested another method based on expectations for linkage disequilibrium expectation between sites that harbor shared polymorphisms and between sites where one shows an ancestral polymorphism and the other a derived polymorphism. The idea is that a hybridization event leaves less time for introgressed polymorphisms to recombine in the recipient population as compared with ancient polymorphisms. Again, coalescent simulations are used to see whether the statistic departs from expectations under a model of no hybridization. One important limitation of this method is that there needs to be recombination within loci and a sufficient number of shared and exclusive polymorphisms to calculate linkage disequilibrium. These conditions are unlikely to be met for some groups of species.

The present approach falls in the same class as these latter approaches, because although it relies on a species tree, the test statistic is not estimated from a tree. It uses a simple statistic—the smallest distance between any pair of sequences from two species—and could be used with a single genetic marker, making the method easily applicable to the study of any group of species. Moreover, the framework we presented here of testing hybridization hypotheses by generating posterior predictive distributions has the advantage of taking into account the uncertainty of the data and incorporating the stochastic nature of the coalescent and the mutation processes.

The method assumes that there is no recombination in the region of interest, which could be an advantage or an inconvenience depending on whether recombination is frequent or not in a group. For example, the present data sets do not exhibit signatures of recombination, a pattern shared by some groups (e.g., Joly and Bruneau 2006) but not others (e.g., Machado et al. 2002). Consequently, the

present method represents a good alternative to the method of Machado et al. (2002) when there is no recombination in the regions of interest. Note that it might be possible to relax the requirement of no recombination, which is a prerequisite of current methods for estimating divergence times and population sizes (e.g., Rannala and Yang 2003). If these parameters are known a priori, it might be possible to apply the test on markers in which recombination has occurred by simulating sequences using the coalescent with recombination (Hudson 1983). However, the performance of the test in the presence of recombination needs to be investigated.

Simulations have shown that the method has relatively good power when population sizes are not unrealistically high. Shorter sequences do reduce the power of the method, which could become problematic when selecting genetic markers that lack recombination. However, it is possible to increase the statistical power of the method by combining information from multiple markers when testing specific hybridization hypotheses. Because different genes can be considered to evolve independently, it is possible to calculate a joint probability of obtaining a given distance between the same two individuals for different genes by taking the product of the probabilities obtained for the genes. For example, consider two individuals for which observed genetic distances for two genes have probability $P = .05$ and $P = .06$ under a scenario of lineage sorting. Although these observations are not significant individually ($P \geq .05$), the joint probability of observing these distances for these individuals for the genes considered is $P = .003$, a value that would support a hybridization hypothesis. Considering independent evidence is thus likely to be particularly useful for identifying recent hybrids from closely related species.

Although the method has been described using examples of nonmonophyletic species, it could also detect hybridization when species are reciprocally monophyletic, as might happen if the evolutionary history of a gene is incongruent with the species phylogeny. For instance, if in the concatenated chloroplast data set the only *Ranunculus crithmifolius* sequence included in the analysis was that from Mount Lyndon—making the species monophyletic—this sequence would still be identified as being introgressed from breeding group 1 (data not shown).

Results from the alpine *Ranunculus* have shown that the method proposed can help in understanding the evolutionary history of a group. However, one must be aware that the results obtained with the present approach could be influenced by several factors when applied to empirical data sets. One is the reliance on a specified species tree, which is a requirement of the method for estimating divergence times and population sizes. Because the choice of a species tree might influence the results, different species trees should be considered where there is uncertainty. Another consideration is prior selection for Bayesian estimation of divergence times and population sizes. One advantage of Bayesian methods over likelihood approaches is that they provide an easy means for assessing whether the information contained by the data is sufficient for estimation of parameter values (Beerli 2006). If the posterior distribution is very similar to the prior distribution, it suggests that the data do not contain enough information for a rigorous estimation. But clearly the priors can affect the posterior distributions, and they need to be selected carefully. In the present example, we attempted to minimize their effects by fixing biologically realistic priors. Yet, we also explored more extreme priors to see how these affected our inferences of hybridization. We found that even though population size estimates were somewhat influenced by prior choice, these did not affect our conclusions.

The results can also be influenced by the presence of hybrid sequences in the data sets. Indeed, the population size and divergence time parameters estimated by MCMCcoal are strictly valid only in the absence of hybridization (Rannala and Yang 2003). The presence of hybrid sequences would result in smaller t/N_E proportions for branches of the species tree to account for the topological incongruence created by the hybridization event. This would in turn increase the importance of incomplete lineage sorting in the simulated data sets and reduce the power of identifying hybridization events. Consequently, the approach proposed here to detect hybridization in empirical data sets is conservative because hybridization is tested on the same data sets that are used to infer population sizes and divergence times.

Another caveat arises because the method relies on the assumption that there is no recombination within markers. Using gene regions that are subject to recombination/concerted evolution for estimating population sizes and divergence times could potentially affect inferences made using the test that we propose. In this study, we used the nrITS locus for illustrating our method. This is a large gene family that is known to be affected by recombination and concerted evolution (Álvarez and Wendel 2003). In our case, we were unable to detect any evidence for recombination in our data set. However, this finding is a minimum requirement and does not guarantee that concerted evolution is not affecting evolution of the nrITS locus. Some patterns resulting from concerted evolution could potentially bias the estimates of population sizes toward lower values and cause the loci to coalesce faster. For this reason, independent low copy number markers would be the markers of choice when implementing the proposed method.

It is important to emphasize that the approach we de-

scribe will not detect all hybridization events. When an introgressed sequence coalesces before the speciation event, it cannot be distinguished from incomplete lineage sorting (fig. 1*C*). Moreover, some hybridization events may not even result in nonmonophyletic species (fig. 1*D*). Failure to detect significant difference between simulated and observed minimum distances with this approach should never be interpreted as absence of hybridization between two species, but rather that the test is unable to determine whether the incongruence is due to hybridization or to incomplete lineage sorting.

Acknowledgments

We thank Z. Yang for allowing us to use a prerelease version of MCMCcoal (ver. 1.2) and C. Lehnebach for helpful advice and suggestions. We also acknowledge useful criticisms from D. Schoen and two anonymous reviewers on an earlier version of the manuscript. Financial support for this work comes from a Natural Sciences and Engineering Research Council postdoctoral fellowship to S.J. and from a Marsden grant to P.J.L.

APPENDIX

Detailed Analysis of the New Zealand Alpine *Ranunculus*

The method described in this article was used to test whether hybridization could be an explanation for the nonmonophyly of species in gene trees of the New Zealand alpine buttercups (*Ranunculus* L.). The radiation of these buttercups (18–21 species) has accompanied Pliocene mountain building and extreme Pleistocene climatic changes in New Zealand (Batt et al. 2000). Gene tree studies (Lockhart et al. 2001) for these species have found that numerous species are not monophyletic. Because of the recent origin of the group, incomplete lineage sorting might explain this finding. Yet hybridization is also a possible explanation because species are interfertile and hybrids have been found in nature (Fisher 1965). Six well-defined species from two major breeding groups (Fisher 1965) were investigated: *Ranunculus sericophyllus* Hook. f., *Ranunculus haastii* Hook. f., and *Ranunculus lyallii* Hook. f. from group 1 and *Ranunculus enysii* Kirk, *Ranunculus insignis* Hook. f., and *Ranunculus crithmifolius* Hook. f. from group 2. A New Zealand subalpine species, *Ranunculus acaulis* DC., was included to root the species phylogeny (Paun et al. 2005).

The five chloroplast regions constituting the concatenated data set (*trn*C-*trn*D, *trn*L-*trn*F, *psb*A-*trn*H, *trn*D-*trn*T, *rp*L16) were amplified using primers listed in table A1 and sequenced using standard procedures (Lockhart et al. 2001; GenBank accession numbers FJ744168–FJ744237). The J_{SA} and ITS data sets (GenBank accession numbers FJ711776–FJ712023) that were used to estimate the species tree and its population sizes and divergence times are subsets of a larger data set in which all New Zealand alpine *Ranunculus* species were included; these will be described in detail elsewhere. We chose not to concatenate the J_{SA} data set with the other chloroplast genes to maximize the amount of information from which we estimate population sizes and divergence times on the species tree.

Phylogenetic Analyses

All three data sets (TreeBase accession number S2284) were tested for the presence of recombination using the Φ statistic (Bruen et al. 2006). These tests were performed using the PhiPack package (Bruen 2005) with a window size of 100 bp and 10,000 permutations. Recombination was not detected in any of the data sets ($P > .05$); therefore, all were included in the analyses. The concatenated chloroplast data set was analyzed using both likelihood and parsimony optimization criteria. The likelihood search consisted of a heuristic search in PAUP* (ver. 4b10; Swofford 2002) using the nucleotide substitution model TVM + I, which was selected by the Akaike Information Criterion in ModelTest (ver. 3.7; Posada and Crandall 1998), using empirical base frequencies and estimating all other parameters during the analysis. The parsimony search, made using PAUP*, consisted of a branch and bound search keeping all most parsimonious trees.

Because no widely accepted phylogeny exists for these species, a species tree was inferred using gene tree parsimony (Maddison 1997; Maddison and Knowles 2006) under the assumption that differences in the chloroplast and nrITS gene trees were the result of incomplete lineage sorting. The best species tree chosen was the one that minimized the number of deep coalescences, or occurrences of lineage sorting, over all gene trees (Maddison 1997). The species tree was reconstructed from all three data sets in Mesquite (Maddison and Maddison 2008) using SPR tree rearrangement, treating the gene trees as unrooted and automatically resolving polytomies when optimizing gene trees on the species tree. The gene trees used for reconstructing the species trees were the strict consensus of all most parsimonious trees found for each of the three data sets (figs. A1, A2). The search for the concatenated chloroplast data set was as previously described, whereas the tree search for the nrITS and J_{SA} data sets consisted of 10 random addition sequence replicates, saving a maximum of 1,000 most parsimonious trees per replicate (other default settings were chosen). A single species tree was obtained by gene tree parsimony (fig. 5) with a score of 27 deep coalescences when rooted with *Ranun-*

culus acaulis. Alternative species trees had scores of 29 deep coalescences or more.

Estimation of Divergence Times and Population Sizes

Divergence times (τ = years × μ) and population sizes ($\theta = 4N_E\mu$) for the species tree were estimated with the Bayesian method implemented in MCMCcoal (ver. 1.2; Rannala and Yang 2003; Burgess and Yang 2008) from the DNA sequences of the three data sets. For each locus (i), the program estimated a relative mutation rate (r_i) and assumed a heredity scalar (h_i) so that $\theta_i = \theta h_i r_i$ and $\tau_i = \tau r_i$. These parameters allow the data sets to have different mutation rates and acknowledge the difference in effective population size for nuclear and chloroplast markers. The locus-specific mutation rates (r_i) were modeled with a transformed dirichlet distribution ($\alpha = 4$). For the heredity scalar, the program assumed that population size for chloroplast markers was a quarter of that for nuclear markers because species of *Ranunculus* are generally outcrossers (protandry; Zomlefer 1994). Nuclear and chloroplast markers were attributed a heredity scalar (h_i) of 1 and 1/4, respectively.

The biogeographic history of the New Zealand alpine buttercups was considered for fixing priors on divergence times (τ). These groups of alpine species have likely diversified following the formation of the Southern Alps in New Zealand ~5 million years ago (Batt et al. 2000). Arguing otherwise would imply speciation in lowlands followed by convergent adaptation of lineages to mountain habitats, a less parsimonious scenario that is also rejected by previous divergence time estimates (Lockhart et al. 2001). Because the mutation rate (μ) is incorporated in parameters τ and θ, μ needs to be specified when determining the priors for these parameters. If the divergence of the two alpine breeding groups is assumed to be 5 million years, the extent of mean sequence divergence gives $\mu \approx 3 \times 10^{-9}$ substitutions per site per year for the combined chloroplast data set. Although this is a very approximate estimation, it is consistent with published estimates for chloroplast sequences (Wolfe et al. 1987). This estimate is only useful for determining realistic prior distributions, and it does not affect the outcome of the test for hybridization because μ is integrated in the population size and divergence time parameters estimated by MCMCcoal. The priors for τ were modeled by gamma distributions using parameters of table A2. Given $\mu = 3 \times 10^{-9}$, these priors have nonzero probabilities associated with divergence times between 0 and 7 million years for the ancestor of the alpine species and between 0 and 5 million years within the alpine radiation. The prior for the split between *R. acaulis* and the ingroup species spanned a range between 2 and 16 million years, which includes values from previous divergence time estimates (Paun et al. 2005). It is more difficult to place biologically realistic bounds for present and past effective population size parameters (θ). The population sizes of alpine *Ranunculus* species are often very small, yet populations are also isolated from one another, which increases N_E. Therefore, the priors for θ were modeled by a gamma distribution of shape $\alpha = 2$ and scale $\beta = 500$ (the mean and variance are α/β and α/β^2, respectively), which give nonzero probabilities for θ values between 0 and 0.015, or 0 to 1.24×10^6 N_E for $\mu = 3 \times 10^{-9}$. To facilitate comparisons, the same prior for population size was used for all branches.

Values for the mixing tuning parameters in MCMCcoal were fixed to 5, 0.0005, 1.5, 0.003, and 0.5 and 1 for ε_1 to ε_6, respectively, which gave good acceptance ratios for all analyses (i.e., between 0.25 and 0.7). Two chains of 1 million generations were performed simultaneously to confirm convergence of estimates. The simultaneous chains always converged on the same parameter estimates, and the samples from these were combined. The first 50,000 generations of each run were discarded, and the chains were then sampled every 10 generations to generate the posterior distributions. The software Tracer (Rambaut and Drummond 2005) confirmed appropriate mixing of the chains and showed that estimated sample sizes (which take into account the correlation among subsequent samples from the chain) were above 2×10^4 for all parameters. Because branch length parameters are not completely independent, analyses were also performed without data to estimate the true priors (table A3).

The results show that species from the two breeding groups diverged ~4 million years ago, assuming $\mu = 3 \times 10^{-9}$ (table A4; figs. 5, A3). Besides the speciation event that gave rise to *R. sericophyllus*, estimated to have occurred at ~3 million years ago, the remaining speciation events within breeding groups appear to have occurred recently, most likely within the last million years. Population size estimates were quite large, generally around 500,000 N_E for $\mu = 3 \times 10^{-9}$ (table A4; figs. 5, A3).

To investigate whether the results obtained were affected by the priors, four analyses with more extreme priors were also performed. For each of these, either θ or τ was assigned the original prior, and the other was assigned a smaller or higher prior (tables A2, A3). Divergence time estimates were little affected by the different priors (fig. A3; table A4), with the exception of the divergence time between the outgroup *R. acaulis* and all alpine species. The impact of this uncertainty had little consequence for this study because *R. acaulis* was not used in the gene tree simulations. Although population size estimates were more affected by the choice of prior (fig. A3), the general conclusions were unaffected.

Generating Posterior Predictive Distributions

Distributions of the test statistic under a null hypothesis of lineage sorting (no hybridization) were obtained through coalescent simulations that excluded the outgroup species *R. acaulis*. For each data set (i), 10,000 gene trees were simulated on the species tree in MCMCcoal, selecting, for each gene tree, a different set of parameters τ_i, θ_i, and r_i according to their marginal posterior distributions. Because the parameters θ_i and τ_i are not independent of each other (Rannala and Yang 2003), the set of parameters used for simulating each gene tree came from a single point of the MCMC chain. Character matrices were simulated independently for each data set. These had the same length as the originals and were simulated in Seq-Gen (ver. 1.3.1; Rambaut and Grassly 1997) using the nucleotide substitution model that best fitted the original data (table A5), as determined by the Akaike Information Criterion in ModelTest. The shortest Hamming distance between any two sequences for all pairs of species was then calculated for all replicates for the three data sets; these collections of values constituted the posterior predictive distributions.

Table A1: Primers used for amplifying the five chloroplast regions used in this study

Chloroplast region and primer name	DNA sequence	Reference
*psb*A-*trn*H:		
trnHgug	CGC GCA TGG TGG ATT CAC AAT CC	Tate and Simpson 2003
psbA	GTT ATG CAT GAA CGT AAT GCT C	Sang et al. 1997
*rpl*16:		
rpL16F71	GCT ATG CTT AGT GTG TGA CTC GTT G	Small et al. 1998
rpL16R1516	CCC TTC ATT CTT CCT CTA TGT TG	Small et al. 1998
*trn*C-*trn*D:		
trnC-F	CCA GTT CAA ATC TGG GTG TC	Lee and Wen 2004
915R	TGA AAG GGA AAA TGT AAA GAC T	This study
*trn*D-*trn*T:		
trnDpop	ACC AAT TGA ACT ACA ATC CC	This study
trnT	CTA CCA CTG AGT TAA AAG GG	This study
*trn*L-*trn*F:		
tabC	CGA AAT CGG TAG ACG CTA CG	Taberlet et al. 1991
tabF	ATT TGA ACT GGT GAC ACG AG	Taberlet et al. 1991

Table A2: Priors used for the different MCMCcoal analyses

	τ			
	Alpine species		Tree	
Priors (α, β)	Internal nodes	Common ancestor	root	θ
Original prior	2, 350	3, 400	6, 300	2, 500
High τ	2, 150	3, 200	6, 150	2, 500
Low τ	2, 900	3, 900	6, 800	2, 500
High θ	2, 350	3, 400	6, 300	2, 100
Low θ	2, 350	3, 400	6, 300	2, 1,000

Note: α and β parameters for the gamma distribution are given in the format α, β; the mean and variance of the distribution are α/β and α/β^2, respectively. $\tau = t\mu$, where t is the time and μ the mutation rate. $\theta = 4N_E\mu$, where N_E is the effective population size and μ the mutation rate.

Table A3: Mean values and 95% credible sets for the different priors used in the MCMCcoal analysis (*Ranunculus* L.)

Parameters	Original	High τ	Low τ	High θ	Low θ
θ:					
R. crithmifolius	.004 (.000076, .0094)	.0039 (.000077, .0094)	.004 (.000025, .0095)	.02 (.00048, .048)	.002 (.00004, .0048)
R. enysii	.004 (.00011, .0095)	.004 (.000085, .0095)	.004 (.0001, .0096)	.02 (.00051, .047)	.002 (.000038, .0047)
R. haastii	.004 (.000095, .0095)	.004 (.000071, .0095)	.004 (.000084, .0096)	.02 (.00051, .047)	.002 (.000042, .0048)
R. insignis	.004 (.0001, .0096)	.0039 (.000075, .0093)	.004 (.000081, .0096)	.02 (.00073, .047)	.002 (.000057, .0047)
R. lyallii	.0039 (.000082, .0094)	.0039 (.00011, .0094)	.004 (.000084, .0097)	.02 (.00067, .048)	.002 (.000026, .0047)
R. sericophyllus	.004 (.000097, .0094)	.0039 (.000088, .0093)	.0041 (.00018, .0096)	.02 (.0006, .047)	.002 (.000061, .0047)
ei	.004 (.000067, .0095)	.004 (.000098, .0096)	.004 (.000091, .0094)	.02 (.00041, .047)	.002 (.000052, .0048)
cei	.004 (.000083, .0096)	.004 (.0001, .0095)	.004 (.00008, .0096)	.02 (.00059, .048)	.002 (.000043, .0048)
lh	.004 (.000083, .0094)	.004 (.000076, .0095)	.004 (.000076, .0096)	.02 (.00058, .048)	.002 (.000039, .0047)
slh	.004 (.000079, .0095)	.004 (.000073, .0095)	.004 (.000093, .0096)	.02 (.00039, .048)	.002 (.000041, .0048)
ceislh	.004 (.000056, .0096)	.004 (.000082, .0095)	.004 (.000098, .0096)	.02 (.00046, .047)	.002 (.000025, .0048)
aceislh	.004 (.00011, .0096)	.004 (.000076, .0095)	.004 (.000075, .0095)	.02 (.00039, .047)	.002 (.000033, .0047)
τ:					
aceislh	.022 (.0087, .037)	.044 (.012, .074)	.0083 (.0035, .014)	.022 (.0087, .037)	.022 (.0088, .037)
ceislh	.011 (.0043, .018)	.023 (.0097, .038)	.0044 (.0018, .0074)	.011 (.0043, .018)	.011 (.0044, .018)
cei	.0059 (.0012, .011)	.013 (.0029, .024)	.0023 (.00052, .0045)	.0059 (.0012, .011)	.0059 (.0013, .011)
ei	.003 (.00012, .0067)	.0068 (.0003, .015)	.0012 (.000027, .0026)	.0031 (.00013, .0067)	.0031 (.000088, .0067)
slh	.0059 (.0014, .011)	.013 (.0031, .024)	.0023 (.00047, .0044)	.0059 (.0013, .011)	.0059 (.0012, .011)
lh	.0031 (.00011, .0067)	.0069 (.00011, .015)	.0012 (.000067, .0026)	.003 (.00014, .0067)	.003 (.00017, .0066)
r:					
cp data set	1 (.28, 1.76)	1 (.27, 1.75)	1 (.29, 1.78)	1 (.28, 1.77)	1 (.29, 1.77)
nrITS	1 (.29, 1.78)	1 (.28, 1.77)	1 (.28, 1.77)	1 (.28, 1.77)	1 (.28, 1.76)
J_{SA}	1 (.28, 1.77)	1 (.28, 1.77)	1 (.28, 1.76)	1 (.27, 1.76)	1 (.29, 1.78)

Note: These represent the priors for branch lengths (τ), population sizes (θ), and locus rate (r) estimated by running the analysis without data. When there is more than one species involved for a parameter, only the first letter of the epithet is used to identify the species.

Table A4: Mean values and 95% credible sets of the posterior distributions of the parameters estimated in MCMCcoal for all priors (*Ranunculus* L.)

Parameters	Original	High τ	Low τ	High θ	Low θ
θ:					
R. crithmifolius	.0078 (.0035, .013)	.00 (.0036, .013)	.0076 (.0033, .013)	.013 (.0046, .024)	.0056 (.0027, .0089)
R. enysii	.0013 (.00025, .0027)	.0014 (.00025, .0028)	.0013 (.00025, .0026)	.0016 (.00027, .0034)	.0011 (.00022, .0022)
R. haastii	.0054 (.0018, .0098)	.0055 (.0017, .0098)	.0052 (.0017, .0095)	.0096 (.0022, .02)	.0038 (.0014, .0065)
R. insignis	.0076 (.0033, .013)	.0077 (.0035, .013)	.0076 (.0031, .013)	.013 (.0037, .027)	.0057 (.0027, .091)
R. lyallii	.012 (.0075, .017)	.012 (.0075, .017)	.012 (.0072, .017)	.016 (.0093, .024)	.0096 (.0061, .013)
R. sericophyllus	.0072 (.0042, .011)	.0073 (.0043, .011)	.0068 (.004, .01)	.0093 (.0051, .014)	.0065 (.0037, .0096)
ei	.0061 (.00081, .012)	.0062 (.00077, .012)	.0058 (.00077, .012)	.013 (.0011, .031)	.0037 (.00042, .0076)
cei	.014 (.0071, .021)	.014 (.0071, .021)	.013 (.0068, .02)	.021 (.0092, .034)	.01 (.0058, .015)
lh	.0073 (.0024, .013)	.0073 (.0024, .013)	.0074 (.0026, .013)	.014 (.0043, .027)	.0047 (.0015, .0082)
slh	.0063 (.0012, .013)	.0061 (.0011, .012)	.0066 (.0014, .013)	.018 (.002, .04)	.0034 (.00037, .0067)
ceislh	.0054 (.0003, .012)	.0052 (.00024, .011)	.006 (.00038, .013)	.02 (.0026, .041)	.0024 (.000094, .0055)
aceislh	.004 (.00008, .0094)	.0038 (.000084, .0089)	.0059 (.00028, .013)	.018 (.00081, .041)	.0021 (.000043, .0049)
τ:					
aceislh	.023 (.015, .032)	.026 (.016, .036)	.016 (.0092, .022)	.024 (.013, .035)	.022 (.014, .03)
ceislh	.011 (.0081, .014)	.012 (.0086, .015)	.0097 (.007, .012)	.01 (.0069, .013)	.011 (.0087, .014)
cei	.001 (.00044, .0019)	.0011 (.00045, .002)	.00095 (.0004, .0017)	.0013 (.0055, .0021)	.00085 (.00035, .0016)
ei	.00055 (.00022, .0009)	.00058 (.00023, .00094)	.00051 (.00021, .00083)	.0006 (.00024, .001)	.00051 (.0002, .00083)
slh	.0069 (.0047, .0093)	.0073 (.0051, .0099)	.0058 (.004, .0078)	.0064 (.0043, .0087)	.0072 (.0049, .0095)
lh	.002 (.00095, .003)	.0021 (.0001, .0032)	.0017 (.00087, .0026)	.0019 (.00097, .003)	.002 (.001, .0031)
r:					
cp data set	.59 (.46, .74)	.57 (.44, .72)	.64 (.5, .8)	.61 (.45, .78)	.59 (.46, .73)
nrITS	1.36 (1.12, 1.59)	1.4 (1.16, 1.64)	1.27 (1.04, 1.5)	1.5 (1.24, 1.75)	1.28 (1.06, 1.51)
J_{SA}	1.05 (.83, 1.27)	1.03 (.8, 1.25)	1.08 (.88, 1.31)	.9 (.67, 1.13)	1.12 (.91, 1.33)

Note: Branch lengths (τ), population sizes (θ), and locus rate (r) are given for the original priors and for different extreme priors. When there is more than one species involved for a parameter, only the first letter of the epithet is used to identify the species.

Table A5: Nucleotide substitution models used for each data set in generating the predictive posterior distributions

Data set	Length	Substitution model	Nucleotide frequencies				Nucleotide substitution matrix							
			A	C	G	T	A → C	A → G	A → T	C → G	C → T	G → T	α	I
Concatenated cp	4,135	TVM	.3379	.1517	.1737	.3367	1.0327	.7399	.2124	.0000	.7399	1	None	.8537
J_{SA}	480	TIM	.2592	.2086	.1559	.376262	1.0000	3.1640	.3764	.3764	1.7647	1	None	.7917
nrITS	603	SYM	.25	.25	.25	.25	.9567	6.1676	.9901	.0698	2.6941	1	.8264	.5840

Note: α, alpha parameter for the gamma distribution of rate variation across sites; I, proportion of invariable sites.

Figure A1: Phylogram of one most parsimonious tree and parsimony statistics for the nrITS data set. Dotted lines indicate branches that collapse in the strict consensus tree.

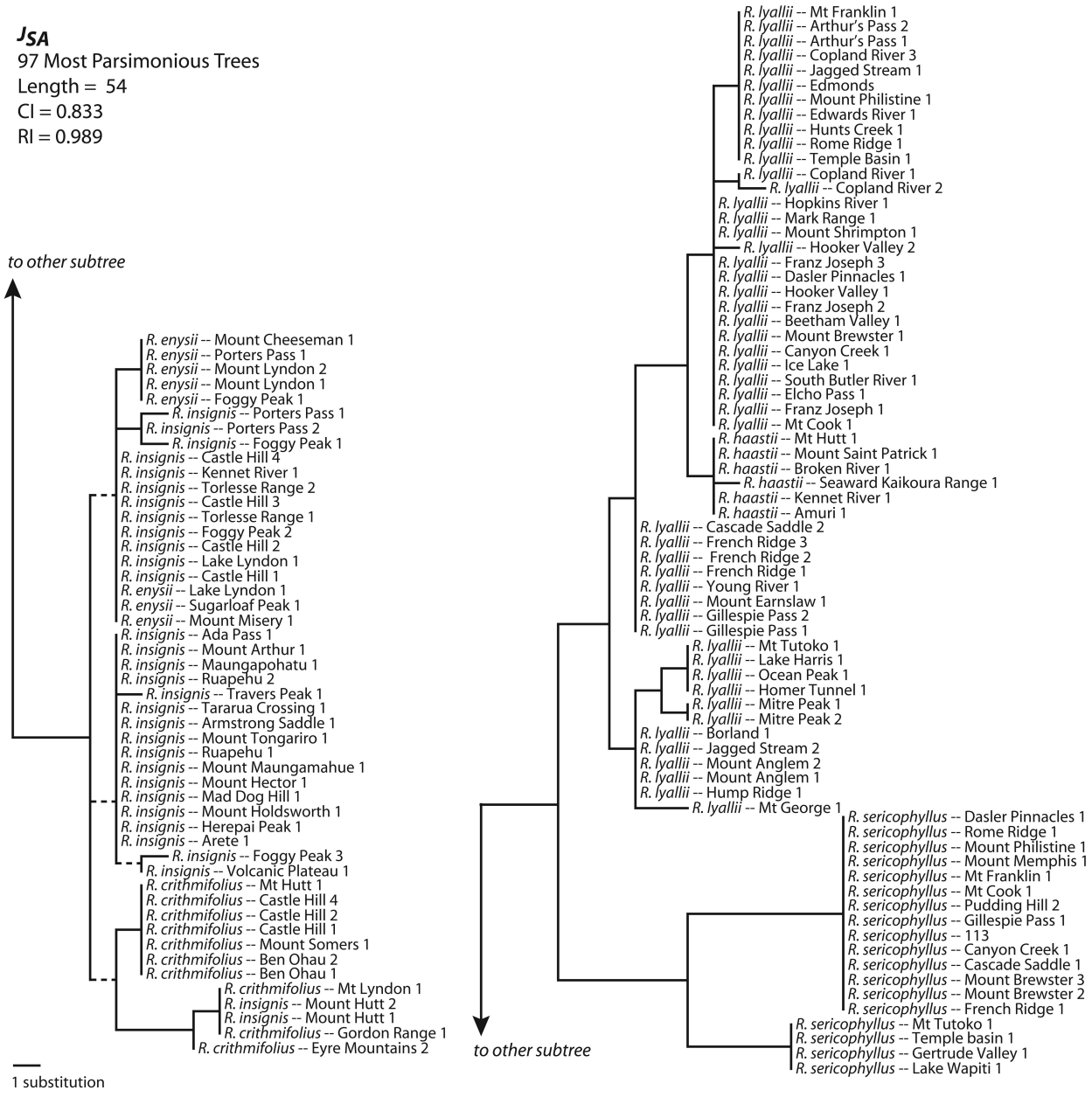

Figure A2: Phylogram of one most parsimonious tree and parsimony statistics for the J_{SA} data set. Dotted lines indicate branches that collapse in the strict consensus tree.

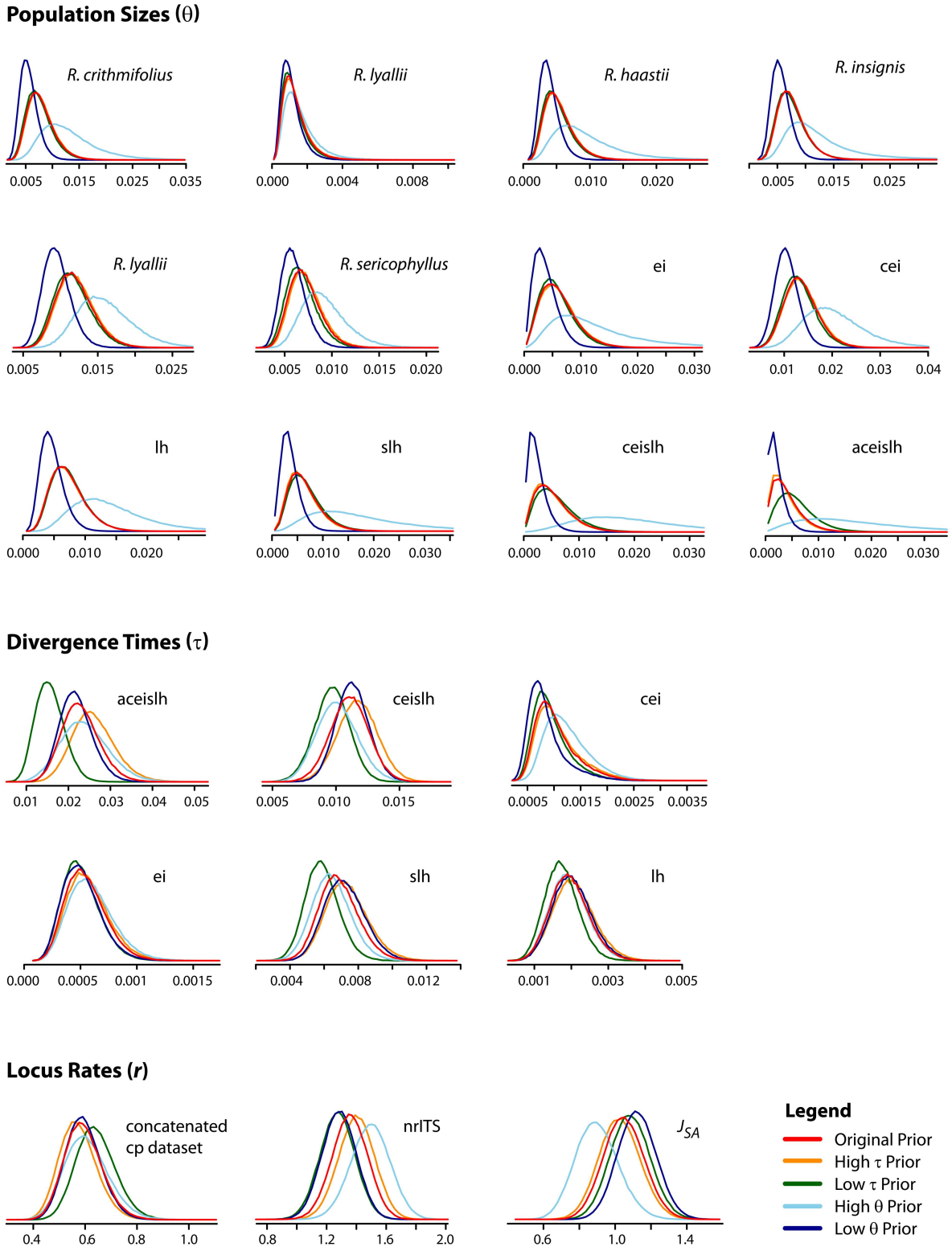

Figure A3: Posterior distributions obtained for branch lengths (τ), population sizes (θ), and locus rates (r) with the different priors used.

Literature Cited

Álvarez, I., and J. F. Wendel. 2003. Ribosomal ITS sequence and plant phylogenetic inference. Molecular Phylogenetics and Evolution 29: 417–434.

Anderson, E. 1949. Introgressive hybridization. Wiley, New York.

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

———. 2004. Transfer and origin of adaptations through natural hybridization: were Anderson and Stebbins right? Plant Cell 16: 562–570.

Barton, N. H. 2001. The role of hybridization in evolution. Molecular Ecology 10:551–568.

Batt, G. E., J. Braun, B. P. Kohn, and I. McDougall. 2000. Thermochronological analysis of the dynamics of the Southern Alps, New Zealand. Geological Society of America Bulletin 112:250–266.

Beerli, P. 2006. Comparison of Bayesian and maximum likelihood inference of population genetic parameters. Bioinformatics 22: 341–345.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57:289–300.

Bordewich, M., S. Linz, K. St. John, and C. Semple. 2007. A reduction algorithm for computing the hybridization number of two trees. Evolutionary Bioinformatics Online 3:86–98.

Bruen, T. C. 2005. PhiPack: PHI test and other tests of recombination. McGill University, Montreal, Quebec.

Bruen, T. C., H. Philippe, and D. Bryant. 2006. A simple robust statistical test for detecting the presence of recombination. Genetics 172:2665–2681.

Buckley, T. R., M. Cordeiro, D. C. Marshall, and C. Simon. 2006. Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (*Maoricicada* Dugdale). Systematic Biology 55:411–425.

Burgess, R., and Z. Yang. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. Molecular Biology and Evolution 25:1979–1994.

Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. PLoS Genetics 2:e68.

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

Ellstrand, N. C., and K. A. Schierenbeck. 2000. Hybridization as a stimulus for the evolution of invasiveness in plants? Proceedings of the National Academy of Sciences of the USA 97:7043–7050.

Ferguson, D., and T. Sang. 2001. Speciation through homoploid hybridization between allotetraploids in peonies (*Paeonia*). Proceedings of the National Academy of Sciences of the USA 98:3915–3919.

Fisher, F. J. F. 1965. The alpine *Ranunculi* of New Zealand. DSIRO, Wellington, New Zealand.

Funk, D. J., and K. E. Omland. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annual Review of Ecology, Evolution, and Systematics 34:397–423.

Grant, B. R., and P. R. Grant. 1996. High survival of Darwin's finch hybrids: effects of beak morphology and diets. Ecology 77:500–509.

Grant, V. 1981. Plant speciation. 2nd ed. Columbia University Press, New York.

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

Holder, M. T., J. A. Anderson, and A. K. Holloway. 2001. Difficulties in detecting hybridization. Systematic Biology 50:978–982.

Holland, B. R., S. Benthin, P. J. Lockhart, V. Moulton, and K. T. Huber. 2008. Using supernetworks to distinguish hybridization from incomplete lineage sorting. BMC Evolutionary Biology 8: 202.

Hudson, R. R. 1983. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology 23:182–201.

Huson, D. H., T. Klöpper, P. J. Lockhart, and M. A. Steel. 2005. Reconstruction of reticulate networks from gene trees. Proceedings of the Ninth International Conference on Research in Computational Molecular Biology, pp. 233–249. Springer, Heidelberg.

Jin, G., L. Nakhleh, S. Snir, and T. Tuller. 2006. Inferring phylogenetic networks by the maximum parsimony criterion: a case study. Molecular Biology and Evolution 24:324–337.

Joly, S., and A. Bruneau. 2006. Incorporating allelic variation for reconstructing the evolutionary history of organisms from multiple genes: an example from *Rosa* in North America. Systematic Biology 55:623–636.

Joly, S., J. R. Starr, W. H. Lewis, and A. Bruneau. 2006. Polyploid and hybrid evolution in roses east of the Rocky Mountains. American Journal of Botany 93:412–425.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–123 *in* H. N. Munro, ed. Mammalian protein metabolism. Academic Press, New York.

Kingman, J. F. C. 1982*a*. The coalescent. Stochastic Processes and Their Applications 13:235–248.

———. 1982*b*. On the genealogy of large populations. Journal of Applied Probability 19:27–43.

Lee, C., and J. Wen. 2004. Phylogeny of *Panax* using chloroplast *trn*C-*trn*D intergenic region and the utility of *trn*C-*trn*D in interspecific studies of plants. Molecular Phylogenetics and Evolution 31:894–903.

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

Linder, C. R., and L. H. Rieseberg. 2004. Reconstructing patterns of reticulate evolution in plants. American Journal of Botany 91: 1700–1708.

Lockhart, P. J., P. A. McLenachan, D. Havell, D. Gleny, D. Huson, and U. Jensen. 2001. Phylogeny, radiation, and transoceanic dispersal of New Zealand alpine buttercups: molecular evidence under split decomposition. Annals of the Missouri Botanical Garden 88: 458–477.

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

Maddison, W. P. 1997. Gene trees in species trees. Systematic Biology 46:523–536.

Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. Systematic Biology 55:21–30.

Maddison, W. P., and D. R. Maddison. 2008. Mesquite: a modular system for evolutionary analysis. Version 2.5. http://mesquiteproject.org.

Noor, M. A. F., K. L. Grams, L. A. Bertucci, Y. Almendarez, J. Reiland, and K. R. Smith. 2001. The genetics of reproductive isolation and the potential for gene exchange between *Drosophila pseudoobscura* and *D. persimilis* via backcross hybrid males. Evolution 55:512–521.

Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. Molecular Biology and Evolution 5:568–583.

Paun, O., C. Lehnebach, J. T. Johansson, P. J. Lockhart, and E. Hörandl. 2005. Phylogenetic relationships and biogeography of *Ranunculus* and allied genera (Ranunculaceae) in the Mediterranean region and in the European Alpine system. Taxon 54:911–930.

Posada, D., and K. A. Crandall. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14:817–818.

Rambaut, A., and A. J. Drummond. 2005. Tracer. Version 1.3. http://tree.bio.ed.ac.uk/software/tracer/.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Computer Applications in the Biosciences 13: 235–238.

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

Rieseberg, L. H. 1997. Hybrid origins of plant species. Annual Review of Ecology and Systematics 28:359–389.

Rieseberg, L. H., O. Raymond, D. M. Rosenthal, Z. Lai, K. Livingstone, T. Nakazato, J. L. Durphy, A. E. Schwarzbach, L. A. Donovan, and C. Lexer. 2003. Major ecological transitions in wild sunflowers facilitated by hybridization. Science 301:1211–1216.

Rosenberg, N. A., and M. Nordborg. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews 3:380–390.

Sang, T., and Y. Zhong. 2000. Testing hybridization hypotheses based on incongruent gene trees. Systematic Biology 49:422–434.

Sang, T., D. J. Crawford, and T. F. Stuessy. 1997. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). American Journal of Botany 84:1120–1136.

Seehausen, O. 2004. Hybridization and adaptive radiation. Trends in Ecology & Evolution 19:198–207.

Small, R. L., J. A. Ryburn, R. C. Cronn, T. Seelanan, and J. F. Wendel. 1998. The tortoise and the hare: choosing between noncoding plastome and nuclear Adh sequences for phylogeny reconstruction in a recently diverged plant group. American Journal of Botany 85:1301–1315.

Small, R. L., R. Cronn, and J. F. Wendel. 2004. Use of nuclear genes for phylogeny reconstruction in plants. Australian Systematic Botany 17:145–170.

Stebbins, G. L. 1959. The role of hybridization in evolution. Proceedings of the American Philosophical Society 103:231–251.

Swofford, D. L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4.0b10. Sinauer, Sunderland, MA.

Taberlet, P., L. Gielly, G. Pautou, and J. Bouvet. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Molecular Biology 17:1105–1109.

Tate, J. A., and B. B. Simpson. 2003. Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploid species. Systematic Botany 28: 723–737.

Wang, R. L., J. Wakeley, and J. Hey. 1997. Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. Genetics 147:1091–1106.

Wolfe, K. H., W.-H. Li, and P. M. Sharp. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proceedings of the National Academy of Sciences of the USA 84:9054–9058.

Zomlefer, W. B. 1994. Guide to flowering plant families. University of North Carolina Press, Chapel Hill.

Associate Editor: Armando Caballero
Editor: Michael C. Whitlock