



Comparison of Methods for Species-Tree Inference in the Sawfly Genus *Neodiprion* (Hymenoptera: Diprionidae)

Citation

Linnen, Catherine R. and Brian D. Farrell. 2008. Comparison of methods for species-tree inference in the sawfly genus *Neodiprion* (Hymenoptera: Diprionidae). *Systematic Biology* 57(6): 876-890.

Published Version

<http://dx.doi.org/10.1080/10635150802580949>

Permanent link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:3138566>

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

Comparison of Methods for Species-Tree Inference in the Sawfly Genus *Neodiprion* (Hymenoptera: Diprionidae)

CATHERINE R. LINNEN AND BRIAN D. FARRELL

Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA;
E-mail: clinnen@oeb.harvard.edu (C.R.L.)

Abstract.—Conifer-feeding sawflies in the genus *Neodiprion* provide an excellent opportunity to investigate the origin and maintenance of barriers to reproduction, but obtaining a phylogenetic estimate for comparative studies of *Neodiprion* speciation has proved difficult. Specifically, nonmonophyly within and discordance between individual gene trees, both of which are common in groups that diverged recently and/or rapidly, make it impossible to infer a species tree using methods that are designed to estimate gene trees. Therefore, in this study, we estimate relationships between members of the *lecontei* species group using four approaches that are intended to estimate species, not gene, trees: (1) minimize deep coalescences (MDC), (2) shallowest divergences (SD), (3) Bayesian estimation of species trees (BEST), and (4) a novel approach that combines concatenation with monophyly constraints (CMC). Multiple populations are sampled for most species and all four methods incorporate this intraspecific variation into estimates of interspecific relationships. We investigate the sensitivity of each method to taxonomic sampling, and, for the BEST method, we assess the impact of prior choice on species-tree inference. We also compare species-tree estimates to one another and to a morphologically based hypothesis to identify clades that are supported by multiple analyses and lines of evidence. We find that both taxonomic sampling and method choice impact species-tree estimates and that, for these data, the BEST method is strongly influenced by Θ and branch-length priors. We also find that the CMC method is the least sensitive to taxonomic sampling. Finally, although interspecific genetic variation is low due to the recent divergence of the *lecontei* group, our results to date suggest that incomplete lineage sorting and interspecific gene flow are the main factors complicating species-tree inference in *Neodiprion*. Based on these analyses, we propose a phylogenetic hypothesis for the *lecontei* group. Finally, our results suggest that, even for very challenging groups like *Neodiprion*, an underlying species-tree signal can be extracted from multi-locus data as long as intraspecific variation is adequately sampled and methods that focus on the estimation of species trees are used. [Bayesian estimation of species trees (BEST); concatenation with monophyly constraints (CMC); gene-tree discordance; hybridization; introgression; lineage sorting; minimize deep coalescences (MDC); shallowest divergences (SD).]

Comparative methods have the potential to detect and explain generalities in the way organisms diversify; however, their successful application to these questions requires a priori evolutionary hypotheses, well-chosen study organisms, and robust phylogenetic estimates. The first two requirements are met by conifer sawflies in the genus *Neodiprion* Rohwer (Hymenoptera: Diprionidae). Because they are strict host specialists and mate on their host plant, multiple authors have argued that sympatric host race formation may be common in *Neodiprion* (Ghent and Wallace, 1958; Alexander and Bigelow, 1960; Knerer and Atwood, 1972, 1973; Bush, 1975a, 1975b; Tauber and Tauber, 1981; Strong et al., 1984). In addition, recent work suggests another reason why *Neodiprion* provides an excellent system to study the origin and maintenance of barriers to reproduction: hybridization is common, yet species have remained distinct in the face of gene flow (Linnen and Farrell, 2007, 2008). However, this work also highlights a major challenge for using *Neodiprion* in comparative studies of speciation: introgression and incomplete lineage sorting make it difficult to obtain a phylogenetic estimate for the genus. These problems are not unique to *Neodiprion* and are expected to be prevalent in any taxon that has undergone recent and/or rapid diversification (Maddison, 1997; Funk and Omland, 2003; Maddison and Knowles, 2006; Kubatko and Degnan, 2007; Belfiore et al., 2008).

Most of the difficulties that are encountered when estimating phylogenies for recently diverged groups stem from the fact that phylogenetic studies typically employ methodologies designed to estimate gene trees, whereas

the history of interest is actually the timing and order of speciation events, or the “species tree” (Maddison, 1997; Edwards et al., 2007). Generally speaking, gene trees are expected to resemble the species trees that contain them, but incomplete lineage sorting and/or introgression can cause gene trees to depart from the underlying species tree in two ways: (1) individual gene-tree topologies may not match the actual order of speciation events (and each other), and (2) gene trees may contain non-monophyletic species (Maddison, 1997; Hudson and Coyne, 2002; Funk and Omland, 2003; Hudson and Turelli, 2003; Maddison and Knowles, 2006). The first of these problems, gene-tree discordance, is generally dealt with by combining data from multiple loci in the hopes of recovering a dominant signal in the data, which is then assumed to be the species phylogeny (either concatenation or consensus tree approaches can be used for these purposes; e.g., de Queiroz, 1993; Huelsenbeck et al., 1996; Rokas et al., 2003; Gadagkar et al., 2005; Gatesy and Baker, 2005; de Queiroz and Gatesy, 2007). Concatenation and consensus tree approaches do not, however, address the problem of inferring a species tree when gene trees contain species that are not reciprocally monophyletic (Carstens and Knowles, 2007). A large number of comparative studies at the species level avoid this issue altogether by utilizing exemplar phylogenies (i.e., a single individual per species); however, failure to adequately sample intraspecific variation can seriously compromise the accuracy of species-tree estimates for closely related taxa (Funk, 1999). In order to extract phylogenetic information from discordant gene trees that contain

non-monophyletic species, phylogenetic analyses must take into account the species-level phenomena responsible for these patterns (i.e., stochastic lineage sorting and introgression).

Several methods that incorporate stochastic lineage sorting into the estimation of species trees have been proposed. The simplest of these methods consider lineage sorting but do not explicitly model it; two such approaches are (1) find the species tree that minimizes the number of deep coalescences (Maddison, 1997; Maddison and Knowles, 2006) and (2) cluster species by their most similar contained sequences (shallowest coalescences; Takahata, 1989; Maddison and Knowles, 2006). These methods, along with a more recent maximum-likelihood approach that does include a stochastic model of lineage sorting (Carstens and Knowles, 2007), do not consider potential error in gene-tree estimation. In contrast, a recently developed Bayesian method employs stochastic models of nucleotide substitution and lineage sorting to simultaneously estimate individual gene trees and the species tree that contains them (Liu and Pearl, 2006, 2007; Edwards et al., 2007; Liu et al., 2008). Encouragingly, analyses of empirical and simulated data suggest that these methods can accurately estimate species trees, even when gene trees are discordant and incomplete lineage sorting is widespread. However, a critical assumption of these methods is that there has been no hybridization between species; violation of this assumption may seriously degrade the accuracy of these methods, even if gene flow rates are low enough that the species phylogeny is still fundamentally a branching process (Maddison and Knowles, 2006; Carstens and Knowles, 2007; Knowles and Carstens, 2007; Edwards et al., 2007; Liu and Pearl, 2007).

In *Neodiprion*, mitochondrial introgression has been rampant, and because their history is dominated by gene flow, mitochondrial genes are unreliable for recovering interspecific relationships (Linnen and Farrell, 2007). In contrast, patterns of gene-tree concordance and estimates of locus-specific gene flow rates suggest that nuclear genes recover predominantly phylogenetic, not introgressive, signal. Phylogenies estimated from concatenated nuclear data are well resolved but are difficult to interpret because they contain multiple non-monophyletic species (Linnen and Farrell, 2007). Because they explicitly consider lineage sorting, the species-tree methods described above might be expected to improve phylogenetic inference in *Neodiprion*. However, even low levels of gene flow may compromise the performance of these methods. By low levels of gene flow, we mean that gene flow is not high enough to erode differences between species (i.e., $Nm < 1$; Wright, 1931) and most within-species genetic variation is the result of ancestral variation and novel mutation, not introgression.

In the absence of species-tree methods that explicitly model both lineage sorting and introgression, we employ a novel strategy to estimate a species tree for *Neodiprion*. Specifically, we suggest a modification of the concatenation approach that utilizes a priori species designations as monophyly constraints in order to allow the

inclusion of multiple individuals per species. This approach provides a way to deal with the ambiguity that results when analysis of concatenated data recovers non-monophyletic species (e.g., Carstens and Knowles, 2007). The primary motivation for using this method over simpler exemplar approaches (in which monophyly constraints are implied by the choice of exemplars) is that it incorporates intraspecific variation into phylogenetic analysis, which has been shown to increase the likelihood of obtaining a topology that is concordant with the species tree (Takahata, 1989; Rosenberg, 2002; Maddison and Knowles, 2006). Moreover, we choose to concatenate multi-locus data because this approach is expected to (1) make more efficient use of limited genetic variation compared to consensus approaches and (2) allow shared phylogenetic signal to swamp out non-phylogenetic “noise” that stems from lineage sorting and introgression (Kluge, 1989; de Queiroz et al., 1995; Baker and DeSalle, 1997; Wiens, 1998; Lerat et al., 2003; de Queiroz and Gatesy, 2007).

In this article, we utilize multiple species-tree methods to obtain a phylogenetic estimate for the *lecontei* group of the genus *Neodiprion*. Along with our “concatenation with monophyly constraints” (CMC) approach, we use three methods that explicitly consider stochastic lineage sorting but might be misled to varying degrees by gene flow between species: (1) “minimize deep coalescences” (MDC; Maddison, 1997); (2) “shallowest divergence” (SD; Takahata, 1989; Maddison and Knowles, 2006), and (3) “Bayesian estimation of species trees” (BEST; Liu and Pearl, 2006, 2007; Edwards et al., 2007). We do not use the maximum-likelihood approach suggested by Carstens and Knowles (2007) because computational constraints limit this approach to a small number of taxa. We investigate the sensitivity of each of these four methods to taxonomic sampling by analyzing different samples of populations within species (Lecointre et al., 1993; Philippe, 1997; Hedtke et al., 2006). We also evaluate the impact of choice of priors on the BEST method. Finally, we compare species-tree estimates obtained using these diverse methods to one another and to a morphologically based hypothesis (Ross, 1955) and, in light of these comparisons, discuss implications for *Neodiprion* phylogeny.

MATERIALS AND METHODS

Neodiprion Samples and DNA Sequence Data

With over 50 described species and subspecies, *Neodiprion* Rohwer is the most diverse of 11 genera in the conifer sawfly family Diprionidae (Hymenoptera: Symphyta). In his 1955 revision of the genus, Ross divided *Neodiprion* into two species groups based on morphology (mesoscutellum sculpture) and geography. The *lecontei* group is found only in eastern North America, the Caribbean (Bahamas and Cuba), and Central America, whereas the *sertifer* group has a Holarctic distribution and appears to be most diverse in western North America. Detailed study of many eastern *Neodiprion* species (several of which are economically important pests, Arnett, 1993) has resulted in a relatively stable

taxonomy for the *lecontei* group (e.g., Ross, 1961; Becker, 1965; Becker et al., 1966; Becker and Benjamin, 1967; Knerer, 1984; Smith, 1988; Knerer and Wilkinson, 1990). By comparison, the *sertifer* group remains inadequately studied (although some progress has been made, e.g., Sheehan and Dahlsten, 1985; Smith and Wagner, 1986), and species boundaries and overall diversity are poorly understood. Because the species-tree methods used in this study require a priori species designations, the analyses described here focus on the *lecontei* group. Previous molecular phylogenetic studies have invariably recovered the *lecontei* clade as a monophyletic group with very high support (100% maximum-likelihood bootstrap, maximum-parsimony bootstrap, and Bayesian posterior probability; Linnen and Farrell, 2007, 2008).

Specimens included in this study were collected in the United States and Canada in 2001 to 2004 as described in Linnen and Farrell (2007), and identifications were based on larvae and reared females (following Atwood and Peck, 1943; Ross, 1955; Becker et al., 1966; Becker and Benjamin, 1967; Wilson, 1977; Knerer, 1984; Dixon, 2004). Included in this study were representatives of 17 described and 2 undescribed *lecontei* group species. These new species, which are referred to here as "N. species 1" and "N. species 2," are morphologically and genetically distinct from all known species (Linnen and Farrell, 2007, 2008) and will be formally described elsewhere. The only *lecontei* group species that were not sampled are two species that are known only from Cuba: *N. cubensis* Hochmut and *N. insularis* (Cresson). Also missing are two subspecies: *N. merkelii maestrensis* Hochmut, which is known only from Cuba, and *N. taedae taedae* Ross, from the southeastern United States.

In choosing specimens to include in molecular studies, multiple populations were included for each species whenever possible, and populations were chosen to maximize the geographical and ecological variation sampled for each species. In total, 125 *lecontei* group individuals representing 1 to 14 populations for each of 19 species were included in this study (Supplementary Table 1; available at <http://www.systematicbiology.org>). In addition, a single *sertifer* group species (*N. autumnalis*) was included to root the *lecontei* group phylogeny. DNA sequence data for each sample were generated, edited, and aligned as described in Linnen and Farrell (2007) for the following nuclear gene regions: a region of the F2 copy of *elongation factor-1 α* (EF1 α) that spanned portions of two exons and a large intervening intron (Danforth and Ji, 1998; Danforth et al., 1999; Nyman et al., 2006); a re-

gion of *rudimentary* (CAD) that spanned portions of two exons and two introns; and an anonymous nuclear locus (ANL43) (GenBank accession numbers EF361837 to EF362376; TreeBASE accession number S2212). Because mitochondrial genes are unreliable for recovering phylogenetic history in *Neodiprion* (Linnen and Farrell, 2007), only nuclear genes were used to estimate relationships between species. The final aligned data set for all nuclear genes and all specimens included in this study was 2779 base pairs (bp) in length (1089-bp EF1 α , 916-bp CAD, 774-bp ANL43).

Species-Tree Estimation

Sampling schemes.—To investigate the sensitivity of each species-tree method to taxonomic sampling, four data sets, representing three different sampling strategies, were constructed (summarized in Table 1). The first data set ("LU" for large and uneven) was analyzed in a previous study (Linnen and Farrell, 2007) and contained many individuals, but sampling was uneven across species (1 to 14 samples per species). To construct a smaller but more evenly sampled data set ("MOD" for moderate size and symmetry), one to three populations were chosen for each species, with the exact number depending on availability of samples and distribution of species (e.g., only a single population was available for some species, whereas the maximum of three populations was used for widely distributed, well-sampled species). Finally, two even, but sparsely sampled, data sets each included only a single individual per species ("ExemA" and "ExemB" for exemplar sets A and B). Individuals for the exemplar sets were sampled arbitrarily from the MOD data set: ExemA included the individual with the lowest collection ID number for each species; ExemB included individuals with the highest collection ID numbers. Collection data for all individuals and data sets are given in Supplementary Table 1.

Although an additional data set that was both large and evenly sampled would have completed the range of possible sampling strategies investigated, such a sample was unavailable despite intensive collecting efforts. However, most groups of organisms contain both rare and common species, and the samples used here are representative of the range of samples typically available in studies of closely related species.

Concatenation with monophyly constraints (CMC).—Conceptually, monophyly constraints fit most naturally into a Bayesian framework for phylogenetic analysis because preexisting taxonomic information (species designations) can be incorporated into the analysis as topological priors. Constrained Bayesian analyses of the three concatenated nuclear genes were performed on the four data sets described in the previous section (LU, MOD, ExemA, and ExemB). For each analysis, sequence data were partitioned by locus (ANL43, EF1 α , and CAD), and for each data set, models for individual loci were selected using the Akaike information criterion (AIC) and MrModelTest version 2.2 (Nylander et al., 2004). Substitution-model parameters were unlinked

TABLE 1. Summary of data sets and sampling schemes used in this study. Numbers refer to ingroup taxa; range refers to the range of individuals sampled per species for a given data set.

Data set	Sampling scheme	No. of species	No. of individuals	Range
LU	Large; uneven	19	125	1–14
MOD	Moderate size and symmetry	19	38	1–3
ExemA	Exemplar; small and even	19	19	1
ExemB	Exemplar; small and even	19	19	1

across data partitions and among-partition rate variation was accommodated using rate multipliers (option: `prset ratepr=variable`; see Marshall et al., 2006). Bayesian searches were performed in MrBayes version 3.1 (Ronquist and Huelsenbeck, 2003) and consisted of two concurrent runs (each with four to five Markov chains and temperatures between 0.1 and 0.2), 10 to 15 million generations (sampled every 1000 generations), and a 25% burn-in. Runs were considered to have converged on the stationary distribution when there were no obvious trends in generation versus log-likelihood plots and the average standard deviation of split frequencies was below 0.01 (Ronquist et al., 2005). Finally, for comparison with constrained Bayesian analyses, unconstrained searches were performed using the same models and run conditions as the constrained analyses.

To investigate the impact of partition choice on Bayesian analyses, two additional partitioning schemes were used: (1) a five-partition scheme in which protein-coding genes (*EF1 α* and *CAD*) were partitioned into introns and exons, and (2) a seven-partition scheme in which exons were further divided into first+second and third codon positions. Models for each partition were chosen using the AIC and MrModelTest. Finally, because Bayesian inference is more sensitive to model under-specification than over-specification, additional analyses in which a complex model of nucleotide substitution (GTR+I+ Γ) was chosen for each partition were performed (Huelsenbeck and Rannala, 2004).

Constrained and unconstrained analyses were also performed using maximum likelihood (ML) and maximum parsimony (MP) to investigate the impact of analysis method. Monophyly constraints for constrained ML and MP analyses were constructed in MacClade version 4.05 (Maddison and Maddison, 2000); for the exemplar sets, monophyly constraints are implicit in the choice of samples (i.e., one individual per species). ML searches for the MOD, ExemA, and ExemB data sets were performed on concatenated nuclear data sets in PAUP* 4.0b10 (Swofford, 2000) with 1000 random addition sequences (RAS), tree bisection-reconnection (TBR) branch swapping, and the "MulTrees" option. Models and parameters for concatenated nuclear data were chosen for each data set using MrModelTest and the AIC. ML bootstrap analyses consisted of 500 replicates, each with 10 random addition sequences and TBR branch swapping ("MulTrees" option in effect). ML searches for the LU data set were performed in GARLI version 0.951 (Zwickl, 2006) using the model of sequence evolution selected by MrModelTest and the AIC (model parameters were estimated by GARLI). Automatic termination was enforced and runs were stopped when 250,000 generations had passed without a significantly better scoring topology. To ensure consistency of topologies and likelihood scores, GARLI runs were repeated until similar topologies and likelihood scores (all within 1 log-likelihood unit) were obtained from three independent runs. The topology with the best likelihood score of these three was then selected as the ML tree (or if different topologies had the same score, a strict consensus was used). Bootstrap analyses

were also performed in GARLI and consisted of 500 replicates with the automatic termination criterion reduced to 5000 generations.

MP searches for all data sets were performed in PAUP*. Searches for the MOD and exemplar sets consisted of 1000 RAS, TBR branch swapping, and "MulTrees." When searches returned multiple equally parsimonious trees, results were summarized using a strict consensus. Bootstrap searches for the MOD and exemplar data sets consisted of 1000 replicates, each with 100 RAS, TBR branch swapping, and "MulTrees." MP searches and bootstrap analyses for the LU data set were the same as for the other data sets, except that each bootstrap replicate consisted of 10 RAS and no more than 10 trees were saved for a given replicate to reduce computation time.

Minimize deep coalescences (MDC).—The MDC method seeks the species tree that minimizes the number of incomplete lineage sorting (deep coalescence) events that must be inferred to explain observed gene trees (Maddison, 1997). Individual gene trees for the MDC method were estimated for each data set using maximum likelihood (model parameters for each gene and data set were estimated as described in "Concatenation with Monophyly Constraints"). ML analyses for MOD, ExemA, and ExemB were performed in PAUP* with 1000 RAS, TBR branch swapping, and the "MulTrees" option. For the LU data set, individual gene trees were estimated in GARLI with three independent runs per gene, each with a termination threshold of 250,000 generations. When analyses (PAUP* and GARLI) returned multiple, equally likely trees, a strict consensus was computed in PAUP* for use in MDC analyses.

Once ML gene-tree estimates were obtained, MDC searches were performed for all four data sets in Mesquite version 1.12 (Maddison and Maddison, 2006) with the following options: subtree pruning and regrafting (SPR) branch swapping, MAXTREES set to 100, scores were computed using "Deep Coalescence Multiple Loci," contained gene trees were treated as rooted (the *sertifer* group species, *N. autumnalis*, was used as an outgroup), and gene-tree polytomies were automatically resolved to minimize incompleteness of lineage sorting. To investigate the impact of the "auto-resolve" option, an additional set of MDC searches was performed without resolving polytomies.

Shallowest divergences (SD).—The SD method, which is based on Takahata's (1989) demonstration that the order of interspecific coalescences within a group has a high probability of matching the actual order of speciation events, utilizes a clustering algorithm to group species and clades not by average pairwise sequence divergences but by their most similar (i.e., fewest number of nucleotide differences) pair of contained DNA sequences (Maddison and Knowles, 2006). For full and reduced data sets, aligned sequence matrices for the three nuclear genes were imported into Mesquite version 1.12 and the "Cluster Analysis" option was used to perform SD analyses. Following Maddison and Knowles (2006), simple uncorrected DNA (p) distances were used, the "closest" option was chosen as the method to count

distances among contained taxa (i.e., samples from a particular species), and “single linkage” was chosen as the cluster method.

Bayesian estimation of species trees (BEST).—The BEST method utilizes a Markov chain Monte Carlo (MCMC) algorithm to estimate the joint posterior distribution of gene trees and the species tree under a hierarchical Bayesian model. This model consists of (1) the likelihood of the data given a vector of gene trees and substitution-model parameters; (2) the prior distribution of substitution model parameters; (3) the probability distribution of gene trees given the species tree (derived from coalescent theory); (4) the prior distribution of the species tree (BEST version 1 uses a birth-death species-tree prior; BEST version 2 uses a uniform prior); and (5) the prior distribution of θ (Edwards et al., 2007; Liu and Pearl, 2007; Liu et al., 2008).

BEST analyses were performed for each data set (ExemA, ExemB, MOD, and LU) in BEST version 2 (Edwards et al., 2007; Liu and Pearl, 2007; Liu et al., 2008), which is a modification of MrBayes version 3.1.2 (Ronquist and Huelsenbeck, 2003). Nucleotide-substitution models were selected for each locus and data set as described in “Concatenation with Monophyly Constraints” and parameters were unlinked across data partitions (loci). To investigate the impact of prior choice on species-tree inference, multiple θ and branch-length priors were used. Specifically, an inverse gamma prior with $\alpha = 3$ and three different values of β (0.006, 0.02, and 1) was used for θ . These values were chosen to cover a biologically realistic range of values of θ for *Neodiprion* and are consistent with sequence-based estimates of θ calculated in DNAsp (Rozas et al., 2003; data not shown). In addition, for each of the three β values, two branch-length priors were used: (1) an exponential prior, in which gene-tree branch lengths are unconstrained and follow an exponential distribution with the default parameter 10 (non-clock gene trees are then converted to clock-like trees using the ad-hoc method described in Edwards et al., 2007); and (2) a coalescent prior, in which gene trees are constrained to be clock-like (no ad hoc adjustment is needed).

In total, 24 sets of BEST analyses representing every possible combination of data set (four total), θ prior (three total), and branch-length prior (two total) were run. For each data set/prior combination, a minimum of 10 independent BEST runs were performed, each consisting of a single Markov chain and 50 million generations, the first 40 million of which were discarded as burn-in. If, after 10 runs, no two analyses converged on a similar topology (as judged by the average standard deviation of split frequencies), results were summarized across all runs using the “sumt” command in BEST version 2 to assess areas of agreement/disagreement across independent analyses.

Comparison with Morphologically Based Hypotheses

Based on a detailed study of larval and adult morphology, Ross (1955) named five “complexes” of species (*virginianus*, *pratti*, *pinusrigidae*, *lecontei*, and *abbottii*) within the *lecontei* group and suggested that one of the five, the

abbottii complex, was sister to remaining species in the group (Fig. 1). We asked whether these six hypothesized groups were recovered in our species-tree estimates to (1) draw direct comparisons between relationships implied by different data set/method combinations and (2) identify relationships that are both robust to method choice and supported by multiple lines of evidence (i.e., morphological and molecular). Because their morphology has not yet been formally described and no a priori hypotheses for their relationships to other *Neodiprion* species exist, N. species 1 and N. species 2 were pruned from species trees before scoring the presence/absence of clades.

For trees estimated using the MDC and SD methods, we simply recorded the presence/absence of the six *lecontei* group clades in Figure 1 for each data set; when there were multiple MDC trees, the percentage of trees containing each clade was recorded (SD analyses returned only a single tree per data set). For the CMC and BEST methods, Bayesian tests of monophyly were performed. Specifically, for each data set/method combination, each hypothesized clade in Figure 1 was evaluated and rejected if present in less than 5% (0.8% after Bonferroni correction for $n = 6$ tests) of the post-burn-in set of trees for that combination (e.g., Miller et al., 2002; Buschbom and Barker, 2006; Linnen and Farrell, 2007).

RESULTS

Constrained and Unconstrained Analyses

The lengths, percentages of variable sites, and models chosen by MrModelTest and the AIC are given in Table 2 for each locus and data set—it is clear from this table that increasing the number of individuals sampled also increases the proportion of sites that are variable and informative. These results demonstrate that the added individuals in the larger data sets contribute unique information for inferring relationships between *Neodiprion* species.

TABLE 2. Length, percentage of variable sites, percentage of parsimony-informative (PI) sites, and models selected by MrModelTest for each locus and data set.

Locus	Data set	Length (bp)	Variable sites (%)	PI sites (%)	Model
EF1 α	LU	1089	10.19	7.35	HKY+I
	MOD	1089	8.26	3.95	HKY+I
	ExemA	1089	7.07	2.30	HKY+I
	ExemB	1089	6.98	2.39	HKY+I
CAD	LU	916	5.68	3.71	GTR+I
	MOD	916	4.15	1.86	HKY+I
	ExemA	916	2.84	1.20	HKY+I
	ExemB	916	3.38	0.98	HKY+I
ANL43	LU	774	13.70	9.04	GTR+I+ Γ
	MOD	774	11.89	5.43	GTR+I+ Γ
	ExemA	774	8.27	2.45	GTR+I+ Γ
	ExemB	774	9.69	2.33	GTR+I+ Γ
All	LU	2779	9.68	6.62	GTR+I+ Γ
	MOD	2779	7.92	3.67	GTR+I+ Γ
	ExemA	2779	6.01	1.98	GTR+I+ Γ
	ExemB	2779	6.55	1.91	GTR+I+ Γ

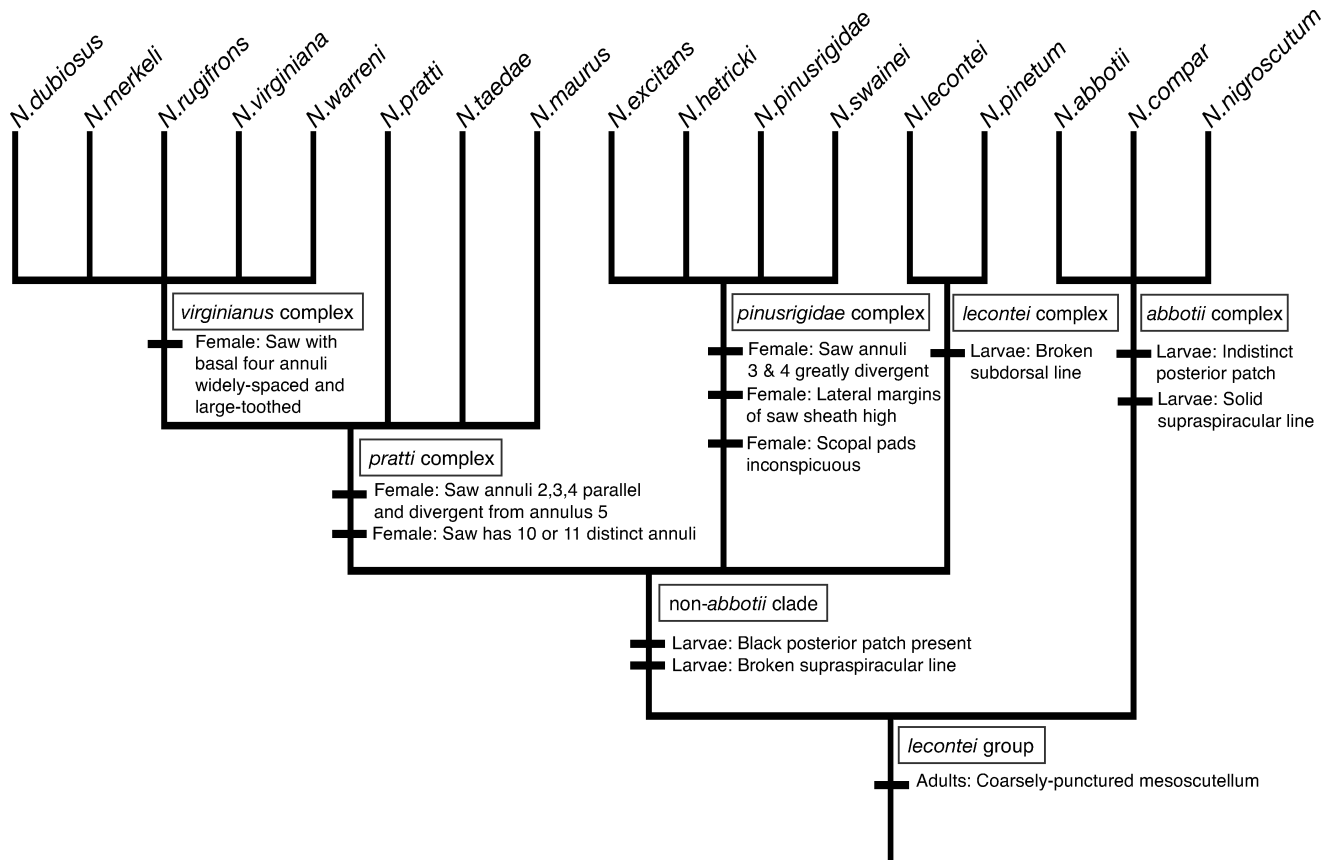


FIGURE 1. *Neodiprion* relationships proposed by Ross (1955) based on morphology. The key morphological characters (and the life-history stage to which they apply) that Ross (1955) used as a basis for these groupings are given for each clade.

For both constrained and unconstrained Bayesian analyses, there were no obvious trends in the generation versus log-likelihood plots and average standard deviation of split frequencies were all below 0.01, which suggests that these searches converged on the stationary distribution (Ronquist et al., 2005). Also, results obtained in Bayesian analyses were robust to choice of partitioning schemes and substitution models (results not shown); therefore, only results from the three-partition (ANL43, CAD, EF1 α) analysis are shown.

As was the case with a previous analysis of the LU data set (Linnen and Farrell, 2007), the unconstrained phylogeny contains several strongly supported clades (>95% posterior probabilities); however, there are some unresolved (or poorly supported) nodes and five species were recovered as non-monophyletic (Supplementary Fig. 1; all supplemental figures are available at <http://www.systematicbiology.org>). Examination of the ML estimates and patterns of variation for individual genes (Supplementary Figs. 2 to 4; Table 2) reveals that these difficulties likely stem from a combination of low levels of informative variation at individual loci, discordance between gene trees, and a lack of reciprocal monophyly within individual gene trees.

Comparing unconstrained and constrained analyses for the largest data set (LU), it can be seen that both

analyses recovered the same overall tree structure with similar levels of support (Supplementary Fig. 1; Fig. 2a), which suggests that the inclusion of monophyly constraints does not have a large impact on the inference of interspecific relationships. Bayes factors calculated from the harmonic means of the likelihoods from unconstrained and constrained analyses suggest that the more complex model (i.e., unconstrained) should be preferred for the MOD and LU data sets ($2\ln(B_{10}) > 10$; Table 3; Kass and Raftery, 1995; Nylander et al., 2004). This does not imply, however, that the three nuclear genes lack a shared history. In fact, explicitly modeling species monophyly and a shared history (i.e., the hierarchical model implemented in BEST) resulted in a dramatic improvement in likelihood scores for all data sets. For example, the BEST model with the coalescent prior was decisively favored over the model employed in the unconstrained Bayesian analyses ($2\ln(B_{10}) > 100$; Table 3).

Results obtained from constrained Bayesian, parsimony, and likelihood analyses of the three nuclear genes (EF1 α , CAD, and ANL43) for the four data sets (ExemA, ExemB, MOD, and LU) are summarized in Table 3, Figure 2a, and Supplementary Figure 5. Across all four data sets, there was strong agreement between constrained Bayesian, ML, and MP analyses. Specifically, most nodes were agreed upon by all three methods

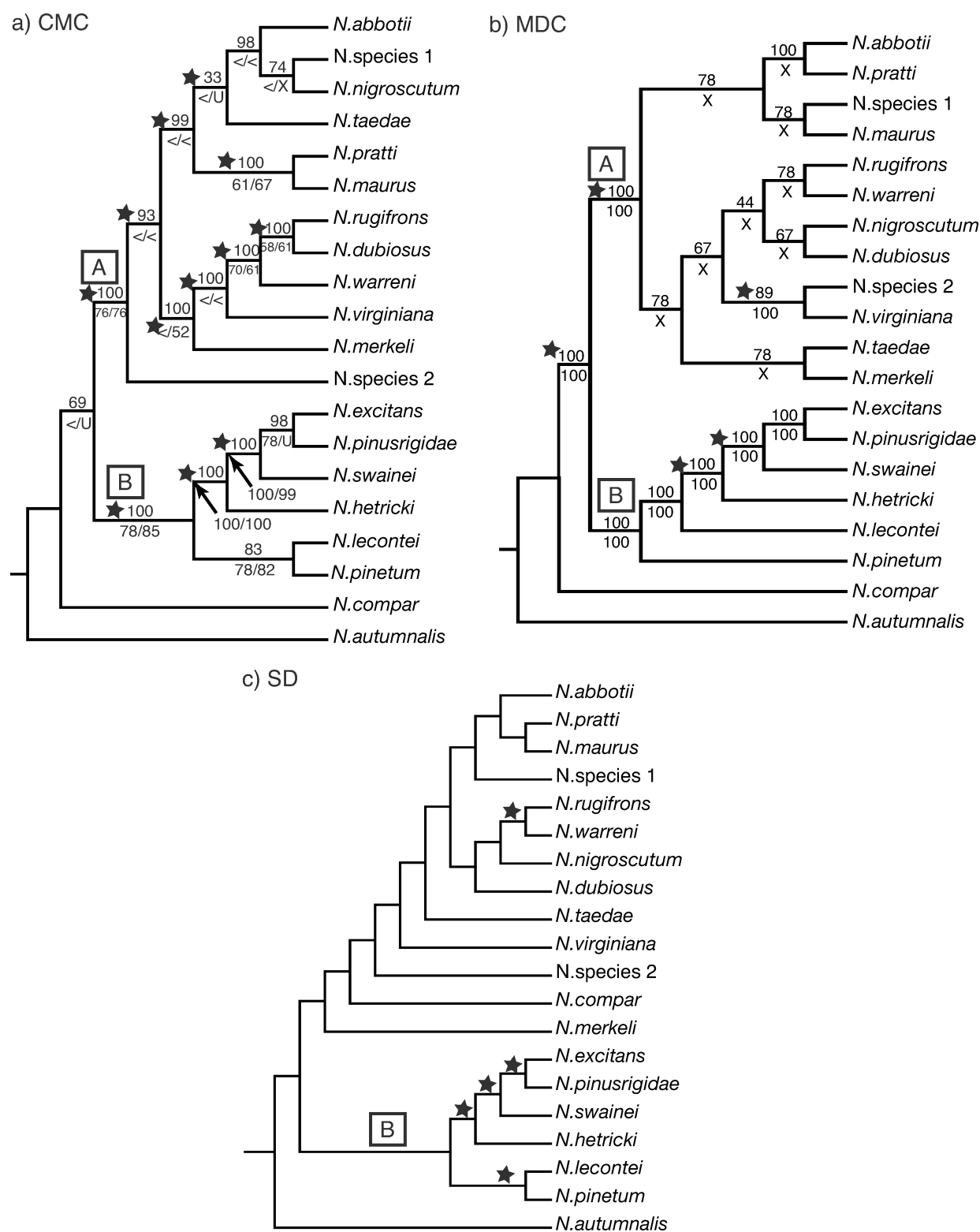


FIGURE 2. Species trees estimated for the LU data set using (a) concatenation with monophyly constraints (CMC), (b) minimize deep coalescences (MDC), and (c) shallowest divergences (SD). Stars indicate clades that were recovered across all four data sets. Support for individual branches is indicated in (a) as follows: Bayesian posterior probabilities are above each node; maximum-likelihood (ML) bootstrap/maximum-parsimony (MP) bootstrap values are below each node (in that order); a "<" indicates nodes that were present in ML or MP analyses but received less than 50% bootstrap support; a "U" indicates nodes that were unresolved (but not conflicting); an "X" indicates that a conflicting relationship was recovered by ML or MP with less than 50% bootstrap support. In (b), numbers above and below nodes indicate the percentage of MDC trees that contained that clade for the "auto-resolve" (above) and "no auto-resolve" (below) analyses; an "X" indicates that a conflicting relationship was recovered using the "no auto-resolve" option. Clade labels "A" and "B" denote clades that are discussed further in the text.

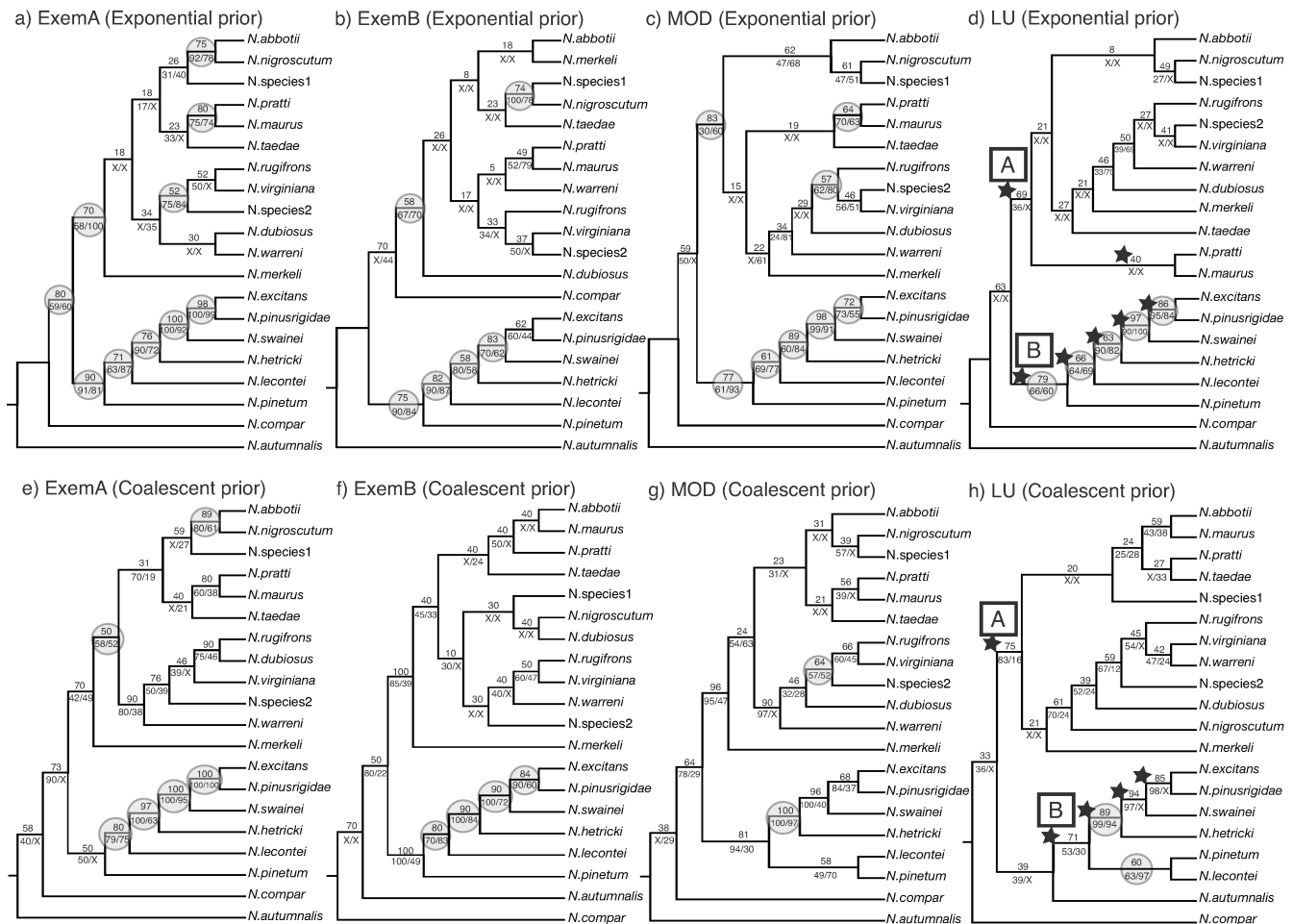


FIGURE 3. Species-tree estimates for the BEST method (using $\beta = 0.02$ for the θ prior) for all data sets and branch-length priors. Numbers above each node are Bayesian posterior probabilities for $\beta = 0.02$; posterior probabilities for $\beta = 0.006$ and $\beta = 1$ are given below each node (in that order); an "X" indicates that a conflicting relationship was recovered under the different β value. For a given data set/branch-length combination, clades that were recovered with posterior probabilities of 50% or above by all θ priors are indicated by gray circles. Stars in the LU trees indicate clades that were recovered by all data sets for a given set of priors. Clade labels "A" and "B" denote clades that are discussed further in the text.

(although some of these received $<50\%$ ML and MP bootstrap support) and there were no strongly supported conflicts (i.e., in the three cases where one method conflicted with the other two, bootstrap support for this node was less than 50%; Fig. 2a, Supplementary Fig. 5). In addition, although sampling did have some impact on CMC analyses (e.g., position of *N. compar*; relationships within the clade containing *N. abbotii* + *N. species1* + *N. nigroscutum* + *N. taedae*), most nodes (12) were identical across all data sets (see starred clades in Fig. 2a).

MDC, SD, and BEST

The ML gene trees used in MDC analyses are given in Supplementary Figures 2 to 4 (ML trees are shown for the largest data set only), and the extent of incomplete lineage sorting (inferred from MDC scores; Table 3) in the species trees inferred from these gene trees is comparable to values observed for simulated species trees with

a shallow depth (i.e., recent divergence; Maddison and Knowles, 2006). MDC scores for the analyses that automatically resolved polytomies were substantially better (lower) than scores for analyses that did not (Table 3). In addition, for every data set, the auto-resolve option had a large impact on the species-tree estimated using the MDC approach (see clades denoted by "X" in Fig. 2b and Supplementary Fig. 6). These observations suggest that how uncertainty in individual gene trees is handled can have a large impact on MDC species-tree estimates. The SD method returns a single species-tree estimate and these are given in Figure 2c (LU data set) and Supplementary Figure 7 (remaining data sets). Taxonomic sampling had a large impact on the MDC and SD methods—for both methods, only five nodes were identical across data sets (see starred clades in Figs. 2b and 2c).

The results of all BEST analyses are summarized in Table 3 and Figure 3. With the exception of two combinations of data set/ θ prior/branch-length prior (the

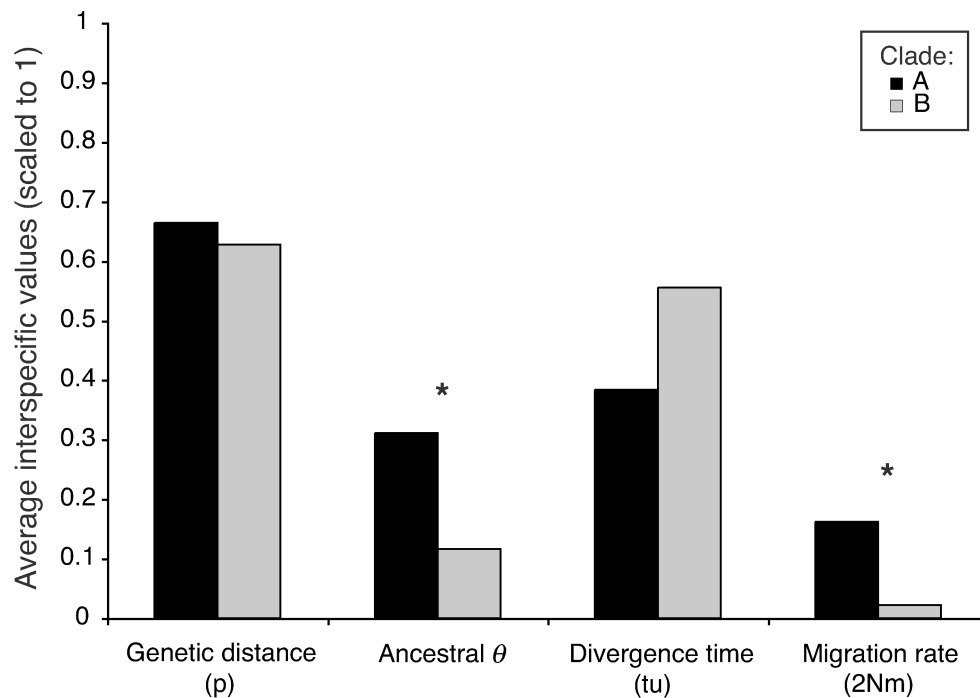


FIGURE 4. Average interspecific genetic distance (uncorrected p), ancestral θ (4 Nu), divergence time (tu), and migration rate for the A and B clades (clades correspond to labeled clades in Figs. 2 and 3). Average interspecific values were estimated as described in Table 5 and in the text. To facilitate comparisons between different parameters, the highest value for each of the four parameters was assigned a value of one and all other estimates for that parameter were scaled accordingly. Asterisks indicate parameters for which there was a significant difference between the two clades ($P < 0.05$; one-tailed Mann-Whitney U -test).

TABLE 3. Summary of tree scores for all data set/method combinations. Analysis abbreviations are as follows: constrained maximum parsimony (CMP), constrained maximum likelihood (CML), minimize deep coalescences (MDC), unconstrained Bayesian (UB), concatenation with monophyly constraints (CMC), Bayesian estimation of species tree (BEST). Scores are given for MDC analyses conducted with ("auto-res.") and without ("no auto-res.") automatic resolution of polytomies. Priors for BEST analyses are given in parentheses as follows: numbers indicate values of β used for the θ prior and letters indicate whether a coalescent ("C") or exponential ("E") branch-length prior was used. Negative log-likelihood ($-\ln L$) values are given for all Bayesian analyses (UB, CMC, BEST) and are the harmonic means of the estimated marginal likelihoods of all post-burn-in trees. For each data set, the highest log-likelihood score is indicated in bold.

	Data set			
	ExemA	ExemB	MOD	LU
No. of CMP trees	17	14	4	2200
CMP tree length	218	239	305	431
No. of CML trees	1	1	1	1
CML $-\ln L$	5350.66	5505.81	6070.70	7092.39
No. of MDC trees (auto-res.)	100	20	18	9
MDC score (auto-res.)	29	38	57	127
No. of MDC trees (no auto-res.)	1	8	8	2
MDC score (no auto-res.)	71	68	168	300
UB $-\ln L$	5370.79	5524.42	6167.93	7307.67
CMC $-\ln L$	5370.79	5524.42	6187.67	7323.03
BEST $-\ln L$ (0.006, E)	5564.90	5714.21	6391.28	8333.73
BEST $-\ln L$ (0.02, E)	5576.92	5676.14	6397.46	8358.98
BEST $-\ln L$ (1, E)	5582.94	5719.30	6373.59	8329.90
BEST $-\ln L$ (0.006, C)	5310.21	5428.64	5975.17	7281.66
BEST $-\ln L$ (0.02, C)	5298.00	5423.99	5952.01	7134.09
BEST $-\ln L$ (1, C)	5305.45	5430.35	5942.25	6979.55

ExemA and ExemB data sets, both with coalescent prior and $\beta = 1$), none of the BEST analyses had an average standard deviation of split frequencies below 0.01 (range: 0.015 to 0.283), which indicates that independent runs for each of these combinations did not converge on the same topology. This failure to converge could be due to a combination of insufficient run times and inefficient searching of tree-space; however, this explanation seems unlikely given that (1) post-burn-in log-likelihood versus generation plots were flat (i.e., increased run times would have been unlikely to produce a better solution) and (2) altering run conditions (e.g., increasing the number of Markov chains and altering the "propTemp" and "poissonmean" search parameters) did not result in improved log-likelihood scores (results not shown). It seems more likely that the observed lack of convergence stems from the existence of multiple species-tree solutions that provide equally good explanations for the data under the given models. Additional data may therefore be required to improve convergence and species-tree inference using BEST, and results from analyses that did not converge should be interpreted with caution and reevaluated as additional loci and populations are sampled. Nevertheless, when results were summarized across 10 independent runs for a given combination, several clades received high support, indicating that there were at least some consistencies across independent runs (Fig. 3). Further, when a second set of 10 runs was performed for a

subset of data set/prior combinations, results were comparable to those obtained by the first set of 10 runs (results not shown).

Both branch-length and θ priors influenced species-tree estimates obtained by BEST analyses (Fig. 3). For each data set, different relationships were recovered by exponential (Fig. 3a to d) and coalescent (Fig. 3e to h) priors, and a comparison of the log-likelihood scores obtained under these two priors suggests that the coalescent prior provides a better explanation of the data (Table 3). Likewise, for each data set/branch-length prior combination, the species-tree estimate depended to some degree on the θ prior. This finding contrasts with a previous study in which the species-tree topologies estimated from two multi-locus data sets were found to be robust to choice of θ prior (Liu and Pearl, 2007). Nevertheless, clades that received high posterior probabilities ($>70\%$) for one θ prior tended to be recovered with some

support ($>50\%$ posterior probabilities) by other θ priors (see circled nodes in Fig. 3). Surprisingly, the θ prior with the highest, and perhaps most unrealistic (based on sequence-based estimates of θ for *Neodiprion*), β value ($\beta = 1$) consistently returned the best likelihood scores for the multiple-allele data sets (MOD and LU), whereas the exemplar data sets (ExemA and ExemB) preferred lower values for the θ prior ($\beta = 0.006$ or 0.02 ; Table 3). A potential explanation for this pattern is considered in the discussion.

As was the case for the other species-tree methods investigated, BEST species-tree estimates were dependent on taxonomic sampling, and some nodes were recovered by all data sets (see starred clades in Fig. 3d and h). Interestingly, posterior probabilities obtained using the BEST method were consistently lower than those obtained using the CMC method (Figs. 2 and 3; Table 4). Belfiore and colleagues (2008) noted a similar pattern

TABLE 4. Comparison of morphologically based hypotheses to results obtained across species-tree methods and data sets. Method and data set abbreviations are as described in the text and in Table 3. Priors for BEST analyses are given in parentheses: numbers refer to the value of β for the θ prior and “exp.” and “coal.” refer to the exponential and coalescent branch-length priors. Clade names correspond to those in Figure 1 and numbers indicate the percentage of trees from a given method/data set combination that contained a particular clade. For MDC and SD analyses, an “X” indicates clades that were absent; for CMC and BEST analyses, an “X” indicates clades present in fewer than 5% (0.8% after Bonferroni correction for $n = 6$ tests) of all post-burn-in trees.

Method	Data set	<i>Virginianus</i> complex	<i>Pratti</i> complex	<i>Pinusrigidae</i> complex	<i>Lecontei</i> complex	<i>Abbotii</i> complex	non- <i>Abbotii</i> clade
CMC	LU	100	X	100.00	83.20	X	X
	MOD	99.99	X	100.00	11.42	X	X
	ExemA	99.87	X	100.00	1.85	X	X
	ExemB	99.37	X	100.00	8.53	X	X
MDC	LU (auto-res.)	X	X	100.00	X	X	X
	LU (no auto-res.)	X	X	100.00	100	X	X
	MOD (auto-res.)	X	X	100.00	100	X	X
	MOD (no auto-res.)	X	X	100.00	50	X	X
	ExemA (auto-res.)	X	X	100.00	100	X	X
	ExemA (no auto-res.)	X	X	100.00	100	X	X
	ExemB (auto-res.)	X	X	100.00	100	X	X
	ExemB (no auto-res.)	X	X	100.00	100	X	X
SD	LU	X	X	100.00	100	X	X
	MOD	X	X	100.00	100	X	X
	ExemA	X	X	100.00	100	X	X
	ExemB	X	X	100.00	100	X	X
BEST	LU (0.006, exp.)	11.27	X	90.03	12.53	X	X
	LU (0.02, exp.)	22.06	X	63.30	11.99	X	X
	LU (1, exp.)	31.62	1.32	82.31	19.04	X	X
	LU (0.006, coal.)	1.35	X	98.57	62.56	X	X
	LU (0.02, coal.)	5.36	X	88.83	59.94	X	X
	LU (1, coal.)	1.62	X	94.03	96.77	X	X
	MOD (0.006, exp.)	12.58	X	59.90	22.66	X	X
	MOD (0.02, exp.)	23.95	15.46	88.74	26.13	X	X
	MOD (1, exp.)	61.27	2.83	84.38	18.35	X	X
	MOD (0.006, coal.)	20.56	2.95	99.74	48.59	X	X
	MOD (0.02, coal.)	21.85	2.51	99.17	57.92	X	X
	MOD (1, coal.)	X	X	97.07	69.56	X	X
	ExemA (0.006, exp.)	8.33	25.00	89.81	7.65	8.33	X
	ExemA (0.02, exp.)	5.68	X	75.65	10.00	X	X
	ExemA (1, exp.)	16.38	8.52	72.11	10.00	X	X
	ExemA (0.006, coal.)	10.00	X	100.00	1.37	X	X
	ExemA (0.02, coal.)	30.00	X	96.97	X	X	X
	ExemA (1, coal.)	1.50	0.92	63.23	6.08	X	X
	ExemB (0.006, exp.)	X	X	80.00	X	X	X
	ExemB (0.02, exp.)	X	X	58.33	X	X	X
	ExemB (1, exp.)	17.84	2.53	58.06	1.33	X	X
	ExemB (0.006, coal.)	15.34	X	100.00	10.00	X	X
	ExemB (0.02, coal.)	20.00	X	90.00	X	X	X
	ExemB (1, coal.)	3.60	X	84.28	12.22	X	X

in their analysis of relationships within the rodent genus *Thomomys*. However, in contrast to the *Thomomys* study, support for nodes within *Neodiprion* tended to decrease rather than increase as individuals were added to BEST analyses (Fig. 3).

In summary, species-tree estimates were heavily dependent on both analysis method and sampling. The impact of analysis method can be seen by examining all phylogenies obtained for a given data set—for all data sets examined, each method recovered a different species phylogeny (Figs. 2 and 3; Supplementary Figs. 5 to 7). Likewise, the impact of sampling becomes obvious when all topologies obtained by a given method are compared (Figs. 2 and 3; Supplementary Figs. 5 to 7).

Comparison with Morphologically Based Hypotheses

Table 4 summarizes the agreement between each method/data set combination and Ross's (1955) six hypothesized *lecontei* group clades (Fig. 1). One clade—the *pinusrigidae* complex—was present in all MDC and SD analyses and received high posterior probabilities in Bayesian analyses (CMC and BEST). Although posterior probabilities tended to be lower than for the *pinusrigidae* complex, the *lecontei* complex was nevertheless present/not rejected by most analysis/data set combinations. Support across methods/data sets for the *virginianus* and *pratti* complexes was more mixed. The *virginianus* complex was not recovered in any MDC or SD trees but could not be statistically rejected in the majority of CMC and BEST analyses. In contrast, the *pratti* complex was rejected/absent in all CMC, MDC, and SD analyses, but several BEST analyses could not statistically reject this clade. Finally, two clades—the non-*abbottii* group and the *abbottii* complex—were rejected/absent in all (or nearly all) analyses.

Outside of Ross's (1955) hypotheses, several additional areas were concordant across methods. Most notably, with the exception of two SD trees (MOD and LU data sets) and two MDC trees (ExemA and MOD), all methods/data sets recovered two major clades within *Neodiprion* (clades "A" and "B" in Figs. 2 and 3); in BEST and CMC analyses, these clades received high support. In addition, relationships within the *pinusrigidae* complex were identical in nearly all analyses and relationships between *N. maurus* + *N. pratti* and between *N. abbotii* + *N. nigroscutum* + *N. species 1* were recovered in many of the Bayesian analyses (Figs. 2 and 3).

Comparison of results obtained from different methods and data sets also reveals that some areas of the *Neodiprion* tree are more impacted by sampling than others. Specifically, for all but the CMC method, relationships within clade A tended to depend more heavily on the individuals sampled than did relationships within clade B (this pattern can be seen by looking at the distribution of starred nodes in Figs. 2 and 3). A difference in the amount of phylogenetically informative variation does not appear to explain these differences because interspecific genetic distances are similar in these two clades (Table 5 and Fig. 4; uncorrected p distances were

TABLE 5. Average interspecific genetic distance, ancestral θ , divergence time, and population migration rate for clades A and B (see Figs. 2 and 3). Pairwise genetic distances (p) are uncorrected and were calculated in Mesquite. Remaining parameter estimates are from Linnen and Farrell (2007; see also Linnen, 2007) and were obtained using the program IM (Hey and Nielsen, 2004). Ancestral θ (per-locus) and divergence time estimates are scaled by the mutation rate u , and population migration rates are averaged across nuclear loci. Only pairwise comparisons that returned complete distributions for all IM parameters are included in clade averages.

Clade	Genetic distance (p)	Ancestral θ (4 Nu)	Divergence time (tu)	Migration rate (2 Nm)
A	0.00775	5.56	5.32	0.368
B	0.00732	2.08	7.70	0.0510

calculated for all species pairs in each clade in Mesquite v. 1.12; these distances were not significantly different according to a one-tailed Mann-Whitney *U*-test, $U = 495.0$, $p = 0.50$). Three additional factors that could explain the observed differences between clades A and B are ancestral population sizes, divergence times, and gene-flow rates. Specifically, larger ancestral population sizes, shorter divergence times, and higher levels of interspecific gene flow would all be expected to reduce the match between gene trees and the underlying species trees in, and therefore make phylogenetic inference more difficult for, the A clade. Estimates for these parameters were available from a previous investigation that used the program IM (Hey and Nielsen, 2004) to compare mitochondrial and nuclear gene flow between *lecontei* group species pairs (Linnen and Farrell, 2007; Linnen, 2007). Using only those comparisons that returned complete posterior distributions for all parameters in the Isolation with Migration model (Nielsen and Wakeley, 2001; Hey and Nielsen, 2004), we found that clades A and B differed significantly with respect to ancestral population size (θ_A ; $U = 173$; $p = 0.042$) and nuclear gene flow rates (2Nm; $U = 178$; $p = 0.027$) but not divergence times (tu; $U = 161$; $p = 0.099$; all tests are one-tailed Mann-Whitney *U*-tests; see Fig. 4 and Table 5).

DISCUSSION

Incomplete lineage sorting, introgression, and low levels of genetic variation, all of which are expected in groups that have diverged rapidly and recently, complicate phylogenetic inference in the genus *Neodiprion*. In this study, we employed four strategies for species-tree inference, including a novel approach that combines monophyly constraints with concatenation to permit the inclusion of multiple individuals per species. We found that sampling of individuals, choice of method, and, for the BEST method, choice of priors all impacted our results. Comparing methods, we found that the CMC method was the least sensitive to taxonomic sampling (i.e., for the most part, the same relationships were recovered by all four data sets). We also found that, although interspecific genetic variation is low due to the recent divergence of the *lecontei* group, incomplete lineage sorting and interspecific gene flow appear to be

the main factors complicating species-tree inference in *Neodiprion*. Despite these difficulties, we identified multiple clades that were robust to both method choice and sampling and several of these clades corresponded to relationships supported by morphological evidence. Potential explanations for these patterns, as well as implications for *Neodiprion* phylogeny, are discussed in more detail below.

Comparison of Species-Tree Methods

Of all the methods examined, CMC was the least sensitive to sampling—12 out of 17 nodes were identical across all data sets, compared to 5 to 7 out of 17 for the MDC, SD, and BEST methods (Figs. 2 and 3). For the MDC and SD methods, one possible explanation for the strong dependency on sampling we observed is that these methods do not account for uncertainty in individual gene-tree estimates. Several observations are consistent with this explanation, including low levels of variation (Table 2), low bootstrap support in ML gene-tree estimates (Supplementary Figs. 2 to 4), and the observation that MDC estimates were heavily influenced by whether polytomies were automatically resolved (Table 3; Fig. 2; Supplementary Fig. 6). For the BEST method, one possible explanation for the observed dependence on both sampling and priors is that the majority of data set/prior combinations failed to converge on a single species-tree solution—it is possible, then, that additional data, longer run times, and/or different run settings might have produced a well-supported topology that was robust to sampling and priors. In support of this argument, agreement between the two data sets that did converge was closer (10 shared nodes) to what was observed for the CMC method. However, these explanations do not shed light on our observation that, despite similar levels of interspecific genetic divergence, clade A appeared to be much more dependent on sampling than clade B for the MDC, SD, and BEST methods but not the CMC method (Figs. 2 and 3).

Impact of gene flow on species-tree methods.—Another reason why the CMC method was less dependent on taxonomic sampling than the BEST, SD, and MDC methods might be that the latter are more sensitive to violations of the assumption of no gene flow. In support of this hypothesis, gene-flow rates appear to be higher in the A clade, within which these three methods seem to have the most difficulty resolving relationships, than in the B clade (Fig. 4; Table 5). All four methods assume that there has been no gene flow following speciation, and this assumption is clearly violated in *Neodiprion* (Linnen and Farrell, 2007); however, low levels of gene flow may impact these methods in different ways. First, as Maddison and Knowles (2006) point out, the SD method may be particularly prone to misinterpret recently introgressed alleles as evidence for close relationships between species. Second, when the MDC method is used, introgression events will be erroneously interpreted as variation shared between species due to deep coalescence. Different individuals may uncover evidence of different introgression

(and coalescent) events and, therefore, different data sets may result in dissimilar MDC trees.

Third, the model implemented in BEST assumes that all gene divergences pre-date speciation events (i.e., no interspecific gene flow); violations of this assumption will place strong restrictions on species-tree branch lengths and may mislead species-tree inference (Liu and Pearl, 2007). More specifically, when interspecific gene flow has been prevalent (i.e., gene-tree divergences post-date species divergences), BEST is expected to favor species trees with large effective population sizes and short divergence times. This may explain, in part, why higher values of θ gave the best results for the larger data sets (Table 3): as sampling was increased within species, more introgression events were uncovered, which caused BEST to prefer models with larger effective population sizes. Moreover, because the signal of past introgression is dependent on the number and identity of individuals sampled, different data sets may place different restrictions on species-tree branch lengths and, therefore, produce different BEST topologies.

Fourth, and finally, the CMC method differs from the BEST and MDC methods in that data are concatenated, which means that phylogenetic signal that is consistent across loci will be retained, whereas instances of introgression at single loci and/or single individuals (or incomplete lineage sorting for that matter) may be overcome by a more dominant, presumably phylogenetic, signal in the data (Baker and DeSalle, 1997; de Queiroz et al., 1995; Kluge, 1989; Wiens, 1998; Lerat et al., 2003; de Queiroz and Gatesy, 2007). This swamping effect may explain why the CMC method was more robust to sampling compared to the other methods, but because we do not know the “true tree,” we do not know how well this decreased sensitivity corresponds to accuracy. It is possible, for example, that the dominant signal in the data does not match the species tree. However, as we discuss below, the inclusion of multiple individuals per species should, in theory, improve the accuracy of the CMC method.

Concatenation, anomalous gene trees, and taxon sampling.—A critical assumption of the concatenation approach is that the predominant phylogenetic signal in the data accurately reflects the underlying species tree. Unfortunately, when internal branches in the species tree are sufficiently short in comparison to external branches, this assumption is likely to be violated because gene trees that do not match the species tree are more probable than matching gene trees (these non-matching trees have been dubbed “anomalous gene trees” or “AGTs”; Degnan and Rosenberg, 2006; Rosenberg and Tao, 2008). Not surprisingly, concatenation has been shown to perform poorly under conditions conducive to AGTs (Edwards et al., 2007; Kubatko and Degnan, 2007). However, only a single individual was sampled per species in these studies, and, as some authors have pointed out, sampling multiple individuals per species might lessen the impact of AGTs (Degnan and Rosenberg, 2006; Kubatko and Degnan, 2007). This suggestion is supported by simulation studies that have

demonstrated that increasing the number of individuals sampled per species increases the match between gene trees and species trees (Takahata, 1989; Rosenberg, 2002) and the accuracy of species-tree methods (Maddison and Knowles, 2006). When multiple individuals are sampled, however, species-trees estimated using the concatenation approach become difficult to interpret because recently diverged species may be non-monophyletic at most loci (Hudson and Coyne, 2002; Funk and Omland, 2003; Carstens and Knowles, 2007). As we have demonstrated here, this ambiguity can be removed by incorporating monophyly constraints into the analysis of concatenated data. In summary, we expect the CMC method to perform well in species-tree estimation even when gene flow and incomplete lineage sorting are both prevalent because (1) shared signal in the data swamps out non-phylogenetic noise, (2) the probability that this shared signal results from an anomalous gene tree is lessened by the inclusion of multiple individuals, and (3) monophyly constraints produce a result that can be interpreted as a species tree. We acknowledge, however, that simulation studies are needed to test these intuitions regarding the performance of the CMC method.

Assumptions of Species-Tree Methods

All species-tree methods used in this study make two assumptions: (1) species are accurately delimited according to whatever species definition one chooses, and (2) the species tree is bifurcating. The first assumption is inherent in any attempt to construct a species tree for individuals sampled from nature (including exemplar approaches). The second assumption is likely valid for many groups of organisms. Even when interspecific gene flow is present, species histories can be described as bifurcating if most intraspecific variation stems from novel mutation and ancestral variation from a single population. To account for these situations, species-tree methods could incorporate models that describe the occasional leakage of alleles across species boundaries into an otherwise bifurcating history (e.g., Nielsen and Wakeley, 2001; Hey and Nielsen, 2004, 2007). In contrast, one decidedly nonbifurcating process is hybrid species formation, in which a significant portion of intraspecific variation originates from multiple ancestral populations (parental species). There is growing evidence that hybrid speciation has occurred in a wide range of plant and animal lineages (Arnold, 2006; Mallet, 2007), and, in *Neodiprion*, a hybrid origin has been proposed for one species (*N. merkei*; Ross, 1961). Although this hypothesis remains to be tested (more *N. merkei* individuals and more loci are needed), it suggests that a dichotomously branching species tree may not be an accurate representation of *Neodiprion* history. At present, however, our assumption that the *Neodiprion* tree is bifurcating is supported by three observations: (1) interspecific nuclear gene flow is low ($2Nm < 1$) (Linnen and Farrell, 2007; Table 5), (2) species remain morphologically distinct in spite of this gene flow, and (3) the hierarchical model implemented

in BEST explains the nuclear data better than a model that does not take shared history into account (Table 3). Nevertheless, this assumption should be reevaluated as new evidence and new methods become available. Fortunately, model-based species-tree methods provide a framework within which multiple types of diversification models (including non-bifurcating ones) could be incorporated and tested.

Implications for *Neodiprion* Phylogeny

Although we have discussed issues of species-tree estimation at length, the ultimate goal of this study was to generate a phylogenetic estimate to be used in future comparative studies of *Neodiprion* speciation. Given currently available methods and data, we propose that the phylogeny in Figure 2a is our best estimate of *lecontei* group relationships. We choose this phylogeny because (1) the CMC method was much less sensitive to taxonomic sampling than other species-tree methods, and (2) this phylogeny was estimated using all available samples (Rosenberg, 2002). Like any phylogenetic hypothesis, this species tree should be reevaluated as new data and methods become available. Also, any study that relies on this phylogeny should consider the sensitivity of conclusions to the presence/absence of clades that were unstable across data sets and/or analysis methods. Fortunately, by combining results obtained from these four diverse methods, we have identified several portions of the *Neodiprion* phylogeny that are robust to taxonomic sampling and method choice (Kim, 1993; Miyamoto and Fitch, 1995; Knowles and Carstens, 2007). We found, for example, that almost none of the analyses supported Ross's hypothesis that *N. compar*, *N. nigros-cutum*, and *N. abbotii* form a monophyletic group (Fig. 1; Table 4)—this suggests that the larval coloration Ross (1955) used to group these species is the result of convergent evolution, not shared ancestry. Also, all (or nearly all) method/data set combinations recovered the *pinus-rigidae* complex, *lecontei* complex, and several additional groupings (e.g., clades A and B in Figs. 2 and 3).

CONCLUSIONS

Our efforts to estimate a species tree for *Neodiprion* illustrate the challenge that is faced when a phylogenetic estimate is sought for a group that diverged rapidly and recently. Fortunately, recent work has shown that there is considerable information that can be extracted from gene trees, even (or perhaps especially) when they are discordant with one another and contain non-monophyletic species (e.g., Hey and Nielsen, 2004, 2007; Maddison and Knowles, 2006; Carstens and Knowles, 2007; Belfiore et al., 2008). To take advantage of this information, multiple individuals must be sampled per species and phylogenetic methods must focus on the processes that have shaped variation within and between species. However, species-tree methods are still in their infancy and we are a long way from models that adequately describe diversification in real organisms. In particular, although methods that include stochastic models of lineage sorting

have been developed (Liu and Pearl, 2006; Edwards et al., 2007; Carstens and Knowles, 2007), these need to be expanded to include post-divergence gene flow and diversification histories that are not strictly bifurcating. Until such methods are available, we suggest using multiple species-tree methods to inform confidence in species-tree estimates. We agree with others who have argued that a different approach to phylogenetic analysis is required when the species tree, not the contained gene trees, is the parameter of interest (e.g., Maddison, 1997; Maddison and Knowles, 2006; Carstens and Knowles, 2007; Edwards et al., 2007; Liu and Pearl, 2007). We look forward to future developments in this exciting field and their application to long-standing questions regarding the ecology and geography of speciation in groups that have undergone rapid diversification (e.g., crater lake cichlids, Hawaiian silverswords, *Anolis* lizards, columbines, and Galápagos finches).

ACKNOWLEDGEMENTS

We are grateful to B. Bossert, D. Haig, N. Pierce, and J. Wakeley for comments and discussion on an earlier version of the manuscript, to D. Smith for assistance with identifications, and to L. Liu and S. Edwards for advice on using the program BEST. We also thank L. Kubatko, J. Sullivan, and an anonymous reviewer for suggestions that greatly improved the manuscript. Funding for this research was provided by a Graduate Research Fellowship and a Dissertation Improvement Grant (DEB-0308815) from the National Science Foundation, a Science to Achieve Results Graduate Fellowship from the Environmental Protection Agency, the Putnam Expeditionary Fund at the Museum of Comparative Zoology, the Theodore Roosevelt Memorial Fund at the American Museum of Natural History, and the Department of Organismic and Evolutionary Biology at Harvard University.

REFERENCES

- Alexander, R. D., and R. S. Bigelow. 1960. Allochronic speciation in field crickets, and a new species, *Acheta veletis*. *Evolution* 14:334–346.
- Arnett, R. H. 1993. American insects: A handbook of the insects of America north of Mexico. Sandhill Crane Press, Gainesville, Florida.
- Arnold, M. L. 2006. Evolution through genetic exchange. Oxford University Press, New York.
- Atwood, C. E., and O. Peck. 1943. Some native sawflies of the genus *Neodiprion* attacking pines in eastern Canada. *Can. J. Res. D* 21:109–144.
- Baker, R. H., and R. DeSalle. 1997. Multiple sources of character information and the phylogeny of Hawaiian *Drosophilids*. *Syst. Biol.* 46:654–673.
- Becker, G. C. 1965. A biological-taxonomic study of the *Neodiprion virginianus* complex in Wisconsin. PhD dissertation. University of Wisconsin, Madison.
- Becker, G. C., and D. M. Benjamin. 1967. Biology of *Neodiprion nigroscutum* (Hymenoptera: Diprionidae) in Wisconsin. *Can. Entomol.* 99:146–159.
- Becker, G. C., R. C. Wilkinson, and D. M. Benjamin. 1966. Taxonomy of *Neodiprion rugifrons* and *N. dubiosus* (Hymenoptera: Tenthredinoidea: Diprionidae). *Ann. Entomol. Soc. Am.* 59:173–178.
- Belfiore, N. M., L. Liang, and C. Moritz. 2008. Multilocus phylogenetics of a rapid radiation in the genus *Thomomys* (Rodentia: Geomyidae). *Syst. Biol.* 57:294–310.
- Buschbom, J., and D. Barker. 2006. Evolutionary history of vegetative reproduction in *Porpidia* s.l. (lichen-forming Ascomycota). *Syst. Biol.* 55:471–484.
- Bush, G. L. 1975a. Modes of animal speciation. *Annu. Rev. Ecol. Syst.* 6:339–364.
- Bush, G. L. 1975b. Sympatric speciation in phytophagous parasitic insects. Pages 187–207 in *Evolutionary strategies of parasitic insects and mites* (P. W. Price, ed.). Plenum, New York.
- Carstens, B. C., and L. L. Knowles. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: An example from *Melanoplus* grasshoppers. *Syst. Biol.* 56:400–411.
- Danforth, B. N., and S. Q. Ji. 1998. Elongation factor-1 alpha occurs as two copies in bees: Implications for phylogenetic analysis of EF-1 alpha sequences in insects. *Mol. Biol. Evol.* 15:225–235.
- Danforth, B. N., H. Sauquet, and L. Packer. 1999. Phylogeny of the bee genus *Halictus* (Hymenoptera: Halictidae) based on parsimony and likelihood analyses of nuclear EF-1 alpha sequence data. *Mol. Phylogenet. Evol.* 13:605–618.
- Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2:762–768.
- de Queiroz, A. 1993. For consensus (sometimes). *Syst. Biol.* 42:368–372.
- de Queiroz, A., M. J. Donoghue, and J. Kim. 1995. Separate versus combined analysis of phylogenetic evidence. *Annu. Rev. Ecol. Syst.* 26:657–681.
- de Queiroz, A., and J. Gatesy. 2007. The supermatrix approach to systematics. *Trends Ecol. Evol.* 22:34–41.
- Dixon, W. N. 2004. Pine sawfly larvae, *Neodiprion* spp. (Insecta: Hymenoptera: Diprionidae). University of Florida IFAS Extension, Gainesville, Florida.
- Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenation. *Proc. Natl. Acad. Sci. USA* 104:5936–5941.
- Funk, D. J. 1999. Molecular systematics of cytochrome oxidase I and 16S from *Neochlamisus* leaf beetles and the importance of sampling. *Mol. Biol. Evol.* 16:67–82.
- Funk, D. J., and K. E. Omland. 2003. Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annu. Rev. Ecol. Syst.* 34:397–423.
- Gadagkar, S. R., M. S. Rosenberg, and S. Kumar. 2005. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *J. Exp. Zool. B* 304B:64–74.
- Gatesy, J., and R. H. Baker. 2005. Hidden likelihood support in genomic data: Can forty-five wrongs make a right? *Syst. Biol.* 54:483–492.
- Ghent, A. W., and D. R. Wallace. 1958. Oviposition behavior of the Swaine jack-pine sawfly. *For. Sci.* 4:264–272.
- Hedtke, S. M., T. M. Townsend, and D. M. Hillis. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst. Biol.* 55:522–529.
- 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.
- Hey, J., and R. Nielsen. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc. Natl. Acad. Sci. USA* 104:2785–2790.
- Hudson, R. R., and J. A. Coyne. 2002. Mathematical consequences of the genealogical species concept. *Evolution* 56:1557–1565.
- Hudson, R. R., and M. Turelli. 2003. Stochasticity overrules the “three-times rule”: Genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. *Evolution* 57:182–190.
- Huelsenbeck, J. P., J. J. Bull, and C. W. Cunningham. 1996. Combining data in phylogenetic analysis. *Trends Ecol. Evol.* 11:152–158.
- Huelsenbeck, J. P., and B. Rannala. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.* 53:904–913.
- Kass, R. E., and A. E. Raftery. 1995. Bayes factors. *J. Am. Stat. Assoc.* 90:773–795.
- Kim, J. 1993. Improving the accuracy of phylogenetic estimation by combining different methods. *Syst. Biol.* 42:331–340.
- Kluge, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Syst. Zool.* 38:7–25.
- Knerer, G. 1984. Morphological and physiological clines in *Neodiprion pratti* (Dyar) (Symphyta: Diprionidae) in eastern North America. *Z. Angew. Entomol.* 97:9–21.
- Knerer, G., and C. E. Atwood. 1972. Evolutionary trends in subsocial sawflies belonging to the *Neodiprion abietis* complex (Hymenoptera: Tenthredinoidea). *Am. Zool.* 12:407–418.
- Knerer, G., and C. E. Atwood. 1973. Diprionid sawflies: Polymorphism and speciation. *Science* 179:1090–1099.

- Knerer, G., and R. C. Wilkinson. 1990. The biology of *Neodiprion pratti* (Dyar) (Hymenoptera: Diprionidae), a winter sawfly in west Florida. *Z. Angew. Entomol.* 109:448–456.
- Knowles, L. L., and B. C. Carstens. 2007. Estimating a geographically explicit model of population divergence. *Evolution* 61:477–493.
- Kubatko, L. S., and J. H. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56:17–24.
- Lecointre, G., H. Philippe, H. L. V. Le, and H. Le Guyader. 1993. Species sampling has a major impact on phylogenetic inference. *Mol. Phylogenet. Evol.* 2:205–224.
- Lerat, E., V. Daubin, and N. A. Moran. 2003. From gene trees to organismal phylogeny in prokaryotes: The case of the α -Proteobacteria. *PLoS Biol.* 1:101–109.
- Linnen, C. R. 2007. Adaptation, speciation, and hybridization in the genus *Neodiprion* (Hymenoptera: Diprionidae). PhD dissertation, Harvard University, Cambridge.
- Linnen, C. R., and B. D. Farrell. 2007. Mitonuclear discordance is caused by rampant mitochondrial introgression in *Neodiprion* (Hymenoptera: Diprionidae) sawflies. *Evolution* 61:1417–1438.
- Linnen, C. R., and B. D. Farrell. 2008. Phylogenetic analysis of nuclear and mitochondrial genes reveals evolutionary relationships and mitochondrial introgression in the *sertifer* species group of the genus *Neodiprion* (Hymenoptera: Diprionidae). *Mol. Phylogenet. Evol.* 48:240–257.
- Liu, L., and D. K. Pearl. 2006. Reconstructing posterior distributions of a species phylogeny using estimated gene tree distributions. Biosciences Institute Technical Report #53. Biosciences Institute Technical Report, The Ohio State University, Columbus, Ohio.
- Liu, L., and D. K. Pearl. 2007. Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.* 56:504–514.
- Liu, L., D. K. Pearl, R. T. Brumfield, and S. V. Edwards. 2008. Estimating species trees using multiple-allele DNA sequence data. *Evolution* 62:2080–2091.
- Maddison, W. P. 1997. Gene trees in species trees. *Syst. Biol.* 46:523–536.
- Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 55:21–30.
- Maddison, W. P., and D. R. Maddison. 2000. *MacClade*. Version 4.0. Sinauer Associates, Sunderland, Massachusetts.
- Maddison, W. P., and D. R. Maddison. 2006. *Mesquite: A modular system for evolutionary analysis*. Version 1.12. <http://mesquiteproject.org>.
- Mallet, J. 2007. Hybrid speciation. *Nature* 446:279–283.
- Marshall, D. C., C. Simon, and T. R. Buckley. 2006. Accurate branch length estimation in partitioned Bayesian analyses requires accommodation of among-partition rate variation and attention to branch length priors. *Syst. Biol.* 55:993–1003.
- Miller, R. E., T. R. Buckley, and P. S. Manos. 2002. An examination of the monophyly of morning glory taxa using Bayesian phylogenetic inference. *Syst. Biol.* 51:740–753.
- Miyamoto, M. M., and W. M. Fitch. 1995. Testing species phylogenies and phylogenetic methods with congruence. *Syst. Biol.* 44:64–76.
- Nielsen, R., and J. Wakeley. 2001. Distinguishing migration from isolation: A Markov chain Monte Carlo approach. *Genetics* 158:885–896.
- Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. *Syst. Biol.* 53:47–67.
- Nyman, T., A. G. Zinoviev, V. Vikberg, and B. D. Farrell. 2006. Molecular phylogeny of the sawfly subfamily Nematininae (Hymenoptera: Tenthredinidae). *Syst. Entomol.* 31:569–583.
- Philippe, H. 1997. Rodent monophyly: Pitfalls of molecular phylogenies. *J. Mol. Evol.* 45:712–715.
- Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798.
- Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
- Ronquist, F., J. P. Huelsenbeck, and P. van der Mark. 2005. MrBayes 3.1 manual. Available at <http://mrbayes.csit.fsu.edu/manual.php>
- Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. *Theor. Popul. Biol.* 61:225–247.
- Rosenberg, N. A., and R. Tao. 2008. Discordance of species trees with their most likely gene trees: The case of five taxa. *Syst. Biol.* 57:131–140.
- Ross, H. H. 1955. The taxonomy and evolution of the sawfly genus *Neodiprion*. *For. Sci.* 1:196–209.
- Ross, H. H. 1961. Two new species of *Neodiprion* from southeastern North America (Hymenoptera: Diprionidae). *Ann. Entomol. Soc. Am.* 54:451–453.
- Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497.
- Sheehan, K. A., and D. L. Dahlsten. 1985. Bionomics of *Neodiprion* species on white fir in northeastern California. *Hilgardia* 53:1–24.
- Smith, D. R. 1988. A synopsis of the sawflies (Hymenoptera, Symphyta) of America south of the United States: Introduction, Xyelidae, Pamphiliidae, Cimbicidae, Diprionidae, Xiphodriidae, Siricidae, Orussidae, Cephidae. *Syst. Entomol.* 13:205–261.
- Smith, D. R., and M. R. Wagner. 1986. Recognition of two species in the pine feeding *Neodiprion fulviceps* complex (Hymenoptera: Diprionidae) of western United States. *Proc. Entomol. Soc. Wash.* 88:215–220.
- Strong, D. R., J. H. Lawton, and R. Southwood. 1984. *Insects on plants: Community patterns and mechanisms*. Harvard University Press, Cambridge, Massachusetts.
- Swofford, D. L. 2000. *PAUP*: Phylogenetic analysis using parsimony (*and other methods)*. Sinauer, Sunderland, Massachusetts.
- Takahata, N. 1989. Gene genealogy in three related populations: Consistency probability between gene and population trees. *Genetics* 122:957–966.
- Tauber, C. A., and M. J. Tauber. 1981. Insect seasonal cycles: Genetics and evolution. *Annu. Rev. Ecol. Syst.* 12:281–308.
- Wiens, J. J. 1998. Combining data sets with different phylogenetic histories. *Syst. Biol.* 47:568–581.
- Wilson, L. F. 1977. *A guide to insect injury of conifers in the Lake States*. USDA Forest Service, Washington, DC.
- Wright, S. J. 1931. Evolution in Mendelian populations. *Genetics* 16:97–159.
- Zwickl, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD dissertation, The University of Texas at Austin, Austin.

First submitted 18 February 2008; reviews returned 21 May 2008;

final acceptance 25 October 2008

Associate Editor: Laura Kubatko