CHAPTER 9

SPECIES TREE ESTIMATION FOR COMPLEX DIVERGENCE HISTORIES: A CASE STUDY IN NEODIPRION *SAWFLIES*

Catherine Linnen

9.1 INTRODUCTION

A fundamental goal in evolutionary biology is to understand how and why populations diverge. Achieving this goal in a given taxon requires an accurate depiction of the branching of its populations through time via the process of speciation, or its "species tree." While this history is most often estimated using DNA sequence data, genes have histories that are distinct from and need not necessarily match the underlying species tree due to processes such as incomplete lineage sorting and introgression (Maddison 1997; see Chapter 1). As several chapters in this book demonstrate, recognizing species trees and gene trees as distinct, but related, entities can improve our estimation of the former from the latter. This is especially true for groups that have radiated rapidly and/or recently, which are arguably some of the most exciting lineages in which to study the evolutionary process (e.g., crater lake cichlids, Hawaiian silverswords, *Anolis* lizards, columbines, and Galápagos finches). However, improved accuracy of historical inference is not the only rationale for distinguishing between species trees and gene trees in phylogenetic estimation. A second, and perhaps equally important, motivation for this distinction is that together, these two histories can tell us much more about evolution than either can in isolation (Fig. 9.1).

In this chapter, I describe my efforts to estimate a species tree for the sawfly genus *Neodiprion* Rohwer. I chose this group as a study system for two reasons: (1) *Neodiprion* have life history features (described below), shared by many other plant-feeding insects, that are thought to influence the tempo and mode of diversification, and (2) ample ecological data are available for the genus (largely because several species are economically important pests; Arnett 1993). My primary motivation for constructing a *Neodiprion* species tree was to test a priori hypotheses about the ecology and biogeography of speciation. In retrospect, given that the life history features that drew me to *Neodiprion* are thought to be conducive to rapid divergence with gene flow, I should have anticipated that obtaining this estimate would be difficult. Species tree estimation was indeed complicated by multiple factors, including extensive mitochondrial introgression, low levels of nuclear gene flow, and incomplete lineage sorting. At present, no method deals with all of these issues simultaneously; I therefore used multiple methods, none of which are ideal, to

Estimating Species Trees: Practical and Theoretical Aspects, Edited by L. Lacey Knowles and Laura S. Kubatko

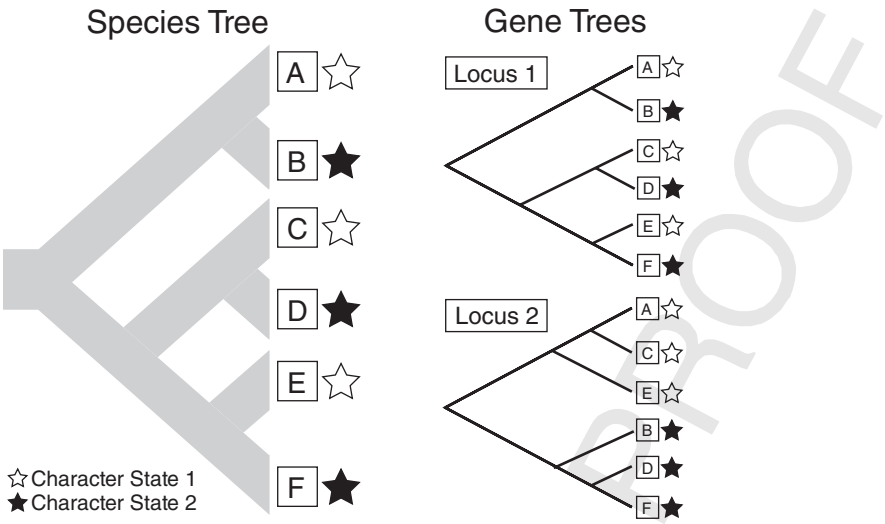

Figure 9.1 Together, species trees and gene trees can tell us more about the evolutionary process than either can in isolation. Here, character states for a phenotype of interest have been mapped onto a species tree for species A–F. With this information, several questions can be addressed, including the following: (1) Are changes in the trait associated with speciation events? (2) Do character states correlate with the environment? (3) Do other traits show correlated evolution? Additional questions can be addressed when gene trees (e.g., locus 1 and locus 2) are also available. For example, we can investigate how reproductive isolation accumulates across the genome: while the match between locus 1 and the species tree suggests that this locus does not cross species boundaries readily, gene flow might be less constrained at locus 2. Moreover, if either locus contributes to the phenotype of interest, we can infer the source adaptive variation. For example, while the species tree suggests that there have been multiple independent transitions between the two character states at the population/species level (perhaps as the result of similar ecological pressures), only the locus 1 gene tree would indicate that this pattern results from independent genetic origins. These are just two of many possible ways in which species and gene trees can be combined to reveal novel insights.

14

disentangle this complicated history. Below, I describe these analyses in detail. In addition, while my initial goal was to obtain a species tree for comparative tests of speciation mode, gene tree estimates revealed insights that would have been unavailable from the species tree alone—an example is given at the end of this chapter.

9.2 STUDY SYSTEM: *NEODIPRION* SAWFLIES

Like many phytophagous (plant-feeding) insects, *Neodiprion* sawflies (order: Hymenoptera; family: Diprionidae) are intimately associated with their host plants throughout their life cycle: eggs are embedded within the host plant tissue; larvae often spend their entire feeding period on their natal host; cocoons are spun on or beneath the host; and mating occurs on the host plant (reviewed in Coppel and Benjamin 1965; Knerer and Atwood 1973). This tight association between *Neodiprion* and their hosts is thought to be responsible for the restricted host range of individual species (Bjorkman and Larsson 1991; Knerer and Atwood 1973; McCullough and Wagner 1993). Specifically, all *Neodiprion* species feed on host plants in the family Pinaceae; most species are found only on hosts

J3

in the genus *Pinus*; and many species will feed only on a single *Pinus* species. Even the more polyphagous *Neodiprion* species (i.e., those that feed on multiple *Pinus* species) tend to have preferences for particular pine species or habitat types (Becker 1965; Benjamin 1955; Coppel and Benjamin 1965; McMillin and Wagner 1993; Smith 1979; Wilson et al. 1992). Because they are strict host specialists and mate on their hosts, it has been hypothesized that changes in host use may initiate the formation of new species in *Neodiprion*, often in the presence of gene flow (Alexander and Bigelow 1960; Bush 1975a; Bush 1975b; Ghent and Wallace 1958; Knerer and Atwood 1972, 1973; Strong et al. 1984; Tauber and Tauber 1981). For this reason, *Neodiprion* is an exciting system in which to explore the ecology and biogeography of phytophage speciation, and, while I do not discuss comparative tests of speciation mode in this chapter, it is with this ultimate goal that I first set out to estimate a species tree for *Neodiprion*.

9.3 SAMPLING STRATEGY

Because hypothesis testing required detailed information on host use and geography and species tree methods required a priori species designations, I focused on the well-studied eastern North American *lecontei* species group, which forms a monophyletic clade nested within North American *Neodiprion* (Linnen and Farrell 2007, 2008b). In total, I collected 19 of 21 *lecontei* group species (the two missing species are found only in Cuba) and sampled multiple populations for most species.

When dividing my sequencing effort between individuals and loci, I decided to maximize the former because simulation studies have shown that, for recently diverged groups such as the *lecontei* species group, increasing the number of individuals sampled per species can increase both the match between gene trees and species trees (Rosenberg 2002; Takahata 1989) and the accuracy of species tree methods (Maddison and Knowles 2006). Moreover, when there is geographic structure within species, as is the case for most organisms (Mayr 1942, 1963), an even greater number of individuals (populations) may be required to achieve accurate species tree estimates (Rosenberg 2002; Wakeley 2000). In choosing specimens to sequence, I therefore included multiple populations whenever possible and chose populations so as to maximize the geographical and ecological variation I had sampled for each species. Because some species are more common and/or have larger ranges than others, this resulted in an uneven number of individuals (mean: 6.6, range: 1–14) per species. Having chosen to include as many individuals/populations as possible, I sequenced a moderate number of loci, which included one mitochondrial locus (COI/COII), and three nuclear loci (Ef1α, CAD, and an anonymous nuclear locus, ANL43). All analyses described below utilized this data set (four loci, 126 individuals sampled from 19 *lecontei* group species, plus one *sertifer* group species [*Neodiprion autumnalis*] as an out-group). At the end of the chapter, I discuss this sampling strategy in light of my experience with species tree methods.

9.4 DETERMINING THE SOURCE OF MITONUCLEAR DISCORDANCE

When I began generating sequence data for *Neodiprion*, I noticed that morphologically distinct species often shared the same or very similar mitochondrial haplotypes. Initially, I thought this was due to recent divergence and incomplete lineage sorting, but as I started sequencing nuclear genes, I quickly realized that species that shared mitochondrial

haplotypes were often quite divergent at nuclear loci. The profound discordance I observed between mitochondrial and nuclear gene trees suggested that one class of markers might be unsuitable for recovering the *Neodiprion* species tree or, alternatively, that a bifurcating tree would not provide an adequate description of *Neodiprion* history. Therefore, before I could estimate a species tree, I needed to determine the source of this discordance.

Two lines of evidence suggested that incomplete lineage sorting alone could not account for mitonuclear discordance in *Neodiprion*. First, compared to the mitochondrial locus, nuclear loci recovered an equal or greater number of monophyletic species (i.e., alleles sampled from a single species were more closely related to one another than to alleles from any other species). This pattern is not expected under incomplete lineage sorting because mitochondrial genes have a smaller effective population size on average than nuclear loci, and coalescent theory predicts an inverse relationship between effective population size and the speed with which reciprocal monophyly is achieved (Ballard and Whitlock 2004; Palumbi et al. 2001). Second, statistical tests of topological congruence (specifically, Shimodaira–Hasegawa tests [Shimodaira and Hasegawa 1999] implemented in PAUP* [Swofford 2000]; see Linnen and Farrell 2007 for details) revealed substantially more conflict between nuclear and mitochondrial data partitions than between individual nuclear partitions—this pattern is also unexpected under incomplete lineage sorting alone. Given these two phylogenetic patterns and a large body of literature documenting the propensity for mitochondria to cross species boundaries (reviewed in Avise 2004; Chan and Levin 2005), I hypothesized that biased mitochondrial introgression, not incomplete lineage sorting, was the primary cause of the discordance I observed.

The mitochondrial introgression hypothesis predicts that mitochondrial gene flow has been consistently higher than nuclear gene flow throughout *Neodiprion*'s evolutionary history. Because this prediction deals with species that have exchanged genes following divergence, an appropriate framework for measuring gene flow is provided by the isolation with migration model (Hey and Nielsen 2004; Nielsen and Wakeley 2001). In this model, an ancestral taxon with an effective population size N_A splits into two descendant taxa (with effective sizes N_1 and N_2) at time t, after which populations 1 and 2 exchange genes at rates m_1 and m_2. Nielsen and Wakeley (2001) developed a likelihood/Bayesian framework that uses a Markov chain Monte Carlo (MCMC) approach to fit single-locus data sets to the IM model, and Hey and Nielsen (2004) extended this method to multiple loci in their program IM.

In order to test the mitochondrial introgression hypothesis, I allowed each of my four loci to have a separate pair of migration rates (option –j 7 in the 9/1/09 version of IM) and asked whether or not mitochondrial gene flow was higher. Because the program IM can only accommodate pairs of taxa, I estimated parameters in the four-locus model for every possible pairwise species comparison (120 total comparisons for the 16 species for which I had sampled multiple individuals) and found that, across all comparisons, mitochondrial gene flow was consistently higher than nuclear gene flow (Fig. 9.2; see Linnen and Farrell 2007 for additional analysis details). This finding suggests that mitochondrial genes are unreliable for recovering phylogenetic history in *Neodiprion* if this additional source of discord is not taken into account. In contrast, gene flow rates at nuclear loci appear low enough that most within-species variation is the result of ancestral variation and novel mutation, not introgression (i.e., $2Nm < 1$; Wright 1931); thus, even with some gene flow, a bifurcating tree is still a reasonable description of the *Neodiprion* species tree, and nuclear loci should be useful for reconstructing this history.

This work demonstrates that IM is a useful tool for determining whether a given locus has experienced exceptionally high (or low) gene flow rates compared to other loci, thus informing the choice of loci for species tree estimation. However, it is unfortunate

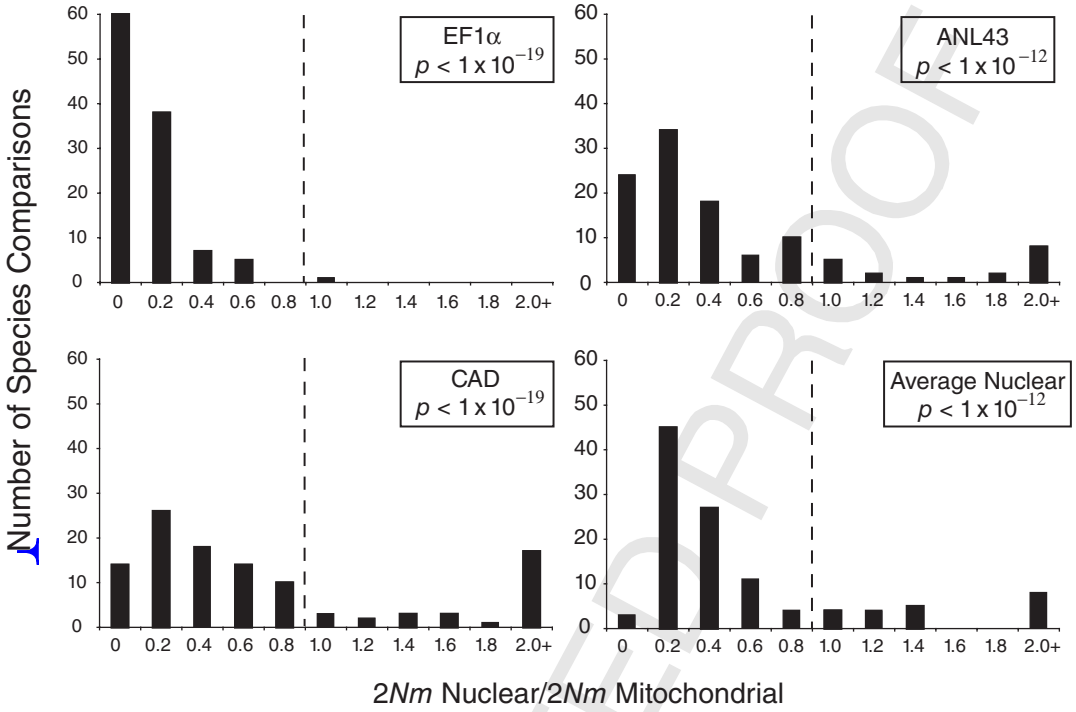

Figure 9.2 Mitochondrial gene flow is consistently higher than nuclear gene flow. Each panel is a frequency distribution of the ratio of nuclear (EF1α, ANL43, CAD, average nuclear) to mitochondrial gene flow (2*Nm*) for *Neodiprion* species pairs. Dashed lines indicate equal mitochondrial and nuclear gene flow. If gene flow rates had been equal, estimates would have been centered on these lines; however, distributions for all nuclear loci were shifted to the left. *p*-Values were calculated using Wilcoxon matched-pairs signed-ranks tests (see Linnen and Farrell 2007 for additional details).

that IM can only accommodate pairs of taxa for several reasons. First, preparing and analyzing data files for a large number of pairwise species comparisons is not only time consuming, but interpreting the results with so many comparisons can be difficult. Even with access to a computing cluster that can analyze many files simultaneously, each comparison must be monitored to ensure that parameter priors are appropriate and that the analysis has converged on the desired posterior distribution (see IM manual and Linnen and Farrell 2007 for additional details; also see discussion of similar issues in ~~the Bayesian Estimation of Species Trees [BEST] section below~~). Also, analyzing all pairwise species comparisons is a clear violation of the IM assumption that the entities examined are sister taxa that have not exchanged genes with a third taxon. Thus, individual gene flow estimates must be interpreted with caution because gene flow may be erroneously inferred between nonhybridizing species due to gene flow with a third species and/or gene flow between ancestral taxa (~~Hey and Nielsen 2006;~~ Won and Hey 2005). The optimal solution to this problem would be to integrate gene flow and species tree estimation into a single analysis. In the absence of such a method, my approach was to interpret IM gene flow estimates in light of a species tree estimated using different methods (Linnen and Farrell 2007; see also ~~Comparison of Gene Trees to Species Trees section~~). Despite these drawbacks, I found the information generated by IM analyses to be indispensable to understanding both difficulties in phylogenetic estimation and details of evolutionary history in *Neodiprion*

(see ~~Comparison of Species Tree Estimates and Comparison of Gene Trees to Species Trees sections~~).

9.5 APPROACHES FOR SPECIES TREE ESTIMATION

IM analyses clearly demonstrated that mitochondrial genes introgress readily and are therefore unreliable for recovering the *Neodiprion* species tree if this additional source of discord is not taken into account, but I was still left with the problem of how to best estimate a species tree from nuclear loci, which after all still showed discord (Fig. 9.3). Incomplete lineage sorting almost certainly explains some of this discord because biogeographic and phylogenetic evidence suggests that the *lecontei* clade originated rapidly and recently (Linnen and Farrell 2008b). There are several species tree methods that account for incomplete lineage sorting, and these methods have been shown to perform well, even when this source of discord is widespread (Carstens and Knowles 2007; Edwards et al. 2007; Kubatko et al. 2009; Maddison and Knowles 2006). However, these methods assume that there has been no introgressive hybridization between species, and IM analyses show that this assumption is violated in *Neodiprion*, even when only nuclear loci are considered (Fig. 9.2). There is some evidence that coalescent-based methods of species tree inference

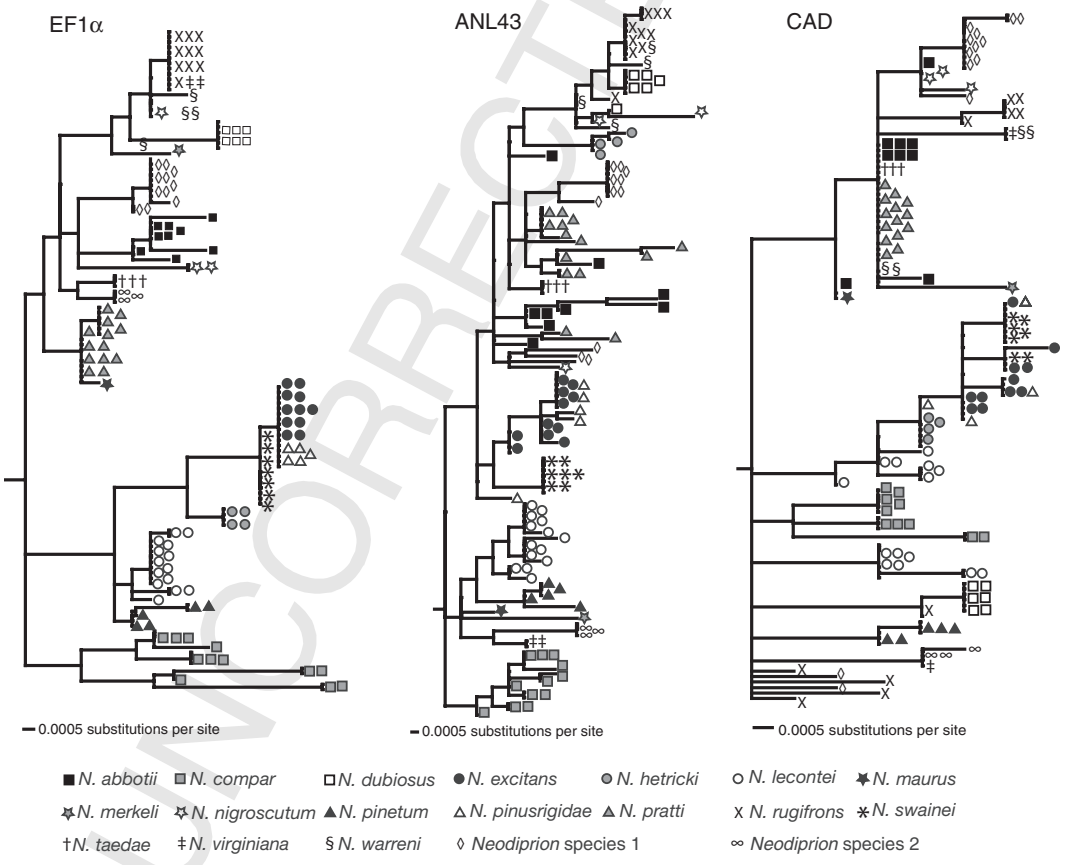

Figure 9.3 Maximum likelihood phylograms for three nuclear genes.

(e.g., minimize deep coalescences [MDC] and BEST; see below) may be robust to low levels of gene flow (Eckert and Carstens 2008; Knowles and Carstens 2007; Liu et al. 2008), but in the absence of a method that explicitly models both incomplete lineage sorting and gene flow, I decided to utilize multiple species tree methods, which are described below, in order to discover which relationships were robust to method choice (details on all analyses are also available in Linnen and Farrell 2008a).

9.5.1 Concatenation with Monophyly Constraints (CMC)

Because species tree methods are a relatively recent development, the most common way of dealing with multilocus data in phylogenetic analysis is still concatenation (i.e., analyze data as if it were a single supergene). This method is popular because it is expected to make more efficient use of multilocus data compared with consensus tree approaches (in which a gene tree is estimated for each locus and a consensus tree is computed across all gene trees) based on the argument that phylogenetic signal across loci will swamp out nonphylogenetic "noise" that stems from incomplete lineage sorting and introgression (Baker and DeSalle 1997; de Queiroz and Gatesy 2007; de Queiroz et al. 1995; Kluge 1989; Lerat et al. 2003; Wiens 1998).

A critical assumption of the concatenation approach is that the predominant 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 nonmatching trees have been dubbed "anomalous gene trees" [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; but see Degnan, Chapter 4). This suggestion is supported by simulation studies that have demonstrated that increasing the number of individuals sampled per species increases the probability of obtaining an estimated topology that is concordant with the underlying species tree (Maddison and Knowles 2006; Rosenberg 2002; Takahata 1989).

When multiple individuals are sampled, however, species trees estimated using the concatenation approach become difficult to interpret because recently diverged species may be recovered as nonmonophyletic (Carstens and Knowles 2007; Funk and Omland 2003; Hudson and Coyne 2002). One way to circumvent this problem is to incorporate topological constraints into phylogenetic analysis, thereby forcing species monophyly. Conceptually, monophyly constraints fit most naturally into a Bayesian framework for phylogenetic analysis because preexisting taxonomic information (i.e., species identity) can be incorporated into the analysis as topological priors. This can be carried out using the "constraint" and "prset" commands in MrBayes (Ronquist and Huelsenbeck 2003). Alternatively, one could produce a species tree from a concatenated gene tree after phylogenetic analysis using the "collapsing" approach described by Rosenberg (2002). Specifically, species trees are constructed by proceeding backward in time along the branches of a gene tree and by grouping species (or clades) in the order of their interspecific (or interclade) coalescences. This approach is thus analogous to the "shallowest divergences" (SD) method described below but differs in that concatenated gene trees, not pairwise genetic distances, are used. In my experience, monophyly constraints and collapsing produce identical (or nearly so) species tree topologies (although it is not clear

whether such trees accurately represent the actual history of divergence). One feature of the former approach, unlike the latter method, is that branch lengths are retained, which might be needed if one wishes to use the species tree as input for other analyses (e.g., a dating analysis in the program r8s; Sanderson 2003).

To summarize, my motivations for using the CMC approach to estimate a species tree in *Neodiprion* were the following: (1) shared signal in the data might swamp out nonphylogenetic noise stemming from gene flow and incomplete lineage sorting; (2) the probability that this shared signal results from an AGT may be lessened by the inclusion of multiple individuals per species; and (3) the monophyly constraints produce a result that can be interpreted as a species tree (and therefore can be compared to estimates obtained with other species tree methods). Ultimately, however, simulation studies are needed to determine the accuracy of this method under different discordance-producing scenarios and different sampling schemes. Also, in contrast to the species tree methods described below, the CMC approach imposes a single genealogical history (which we know does not apply in this case) and does not consider the underlying causes of gene tree discordance, and therefore fails to take full advantage of the information contained in a multilocus data set.

9.5.2 MDC

In contrast to the CMC approach, the MDC method, which is implemented in the program Mesquite (Maddison and Maddison 2006), does consider each locus separately. Specifically, this method takes gene trees as input and seeks the species tree that requires the fewest incomplete lineage sorting (deep coalescence) events to explain, and therefore provides the most parsimonious explanation for, the observed gene trees (Maddison 1997). Both simulation and empirical studies suggest that the MDC method performs well, even when incomplete lineage sorting is prevalent (Maddison and Knowles 2006) and the assumption that there has been no gene flow is violated (Eckert and Carstens 2008; Knowles and Carstens 2007). Thus, I was hopeful that this method might perform well in *Neodiprion*.

One major drawback of the MDC approach, however, is that it does not take error in gene tree estimation into consideration. One way to investigate the impact of this is to repeat the analysis with different gene tree estimates (e.g., Bayesian vs. maximum likelihood [ML] trees, strict vs. majority rule consensus). Unfortunately, low ML bootstrap values indicate that there is considerable uncertainty in *Neodiprion* gene tree estimates (Linnen and Farrell 2008a), and it has been my experience that the gene trees one uses as input can have a large impact on MDC results. Also, when polytomies are present in the gene trees, one must decide how to treat them in the MDC analysis. Mesquite has two options: automatically resolve polytomies to minimize the incompleteness of lineage sorting or do not auto-resolve. I have found that this option can have a large impact on the species tree estimate (Fig. 9.4). Generally speaking, the auto-resolve option produced species tree estimates that were more in line with my a priori expectations, provided that there were not too many polytomies in the gene trees (which could produce some very strange results). A second problem with the MDC approach is that the species trees it produces lack branch lengths, which limits the utility of these estimates in downstream analyses.

9.5.3 SD

SD is a clustering method based on Takahata's (1989) demonstration that, for a given gene tree, the order of interspecific coalescences has a high probability of matching the underly-

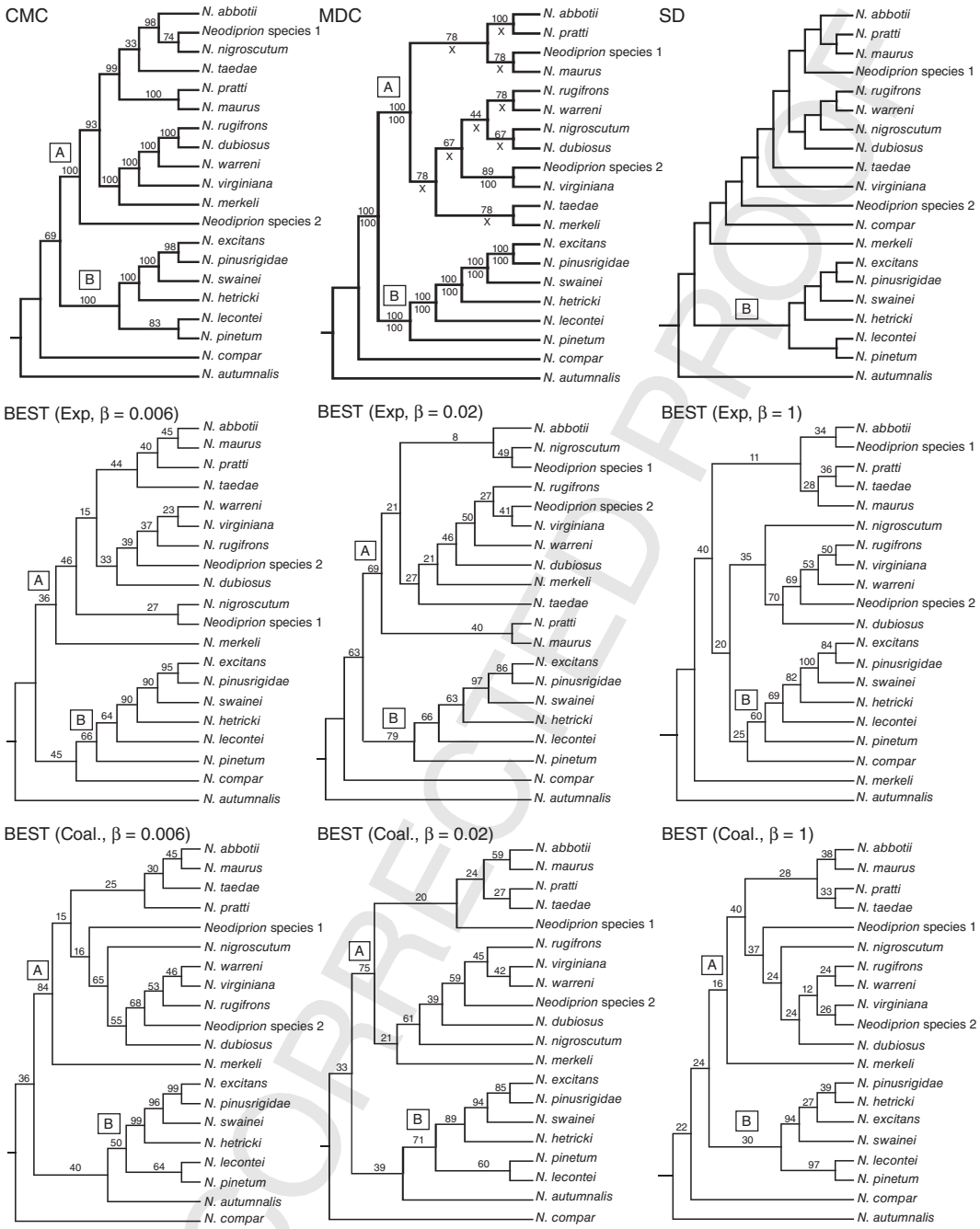

Figure 9.4 Species tree estimates obtained using different methods. Method names and details are abbreviated as described in the text and in Table 9.1. For CMC and BEST trees, Bayesian posterior probabilities are given above each node. For the MDC tree, numbers above and below nodes indicate the percentage of MDC trees that contained that clade for the "auto-resolve" (*top*) and "no auto-resolve" (*bottom*) options, and "X" indicates a conflicting relationship. Clade labels "A" and "B" denote clades that were recovered in most analyses and are discussed further in the text.

TABLE 9.1. Comparison of Topologies Obtained Using Different Species Tree Methods and Two Random Topologies (Default Symmetrical and Default Ladder, Both Generated in Mesquite)

	CMC	MDC AR	MDC NAR	SD	BEST E, 0.006	BEST E, 0.02
CMC	—	7	6	7	7	11
MDC AR	0.75	—	8	8	7	8
MDC NAR	0.63	0.80	—	5	9	8
SD	0.75	0.75	0.60	—	7	7
BEST E, 0.006	0.72	0.82	0.80	0.78	—	9
BEST E, 0.02	0.66	0.76	0.75	0.49	0.76	—
BEST E, 1	0.64	0.79	0.73	0.61	0.84	0.72
BEST C, 0.006	0.66	0.78	0.75	0.72	0.91	0.79
BEST C, 0.02	0.65	0.79	0.73	0.67	0.87	0.79
BEST C, 1	0.60	0.77	0.71	0.66	0.85	0.81
Default symmetrical	0.02	0.06	0.13	–0.04	0.00	0.10
Default ladder	0.07	0.05	0.01	0.06	–0.08	–0.09

Method names are abbreviated as described in the text. Number of shared clades (above diagonal) and pairwise patristic correlation coefficients (below diagonal) were calculated in Mesquite. For MDC, "AR" and "NAR" refer to the auto-resolve and no auto-resolve options. Priors for BEST analyses are given as follows: "E" and "C" refer to the exponential and coalescent branch-length priors, and numbers refer to the value of β in the inverse gamma prior for θ (all with $\alpha = 3$). β-Values were chosen to correspond to a biologically realistic range of θ values for *Neodiprion* (based on sequence-based estimates of θ).

ing species tree (Maddison and Knowles 2006). As it is implemented in Mesquite, however, the SD method does not use actual gene trees; instead, it assumes that the most similar pairs of sequences (i.e., those with the smallest genetic distance) represent the shallowest (most recent) coalescent events (Takahata and Nei 1985). When multiple loci are sampled, the distance between two individuals or clusters of individuals is averaged across all loci. The input for this method is aligned sequence data.

When the only cause of discordance is incomplete lineage sorting, simulation studies suggest that the SD method performs comparably to the MDC approach (Maddison and Knowles 2006). However, when gene flow is also present, the SD method may perform much worse than the MDC method because it considers only the first interspecific coalescences (which might result from recent introgression events rather than shared ancestry), whereas MDC considers all coalescent events (Knowles and Carstens 2007; Maddison and Knowles 2006). My experience with this method is consistent with these expectations—while I do not know the true *Neodiprion* species tree, the trees produced by the SD method deviated the most from my expectations based on morphology and from trees produced by the other methods (Fig. 9.4, Table 9.1). An additional disadvantage of this approach is that, like the MDC method, SD does not produce species trees with branch lengths.

9.5.4 BEST

BEST utilizes an MCMC algorithm to estimate the joint posterior distribution of gene trees and the containing species tree under a hierarchical Bayesian model, which is described in more detail in Chapter 2 (see also Edwards et al. 2007; Liu and Pearl 2007; Liu et al. 2008). This approach has several advantages over the CMC, MDC, and SD approaches. First, unlike MDC and SD, BEST accounts for uncertainty in gene tree estimation. Second, analyses of real and simulated data suggest that the more complex model implemented in BEST explains multilocus data better than the traditional Bayesian analy-

BEST E, 1	BEST C, 0.006	BEST C, 0.02	BEST C, 1	Default Symmetrical	Default Ladder
5	6	6	6	0	0
7	6	6	8	1	0
8	6	6	6	2	0
5	6	5	6	0	0
10	12	11	9	0	0
7	8	8	9	1	0
—	8	10	7	0	0
0.83	—	15	11	0	0
0.89	0.96	—	10	0	0
0.80	0.92	0.88	—	1	0
0.03	0.04	0.06	0.02	—	4
0.06	–0.06	0.00	–0.08	0.55	—

sis of concatenated data (models were compared using Bayes factors; Belfiore et al. 2008; Liu and Pearl 2007). In *Neodiprion*, modeling population-level processes of gene lineage coalescence in addition to nucleotide substitutions (i.e., the hierarchical model implemented in BEST) resulted in higher likelihood scores compared with the traditional Bayesian analysis of concatenated data (with and without monophyly constraints), and Bayes factors calculated from the harmonic means of the likelihoods of the data under the different models decisively favored the BEST model ($2\ln(B_{10}) > 100$; Kass and Raftery 1995; Nylander et al. 2004). Third, BEST estimates species tree branch lengths (divergence times) *and* widths (population sizes). While CMC also produces branch-length estimates, these are gene tree branch lengths, which are distinct from species tree branch lengths due to the coalescent process (Edwards and Beerli 2000).

BEST has the potential to tell us a great deal about the evolutionary history of a group of species, but, like all of the species tree methods discussed in this chapter, it assumes that there has been no hybridization. While simulations suggest that other coalescent-based methods for species tree inference (ESP-COAL and MDC) perform well despite historical gene flow (Eckert and Carstens 2008), the impact of gene flow on BEST has not been explored with simulations. At the very least, introgression is expected to have a large impact on divergence time estimates because the BEST model assumes that all interspecific gene tree coalescences predate speciation events—thus, a recently introgressed allele will be interpreted as evidence for an even more recent speciation event. If introgression has occurred between nonsister species, this may also lead to topological inaccuracies in BEST.

Model violations aside, the two most important practical considerations when using BEST, or any other Bayesian MCMC method (e.g., IM or traditional Bayesian phylogenetic estimation; see Chapter 1 for general discussion of Bayesian methods and MCMC), are ensuring that (1) results are not overly sensitive to choice of priors and (2) the analysis has converged on the desired posterior probability (or stationary) distribution (Huelsenbeck

et al. 2002). In my experience, neither of these are trivial issues in BEST. First, in contrast to the suggestion that topology may be robust to choice of priors (Liu and Pearl 2007), my species tree estimates were heavily dependent on choice of θ and branch-length priors (Fig. 9.4). There are several possible explanations for this discrepancy. There may be insufficient information in my data to estimate the parameters of interest and priors therefore have a large impact on the posterior distribution. Alternatively, this history of divergence may be particularly sensitive to the priors; it is worth noting that the sensitivity of BEST to priors has not been examined across a broad range of divergence histories. Distinguishing between these hypotheses will require additional data. Second, several diagnostics can be used to assess convergence of the BEST runs. Perhaps the simplest method is to examine log likelihood plots to ensure that these reach a steady value (samples from the Markov chain prior to this point are discarded as "burn-in"). Because this likelihood plateau could represent a local optimum, it is also necessary to demonstrate that similar results are obtained across independent runs. When the topology is the parameter of interest, similarity across independent runs can be assessed using the average standard deviation of the split frequencies—runs are considered to have converged if this value is below 0.010 (Ronquist et al. 2005). By this criterion, none of my *Neodiprion* BEST analyses (10 independent runs for each of six different prior combinations) converged (range: 0.088–0.257). Insufficient run times (all runs were 50 million generations) and inefficient tree searching could explain this lack of convergence, but I did not see any improvement upon altering run conditions (e.g., increasing the number of Markov chains and altering the "propTemp" and "poissonmean" search parameters). Instead, I suspect that my convergence problems resulted from a combination of complex posterior distribution shapes (and therefore a tendency to trap Markov chains in local optima) and insufficient data for the highly parameterized BEST model.

Even in the absence of convergence (according to the 0.010 criterion), I used the "sumt" command to summarize results across all runs (10 or more) for a given set of priors to see which, if any, clades were consistently recovered. Encouragingly, I found that there were many areas of agreement across independent runs. I also found that, while prior choice certainly impacted species tree topologies, BEST trees were more similar to one another, on average, than to trees obtained using the other methods (Table 9.1, Fig. 9.4). Nevertheless, because I cannot show that runs converged on the posterior distribution, nor can I demonstrate that results are robust to prior choice, my BEST results should be interpreted with caution and reexamined with additional data.

9.6 COMPARISON OF SPECIES TREE ESTIMATES

Each of the four methods I used to estimate a *Neodiprion* species tree has its advantages and disadvantages; each assumes that there has been no interspecific gene flow; and each method produced a different species tree topology (Fig. 9.4). As discussed above, these differences could stem from any number of reasons, including how the data are treated (i.e., concatenated or not; input is gene trees or aligned sequences), whether and how incomplete lineage sorting is taken into account, and whether error in gene tree estimation is considered. In the absence of a simulation study, however, I did not know which, if any, of these methods was most likely to produce an accurate species tree estimate under my sampling scheme and *Neodiprion*'s divergence history. I was therefore left with the task of making sense out of these different estimates.

First, I asked whether there were any similarities across the different methods. To do so, I looked at two metrics for topological similarity (patristic distance correlation and

TABLE 9.2. Comparison of Six Morphologically Based Hypotheses (from Ross 1955) to Results Obtained across Species Tree Methods

Method	*Virginianus* Complex	*Pratti* Complex	*Pinusrigidae* Complex	*Lecontei* Complex	*Abbotii* Complex	Non-*Abbotii* Clade
CMC	100	X	100.00	83.20	X	X
MDC (AR)	X	X	100.00	X	X	X
MDC (NAR)	X	X	100.00	100	X	X
SD	X	X	100.00	100	X	X
BEST (E, 0.006)	11.27	X	90.03	12.53	X	X
BEST (E, 0.02)	22.06	X	63.30	11.99	X	X
BEST (E, 1)	31.62	1.32	82.31	19.04	X	X
BEST (C, 0.006)	1.35	X	98.57	62.56	X	X
BEST (C, 2)	5.36	X	88.83	59.94	X	X
BEST (C, 1)	1.62	X	94.03	96.77	X	X

Method abbreviations are as described in the text and in Table 9.1. Numbers indicate the percentage of trees from a given method 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 that were statistically rejected because they were present in fewer than 5% (0.8% after Bonferonni correction for $n = 6$ tests) of all post-burn-in trees (see Buschbom and Barker 2006; Linnen and Farrell 2008a; Miller et al. 2002).

number of shared clades, both of which can be calculated in Mesquite) and found that species trees estimated by the different methods were much more similar to one another than to randomly generated trees (Table 9.1). In addition, I asked whether six clades that were proposed by Ross (1955) based on morphological characters were recovered by each of the methods. While there were some discrepancies between the methods (e.g., *virginianus* complex, Table 9.2), for the most part, methods agreed on which morphological hypotheses to reject (*pratti* complex, *abbotii* complex, non-*abbotii* clade) and which to support (*pinusrigidae* complex, *lecontei* complex). Moreover, several additional groupings were recovered by all (or nearly all) methods (e.g., clades "A" and "B," Fig. 9.4). These findings illustrate two points: (1) there is at least some agreement among the different methods, and (2) even when each method recovers a different tree, similarities can be used to inform our understanding of *Neodiprion* relationships.

Having found similarities across the methods, I next tried to determine where they disagreed most and why. A visual examination of the trees in Figure 9.4 illustrates that while there were pronounced differences between the methods regarding relationships within clade A, relationships within clade B were much more robust to method choice. A difference in the amount of phylogenetically informative variation did not explain this observation because average interspecific genetic distances did not differ significantly between the two clades ($p = 0.50$). While sample sizes (number of individuals) were, on average, larger for clade B (7.7 vs. 5.6 individuals per species in clade A), this also does not explain differing sensitivities to method choice because the same (or very similar) relationships were recovered for clade B when only a single individual was sampled per species (Linnen and Farrell 2008a). 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 tree, and therefore make phylogenetic inference more difficult, in the A clade. Based on estimates for these parameters from IM

analyses, clades A and B differ significantly in ancestral population size ($p = 0.042$) and nuclear gene flow rates ($p = 0.027$), but not divergence times ($p = 0.099$), indicating that the different species tree estimates reflect different sensitivities to several aspects of the divergence histories in clade A.

This raises an interesting biological question: why do the divergence histories differ between clades A and B? Because species in clade A use fewer host species, on average, than those in clade B (three hosts vs. six hosts), one possible explanation is that host use has been sufficiently specialized to permit divergence-with-gene-flow speciation in clade A but not in clade B. While additional data are needed to test this hypothesis, these findings illustrate the fact that species tree estimation may be inherently more difficult in some groups than in others and highlights the need for simulation studies that inform strategies for sampling and analysis under different divergence scenarios (e.g., Eckert and Carstens 2008; Maddison and Knowles 2006).

9.7 COMPARISON OF GENE TREES TO SPECIES TREES

In treating the *Neodiprion* species tree and the mitochondrial gene tree as distinct histories, I undoubtedly obtained a more accurate picture of *Neodiprion* relationships than I would have had I simply concatenated nuclear and mitochondrial loci. Because I was able to compare the two histories directly, this approach also provided insight into the biological processes underlying hybridization and mitochondrial introgression in the genus. As Figure 9.5 illustrates, there appears to have been a major episode of mitochondrial introgression that involved five *Neodiprion* species (~~*Neodiprion abbotii*, *Neodiprion dubiosus*, *Neodiprion nigroscutum*, *Neodiprion rugifrons*, and *Neodiprion swainei*~~). These species are morphologically and behaviorally distinct and occur in multiple clades throughout the *Neodiprion* species tree, but all five are monophagous (or nearly so) on jack pine. Furthermore, two other proposed mitochondrial introgression events (involving ~~*Neodiprion pinetum*/*Neodiprion lecontei* and *Neodiprion prattii*/*Neodiprion taedae*~~; see Fig. 9.5 and Linnen and Farrell 2007) also involved species that at least sometimes share hosts. More generally, species that shared at least one host had higher estimated mitochondrial gene flow rates on average than pairs that shared no hosts. These results suggest that host sharing facilitates hybridization and mitochondrial introgression, which implies that divergent host use is an important barrier to gene flow in *Neodiprion*. These conclusions are robust to ~~the species tree method~~—for example, in none of the species trees in Figure 9.4 do the five jack pine feeding species form a monophyletic group.

9.8 CONCLUSIONS AND FUTURE DIRECTIONS

In this chapter, I have outlined one strategy for dealing with complex divergence histories, such as the one that characterizes the genus *Neodiprion*, using currently available methods. To summarize, this strategy consisted of (1) using the program IM to distinguish between lineage sorting and introgression as underlying causes of gene tree discordance and to identify genes with unusual amounts of gene flow, (2) utilizing multiple species tree methods that differ in their underlying assumptions and handling of the data to identify relationships that are robust (or sensitive) to method choice, and (3) taking this information into account when making inferences about *Neodiprion* evolution from species tree estimates. Compared to the traditional phylogenetic paradigm in which data are simply concatenated and analyzed, this strategy has resulted in a much richer understanding

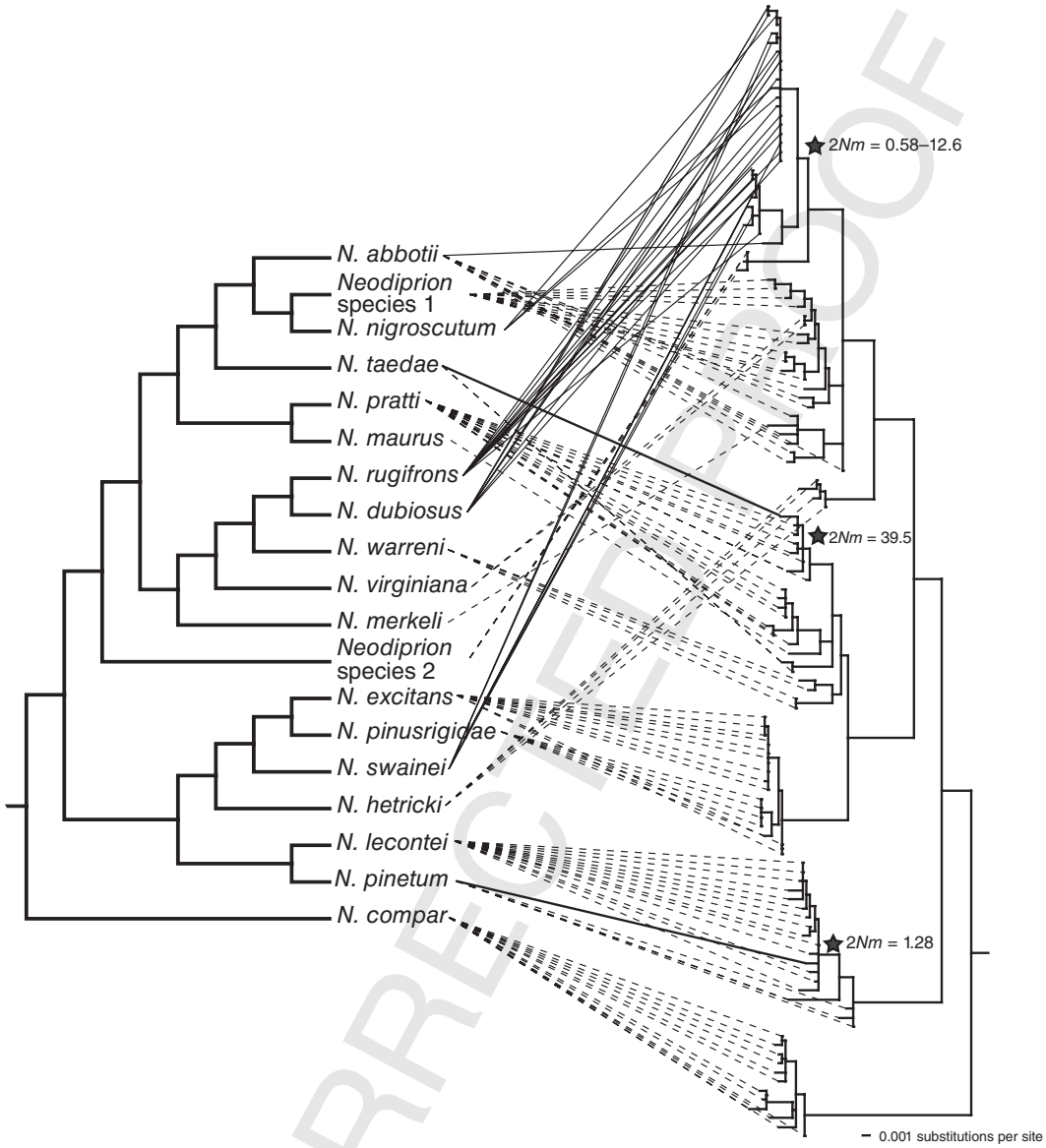

Figure 9.5 *Neodiprion* species tree topology and mitochondrial gene tree. The species tree was estimated using nuclear genes and the CMC method, but conclusions are robust to method choice (see text). The mitochondrial gene tree was estimated in MrBayes (see Linnen and Farrell 2007). Solid lines and stars denote recent mitochondrial introgression events that are discussed further in the text; introgression was inferred from a combination of polyphyly in the gene tree and high mitochondrial gene flow estimates (2*Nm* estimates are from IM analyses).

of *Neodiprion* evolution. Still, these analyses are far from ideal and highlight two main areas for future work.

First, it is clear that additional data are needed to obtain an accurate estimate of *Neodiprion* relationships. Given that I am missing two species and have only sampled a small fraction of the total geographical and ecological variation for some species, some of my future sampling effort should certainly go into increasing the number of individuals

(particularly in the A clade). However, I now think that I will see the most dramatic improvement (in terms of agreement across methods and, presumably, accuracy) if I focus most of my future effort on generating data from many more loci. While simulations suggest that the number of loci I sampled may be sufficient to obtain an accurate species tree estimate when gene tree–species tree discord is low (e.g., as may be the case in clade B), they also indicate that many more loci may be needed when incomplete lineage sorting and gene flow are prevalent (Eckert and Carstens 2008; Edwards et al. 2007; Maddison and Knowles 2006).

Second, groups like *Neodiprion* demonstrate the need for species tree methods that model postdivergence gene flow (e.g., Hey and Nielsen 2004, 2007; Nielsen and Wakeley 2001) in addition to stochastic lineage sorting (e.g., Carstens and Knowles 2007; Edwards et al. 2007; Liu and Pearl 2006). Moreover, a hybrid origin has been proposed for at least one *Neodiprion* species (*N. merkeli*; Ross 1961). While this hypothesis has not yet been tested (more *N. merkeli* individuals and more loci are needed), an increasing number of examples of hybrid speciation in a wide range of plant and animal lineages (reviewed in Arnold 2006; Mallet 2007) demonstrate the need for species tree methods that can also accommodate histories that are not strictly bifurcating (e.g., Meng and Kubatko 2009; Chapter 6). Extending current species tree methods to include gene flow and hybrid species formation will no doubt demand a large amount of data; thus, we will also need simulation studies that tell us how to divide our sampling efforts between individuals and loci. Even in their current form, however, species tree methods can tell us a great deal about the evolutionary process, and as these methods continue to improve, so too will our understanding of how and why populations diverge.

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*. Gainesville, FL: Sandhill Crane Press.

Arnold, M. L. 2006. *Evolution through Genetic Exchange*. New York: Oxford University Press.

Avise, J. C. 2004. *Molecular Markers, Natural History, and Evolution*, 2nd ed. Sunderland, MA: Sinauer.

Baker, R. H. and R. DeSalle. 1997. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. *Systematic Biology* 46:654–673.

Ballard, J. W. O. and M. C. Whitlock. 2004. The incomplete natural history of mitochondria. *Molecular Ecology* 13:729–744.

Becker, G. C. 1965. A biological-taxonomic study of the *Neodiprion virginianus* complex in Wisconsin. PhD dissertation, University of Wisconsin–Madison.

Belfiore, N. M., L. Liang, and C. Moritz. 2008. Multilocus phylogenetics of a rapid radiation in the genus *Thomomys* (Rodentia: Geomyidae). *Systematic Biology* 57:294–310.

Benjamin, D. M. 1955. The biology and ecology of the red-headed pine sawfly. *USDA Technical Bulletin* 118:••–••.

Bjorkman, C. and S. Larsson. 1991. Pine sawfly defense and variation in host plant resin acids: a trade-off with growth. *Ecological Entomology* 16:283–289.

Buschbom, J. and D. Barker. 2006. Evolutionary history of vegetative reproduction in *Porpidia* s.l. (lichen-forming Ascomycota). *Systematic Biology* 55:471–484.

Bush, G. L. 1975a. Modes of animal speciation. *Annual Review of Ecology and Systematics* 6:339–364.

Bush, G. L. 1975b. Sympatric speciation in phytophagous parasitic insects. In P. W. Price, ed. *Evolutionary Strategies of Parasitic Insects and Mites*. New York: Plenum, pp. 187–207.

Carstens, B. C. and L. L. Knowles. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. *Systematic Biology* 56:400–411.

Chan, K. M. A. and S. A. Levin. 2005. Leaky prezygotic isolation and porous genomes: rapid introgression of maternally inherited *DNA*. *Evolution* 59:720–729.

Coppel, H. C. and D. M. Benjamin. 1965. Bionomics of Nearctic pine-feeding diprionids. *Annual Review of Entomology* 10:69–96.

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

de Queiroz, A. and J. Gatesy. 2007. The supermatrix approach to systematics. *Trends in Ecology & Evolution* 22:34–41.

de Queiroz, A., M. J. Donoghue, and J. Kim. 1995. Separate versus combined analysis of phylogenetic evidence. *Annual Review of Ecology and Systematics* 26:657–681.

Eckert, A. J. and B. C. Carstens. 2008. Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. *Molecular Phylogenetics and Evolution* 49:832–842.

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

Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences of the United States of America* 104:5936–5941.

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.

Ghent, A. W. and D. R. Wallace. 1958. Oviposition behavior of the Swaine jack-pine sawfly. *Forest Science* 4:264–272.

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.

10 ~~Hey, J. and R. Nielsen. 2006. IM Documentation.~~

Hey, J. and R. Nielsen. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America* 104:2785–2790.

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

Huelsenbeck, J. P., B. Larget, R. E. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematc Biology* 51:673–688.

Kass, R. E. and A. E. Raftery. 1995. Bayes factors. *Journal of the Americal Statistical Association* 90:773–795.

Kluge, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Systematic Zoology* 38:7–25.

Knerer, G. and C. E. Atwood. 1972. Evolutionary trends in subsocial sawflies belonging to *Neodiprion abietis* complex (Hymenoptera: Tenthredinoidea). *American Zoologist* 12:407–418.

Knerer, G. and C. E. Atwood. 1973. Diprionid sawflies: polymorphism and speciation. *Science* 179:1090–1099.

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. *Systematic Biology* 56:17–24.

Kubatko, L. S., B. C. Carstens, and L. L. Knowles. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25:971–973.

Lerat, E., V. Daubin, and N. A. Moran. 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the a-proteobacteria. *PLoS Biology* 1:101–109.

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. 2008a. Comparison of methods for species-tree inference in the sawfly genus *Neodiprion* (Hymenoptera: Diprionidae). *Systematic Biology* 57:876–890.

Linnen, C. R. and B. D. Farrell. 2008b. Phylogenetic analysis of nuclear and mitochondrial genes reveals evolutionary relationships and mitochondrial introgression in the *sertifer* species group of the genus *Neodiprion* (Hymenoptera: Diprionidae). *Molecular Phylogenetics and Evolution* 48:240–257.

Liu, L. and D. K. Pearl. 2006. Reconstructing posterior distributions of a species phylogney using estimated gene tree distributions. Biosciences Institute Technical Report #53, The Ohio State University, Columbus, OH.

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. *Systematic Biology* 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. *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. 2006. Mesquite: a modular system for evolutionary analysis. Version ~~1.12~~ 2 http://mesquiteproject.org (accessed ~~••, ••, ••~~ 2). 11

Mallet, J. 2007. Hybrid speciation. *Nature* 446:279–283.

Mayr, E. 1942. *Systematics and the Origin of Species*. New York: Columbia University Press.

Mayr, E. 1963. *Animal Species and Evolution*. Cambridge, MA: Belknap.

McCullough, D. G. and M. R. Wagner. 1993. Defusing host defenses: ovipositional adaptations of sawflies to pine resins. In M. Wagner and K. F. Raffa, eds. *Sawfly Life History Adaptations to Woody Plants*. San Diego, CA: Academic Press, pp. 157–172.

McMillin, J. D. and M. R. Wagner. 1993. Influence of stand characteristics and site quality on sawfly population dynamics. In M. R. Wagner and K. F. Raffa, eds. *Sawfly Life History Adaptations to Woody Plants*. San Diego, CA: Academic Press, pp. 333–361.

Meng, C. and L. S. Kubatko. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theoretical Population Biology* 75:35–45.

Miller, R. E., T. R. Buckley, and P. S. Manos. 2002. An examination of the monophyly of morning glory taxa using Bayesian phylogenetic inference. *Systematic Biology* 51:740–753.

J3

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. *Systematic Biology* 53:47–67.

Palumbi, S. R., F. Cipriano, and M. P. Hare. 2001. Predicting nuclear gene coalescence from mitochondrial data: the three-times rule. *Evolution* 55:859–868.

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*. ••: ••.

Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. *Theoretical Population Biology* 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. *Systematic Biology* 57:131–140.

[illegible], H. H. 1961. Two new species of *Neodiprion* from southeastern North America (Hymenoptera: Diprionidae). *Annals of the Entomological Society of America* 54:451–453.

Sanderson, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19: 301–302.

Shimodaira, H. and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* 16: 1114–1116.

Smith, D. R. 1979. Symphyta. In ~~K. V. K. Smith, et al., ed.~~, *Catalog of Hymenoptera in America North of Mexico*. Washington, DC: Smithsonian Institution Press, pp. 3–137.

Strong, D. R., J. H. Lawton, and R. Southwood. 1984. *Insects on Plants: Community Patterns and Mechanisms*. Cambridge, MA: Harvard University Press.

Swofford, D. L. 2000. *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sunderland, MA: Sinauer.

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

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

Tauber, C. A. and M. J. Tauber. 1981. Insect seasonal cycles: genetics and evolution. *Annual Review of Ecology and Systematics* 12:281–308.

Wakeley, J. 2000. The effects of subdivision on the genetic divergence of populations and species. *Evolution* 54:1092–1101.

Wiens, J. J. 1998. Combining data sets with different phylogenetic histories. *Systematic Biology* 47:568–581.

Wilson, L. F., R. C. Wilkinson, and R. C. Averill. 1992. *Redheaded Pine Sawfly: Its Ecology and Management*. Washington DC: USDA.

Won, Y. J. and J. Hey. 2005. Divergence population genetics of chimpanzees. *Molecular Biology and Evolution* 22:297–307.

Wright, S. J. 1931. Evolution in Mendelian populations. *Genetics* 16:97–159.

AUTHOR QUERY FORM

Dear Author

During the preparation of your manuscript for publication, the questions listed below have arisen. Please attend to these matters and return this form with your proof.

Many thanks for your assistance.

Query References	Query	Remarks
	AUTHOR: Note that figures may have been relabelled for readability, please check.	
1.	AUTHOR: Please confirm if the full name of the author is correct or if it should be changed to "Catherine R. Linnen."	
2.	AUTHOR: Please define IM.	
3.	AUTHOR: Please confirm if "the Bayesian Estimation of Species Trees [BEST] section below" is okay or if it should be changed to "Section 9.5.4."	
4.	AUTHOR: Please confirm if "Comparison of Gene Trees to Species Trees section" is okay or if it should be changed to "Section 9.7."	
5.	AUTHOR: Please confirm if "Comparison of Species Tree Estimates and Comparison of Gene Trees to Species Trees sections" is okay or if it should be changed to "Sections 9.6 and 9.7."	
6.	AUTHOR: Please define ESP-COAL.	
7.	AUTHOR: Ross (1955) has not been found in the reference list. Please provide full reference details or delete these citations from the text.	
8.	AUTHOR: Please note that the use of only the species name (without the genus name) has been retained throughout the text as it is assumed to be the author's preference. Please confirm if this is correct.	
9.	AUTHOR: Benjamin 1955: Please provide the page range.	
10.	AUTHOR: Hey and Nielsen 2006: Please provide more reference details.	
11.	AUTHOR: Maddison and Maddison 2006: Please provide the access date (month, day, and year).	

12.	AUTHOR: Ronquist et al. 2005: Please provide the city of publication and the publisher name.	
13.	AUTHOR: Smith 1979: Please confirm if the editor's surname is correct. Also, please provide all the other editors' names and surnames instead of using "et al."	
14.	AUTHOR: Figure 9.1: The questions "Are changes in the trait … Do other traits show …" have been made into a run-on list. Is this okay?