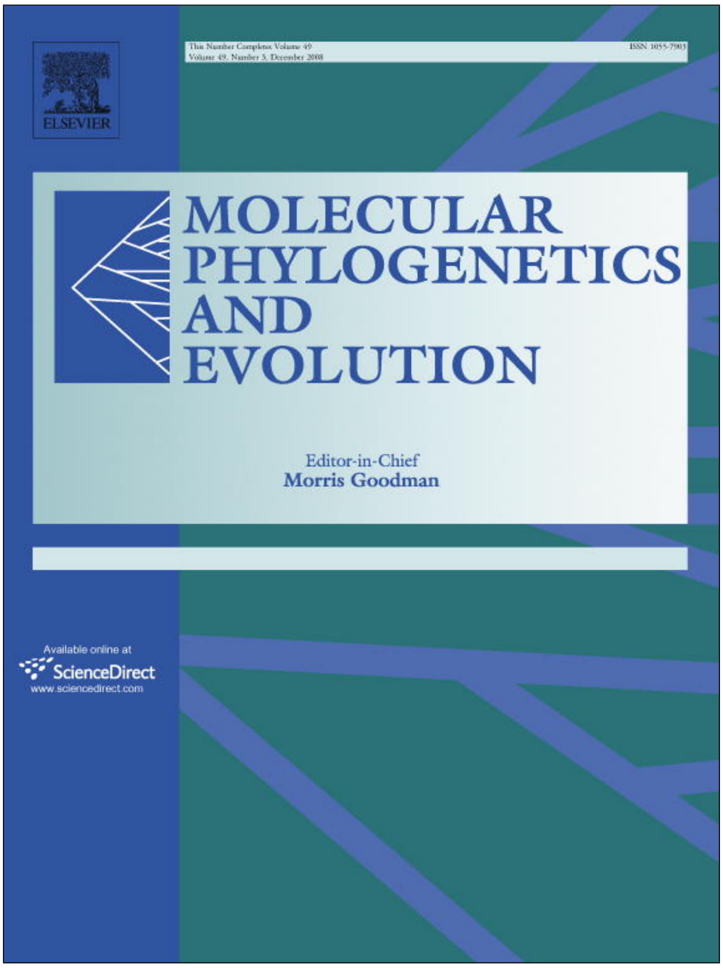

This Number Completes Volume 49
Volume 49, Number 3, December 2008

ISSN 1055-7903

ELSEVIER

MOLECULAR PHYLOGENETICS AND EVOLUTION

Editor-in-Chief
Morris Goodman

Available online at
ScienceDirect
www.sciencedirect.com

Molecular Phylogenetics and Evolution 49 (2008) 832–842

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

# Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow

Andrew J. Eckert [a], Bryan C. Carstens [b,*]

[a] *Section of Evolution and Ecology, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA*
[b] *Department of Biological Sciences, 202 Life Sciences Building, Louisiana State University, Baton Rouge, LA 70803, USA*

ARTICLE INFO

*Article history:*
Received 21 March 2008
Revised 8 September 2008
Accepted 12 September 2008
Available online 21 September 2008

*Keywords:*
COAL
Gene flow
Estimating species phylogenies
Incomplete lineage sorting
Minimizing deep coalescence
*Rhododendron*

## ABSTRACT

Incomplete lineage sorting has been documented across a diverse set of taxa ranging from song birds to conifers. Such patterns are expected theoretically for species characterized by certain life history characteristics (e.g. long generation times) and those influenced by certain historical demographic events (e.g. recent divergences). A number of methods to estimate the underlying species phylogeny from a set of gene trees have been proposed and shown to be effective when incomplete lineage sorting has occurred. The further effects of gene flow on those methods, however, remain to be investigated. Here, we focus on the performance of three methods of species tree inference, ESP-COAL, minimizing deep coalescence (MDC), and concatenation, when incomplete lineage sorting and gene flow jointly confound the relationship between gene and species trees. Performance was investigated using Monte Carlo coalescent simulations under four models (*n*-island, stepping stone, parapatric, and allopatric) and three magnitudes of gene flow ($N_em = 0.01, 0.10, 1.00$). Although results varied by the model and magnitude of gene flow, methods incorporating aspects of the coalescent process (ESP-COAL and MDC) performed well, with probabilities of identifying the correct species tree topology typically increasing to greater than 0.75 when five more loci are sampled. The only exceptions to that pattern included gene flow at moderate to high magnitudes under the *n*-island and stepping stone models. Concatenation performs poorly relative to the other methods. We extend these results to a discussion of the importance of species and population phylogenies to the fields of molecular systematics and phylogeography using an empirical example from *Rhododendron*.

## 1. Introduction

The fundamental goal of systematics is to understand the process of lineage divergence that leads to the formation of new species. Since Maddison (1997) there has been growing acceptance among systematists that gene genealogies are not always congruent with species phylogenies (e.g. the actual pattern of lineage splitting and descent from common ancestors). It is now widely recognized that processes such as gene duplication (Fitch, 1970), lateral transfer (Cummings, 1994) and incomplete lineage sorting (Tajima, 1983; Takahata and Nei, 1985; Hudson, 1992) can lead to incongruence between gene trees and species trees, and empirical examples of each process exist (cf. Syring et al., 2007 for an example of incomplete lineage sorting). This realization has prompted the development of approaches designed to estimate species phylogenies despite the process that presumably caused the incongruence. For example, gene tree parsimony (Slowinski and Page, 1999) was developed to account for gene duplication, while the minimization of deep coalescence (MDC; Maddison, 1997), COAL (Degnan and Salter, 2005), and BEST (Edwards et al., 2007; Liu and Pearl, 2007) were designed in part to estimate species phylogeny when the discord between the gene trees and species tree is a result of the incomplete sorting of ancestral polymorphisms.

At the initial stages of divergence, incomplete lineage sorting is ubiquitous and likely produces the majority of gene-species tree discord among closely related lineages. This is a direct outcome of population-level processes; consequently, the developers of methods have incorporated statistical models derived from the coalescent (Kingman, 1982; Hudson, 1990) into species-level phylogenetic analyses to account for these processes. However, for many empirical systems it is also these lineages that exchange migrants, particularly when they occur in sympatry. Since genetic polymorphism shared among lineages can result from either retained ancestral polymorphism or a gene copy introduced into the population via gene flow (Slatkin and Maddison, 1989), it is often difficult to determine which process produced the shared polymorphism. Fully statistical treatments of coalescence, gene flow,

[*] Corresponding author.
*E-mail addresses:* carstens@lsu.edu, ajeckert@ucdavis.edu (B.C. Carstens).

1055-7903/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2008.09.008

and divergence are currently available only for pairwise comparisons between two lineages (Nielsen and Wakeley, 2001; Hey and Nielsen, 2004, 2007; Hey, 2006).

It is an understatement to suggest that the biologist who wishes to estimate species phylogeny in a system where details such as (a) the number of lineages, (b) the relationship among lineages, and (c) the amount of gene flow are unclear is currently faced with a difficult task. Methods that estimate a species phylogeny using some approach derived from the coalescent must be robust to at least moderate levels of gene flow (e.g. levels that not be easily recognizable) to be of any use to the majority of empirical biologists, or the use of such methods may result in spurious conclusions about the actual pattern of lineage divergence. The data we present in this manuscript were collected out of a desire to explore how the phylogenetic signal contained in DNA sequence data is affected by gene flow in recently diverged lineages. Does gene flow destroy phylogenetic signal entirely, or are some methods able to accurately estimate species phylogeny when some of the shared polymorphisms result from gene flow? In order to explore this issue, we evaluate approaches based on the coalescent that use estimated gene trees as input in an attempt to isolate gene flow as the sole factor affecting phylogenetic accuracy.

## 2. Materials and methods

### 2.1. Statistical inference of species trees from gene trees

A renewed interest exists in the development and interpretation of statistical methods for the inference of species trees from gene trees (Maddison and Knowles, 2006). A myriad of innovative approaches have been developed (Slatkin and Maddison, 1989; Maddison, 1997; Page and Charleston, 1997; Slowinski and Page, 1999; Liu and Pearl, 2006; Edwards et al., 2007; Carstens and Knowles, 2007), as well as applied to empirical questions in phylogeography and systematics (Knowles and Carstens 2007; Brumfield et al., in press; Carling and Brumfield, 2008). Here, we focus on two methods for estimating species phylogenies at relatively low levels of lineage divergence. The first seeks to identify the species tree that maximizes the probability of a set of genealogies given the species tree (Maddison, 1997), as implemented in COAL (Degnan and Salter, 2005) and as applied by Carstens and Knowles (2007). The second method, described by Maddison (1997) and implemented in the Mesquite software package (Maddison and Maddison, 2004) minimizes the amount of deep coalescence to estimate the species phylogeny. Hereafter we refer to these approaches as ESP-COAL and MDC, respectively.

ESP-COAL is a maximum-likelihood approach to the inference of species trees from a set of gene trees. Maddison (1997) noted that the likelihood (L) of a species tree inferred from $n$ independent gene loci could be written as:

$$\mathrm{L(D|ST)} = \prod_{n=1}^{n}\left(\sum_{\mathrm{GT}}[\Pr(D|\mathrm{GT})\Pr(\mathrm{GT}|\mathrm{ST})]\right) \tag{1}$$

where D are the sequence data, ST is the species tree, and GT is the gene tree. Note that the summation is over all possible GT for each of the $n$ loci. The first expression of the inner product is the likelihood of the data given a gene tree, which can be computed by standard phylogenetic software. The second expression of the inner product represents the probability of a gene tree given a species tree. This quantity can only be calculated for some sample configurations using the mathematical theory for the neutral coalescent (Tajima, 1983, 1989; Hudson, 1983; Takahata, 1989; Rosenberg, 2002; Yang, 2002; Wall, 2003). Degnan and Salter (2005), however, devised a combinatoric approach for the calculation of this probability, with the limitation that it results in the probability of the gene tree topology, not considering branch lengths, conditional on a species tree topology with known branch lengths. Rigorous maximization of the likelihood function would require joint searches through the state space of all possible gene and species tree topologies and their branch lengths using some form of importance sampling (Felsenstein, 2004). In order to approximate the maximization of the likelihood function as defined above, we followed the method of Maddison (1997) and Carstens and Knowles (2007), which searches for the ST topology conferring the highest probability for the observed gene trees.

The second approach to estimating species phylogeny (MDC), also described by Maddison (1997), uses a heuristic search to identify the species phylogeny that minimizes the amount of deep coalescence (e.g. incomplete lineage sorting). This approach can be accurate in the absence of gene flow under certain assumptions concerning the species tree topology (Degnan and Rosenberg, 2006; Maddison and Knowles, 2006), but has not been explicitly explored given varying levels of gene flow. Like the ESP-COAL approach, it evaluates the pattern of coalescence without considering the branch lengths of the genealogies.

### 2.2. Parameters and models of gene flow

ESP-COAL is accurate when ancestral polymorphisms are segregating within species that otherwise conform to a bifurcating phylogenetic tree (Carstens and Knowles, 2007), particularly when the depth of the species tree is 3$N_e$ or greater. However, the signature of ancestral polymorphisms segregating within descendant lineages due strictly to genetic drift is complicated when gene flow, either recent or historical, has occurred among lineages (Slatkin and Maddison, 1989).

We devised four basic models of gene flow in order to elucidate the effects of this process on phylogenetic analyses: $n$-island, stepping stone, and two models of historical gene flow (Fig. 1). The models of historical gene flow were formulated to reflect scenarios of either allopatric or parapatric speciation. Historical gene flow occurred strictly between sister lineages and was modeled as a burst of gene flow directly after (parapatric) or 0.5$xN_e$ generations after speciation (allopatric), where $x$ is the length of time between successive speciation events (Fig. 1). The duration of these bursts was controlled by the parameter $d$, and we incorporated a relatively short period of divergence with gene flow (0.1$N_e$) as well as a longer period (0.5$N_e$). For each model, we assumed three different magnitudes of gene flow as measured by the effective number of migrants per generation ($N_em$ = 0.01, 0.10, or 1.00).

As shown previously, the power of the ESP-COAL and MDC depend upon the number of unlinked loci used in the analysis and the depth of the species tree (Maddison and Knowles, 2006; Carstens and Knowles, 2007). Therefore, we varied the number of sampled loci from two to ten and the depth of the species tree (2$N_e$ or 6$N_e$), as well as the effective population sizes ($N_e$ = 10,000 or 100,000). These parameter treatments were considered in a fully factorial design, yielding 72 different treatment combinations, for each of which we analyzed the accuracy of the ESP-COAL and MDC across samples of two to ten loci (Table S1; online Supplemental data).

### 2.3. Canonical species tree

We assumed a single, fully resolved, pectinate species tree with four taxa for our simulations [(((c:0.75(a:0.375, b:0.375):0.375): 0.25, d:1.00)]. The fourth taxon (taxon d) was designated as an outgroup. The ingroup taxa can then be characterized by three possible rooted tree topologies. In all cases, the relative branch lengths conform to a molecular clock and were defined as pictured in Fig. 1.

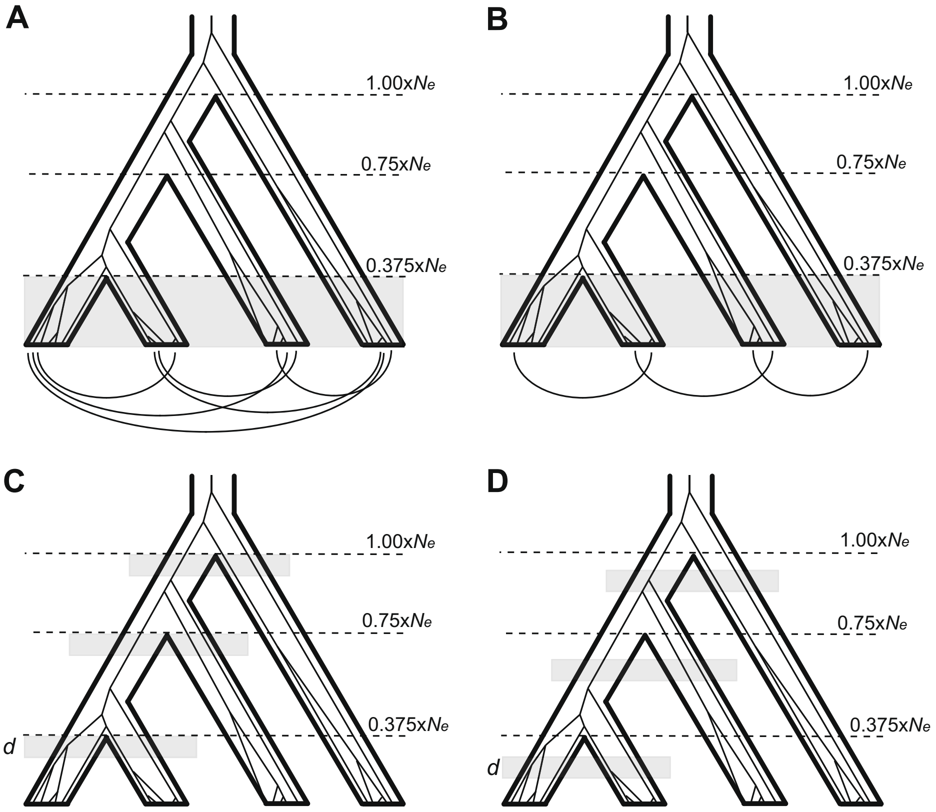

**Fig. 1.** Models of gene flow showing (A) *n*-island, (B) stepping stone, (C) historical allopatric, and (D) historical parapatric gene flow. Shown for each model are the species phylogeny (bold outlined), as well as an example genealogy contained within the species phylogeny. Branch lengths are standardized to a species tree depth of 1.00. The models are differentiated by when gene flow occurs; this is represented on the tree through the use of shaded rectangles. The *n*-island and stepping stone models are differentiated by which lineages exchange migrants. This is shown with the curved lines connecting the terminal lineages.

### 2.4. Monte Carlo coalescent simulations

Monte Carlo coalescent simulations were used to generate simulated genealogies under each treatment. For each of the parameter combinations, we simulated 500 genealogies consisting of five gene sequences sampled from each of four species in the canonical species tree. Using those simulated genealogies, we generated data sets consisting of 2–10 loci selected at random without replacement. The genealogies within these data sets were assumed to be estimated without error, thus negating the need to calculate the quantity Pr(*D*|GT) in ESP-COAL. We also assumed that all lineages were consistent and equal in their effective population size. Furthermore, at each speciation event, daughter lineages were assumed to instantaneously grow to the size of the ancestral population. For example, an ancestral population with an $N_e$ of 10,000 was assumed to speciate into two daughter lineages each with an $N_e$ of 10,000. While this is clearly a simplified model of the process of lineage divergence, our aim here is to explore the effects of gene flow, and only gene flow, on these methods. All coalescent simulations were carried out using Simcoal v. 2.0 (Excoffier et al., 2000).

### 2.5. Performance of ESP-COAL, MDC, and concatenation in estimating species phylogeny

We followed the approach of Carstens and Knowles (2007) to explore the power and accuracy of ESP-COAL. Briefly, this method involves five steps: (1) estimation of the Pr(GT|ST) using COAL for each of the 500 simulated gene trees, (2) sampling between two and ten loci, (3) calculating the sum of the probability of the gene trees given the species tree for each possible species tree, (4) performing approximate likelihood ratio tests (LRTs) to evaluate the significance of the correct species topology vs. the two incorrect topologies (Anisimova and Gascuel, 2006), and (5) counting the number of times the LRTs identified the correct species topology out of the replicated simulations. This process was repeated for each of the 72 parameter combinations.

We used a similar approach to explore the performance of the MDC method given varying types and amounts of gene flow. For each of the 72 treatments, we randomly selected between two and ten genealogies and used Mesquite v. 2.5 (Maddison and Maddison, 2004) to estimate the species phylogeny. The genealogies were treated as unrooted, and a heuristic search with nearest-neighbor interchange branch swapping and maxtrees set to 100 was used to explore species treespace. As above, we used the actual simulated genealogies in the searches.

Lastly, we analyzed the 72 treatments by concatenating data from between two and ten loci and estimating phylogeny using maximum-likelihood (ML). Sequence data were simulated using Simcoal under an HKY model of sequence evolution with *ts*/*tv* = 3.0 and simulations were conditioned on an expectation of $\theta = 20$, which resulted in between 40 and 60 segregating sites in each data simulated data set. For each concatenated data set, PAUP* (Swofford, 2002) was used to estimate the phylogeny using ML with a heuristic search, maxtrees = 10, and the HKY model of sequence evolution used to simulate the data. A tree filter was then used to determine what proportion of estimated phylogenies matched the species phylogeny that was used to simulate the data. Each treatment was replicated 100 times.

### 2.6. An empirical example from Rhododendron

The genus *Rhododendron* contains approximately 1000 species distributed throughout the Northern Hemisphere (Cox and Cox, 1997). Recent phylogenetic work has elucidated the placement of *Rhododendron* within the Ericaceae, defined generic boundaries, and revised the taxonomy within some of the largest subgeneric clades (Kron et al., 2002; Milne, 2004; Goetsch et al., 2005). At the species level, however, many relationships remain ambiguous, even when phylogenies are inferred from multilocus nuclear markers. This ambiguity is likely caused by a combination of historical gene flow and incomplete lineage sorting when the species under consideration are or were geographically proximal.

Here, we concentrate on inferring the species phylogeny of a group of four *Hymenanthes* rhododendrons (*Rhododendron macrophyllum* D. Don ex G. Don, *Rhododendron catawbiense* Michx., *Rhododendron caucasicum* Pall., and *Rhododendron brachycarpum* D. Don ex G. Don) distributed across eastern Asia, Europe, and North America using data from two RNA polymerase genes (*RPB*2 and *RPC*1) and two chloroplast genes (*trn*K and *trn*L-*trn*F). Data for the chloroplast genes were obtained from Milne (2004) and were concatenated due to the uniparental inheritance and lack of recombination for the chloroplast genome. The sample size for each species ranged from one to five gene copies per locus. All gene trees were rooted using *Rhododendron aureum* Georgi as an outgroup.

Estimation of the species tree topology for the four *Rhododendron* species was carried out using the three methods described above. For ESP-COAL, we selected the HKY model of DNA sequence evolution using DT-ModSel (Minin et al. 2003), and estimated gene trees with ML using PAUP∗. Parameters of the HKY model were estimated concomitantly with the gene tree topology and branch lengths. The likelihood score associated with the maximum-likelihood estimate (MLE) of the gene tree was equated to the Pr($D$|GT). The nonparametric bootstrap (Felsenstein, 1985) with 1000 replicates was used to assess the reliability of nodes within inferred genealogies as well as the concatenated tree topology. The topology of the MLE gene tree was evaluated on each of the 15 possible rooted species tree topologies using COAL. The species tree topology conferring the largest probability of the MLE gene tree was taken as the best estimate. Analyses using MDC and concatenation were conducted as described above for these data with gene and species trees being rooted with *R. aureum*.

While our analyses of simulated data used the actual genealogies, the analysis of empirical data utilized estimates of the actual *Rhododendron* genealogies for these loci, and includes an additional source of statistical error associated with the estimation of the gene tree. To explore the potential magnitude of this effect, we chose the simulation treatments likely to mirror the evolutionary history of the four *Rhododendron* species described above (parapatric model, $N_e$ = 10,000, $N_em$ = 0.10, $d$ = 0.10, and ST depth = 2$N_e$), and conducted the ESP-COAL and MDC analyses using estimated rather than the actual genealogies. Data sets were simulated under these conditions based on the conclusions of Milne (2004) who showed that the divergence time among the species considered here ranges from one to two million years ago, or approximately 2$N_e$ generations. Nucleotide data were simulated under the HKY model, maximum-likelihood searches in PAUP∗ were used to estimate the genealogies for 500 simulated data sets, and two to ten gene trees were picked at random and analyzed as described in Section 2.5. Thus, any decrease in the performance of ESP-COAL and MDC likely results from the contribution of the addition source of error in the estimated genealogies.

## 3. Results

### 3.1. Performance of ESP-COAL

The type and magnitude of gene flow affected the ability to infer the correct ST topology using ESP-COAL (Fig. 2A). In general, models of historical gene flow did not greatly degrade the phylogenetic accuracy, regardless of the magnitude ($N_em$ = 0.01–1.00) or duration (0.1x$N_e$ or 0.5x$N_e$ generations) of gene flow. The

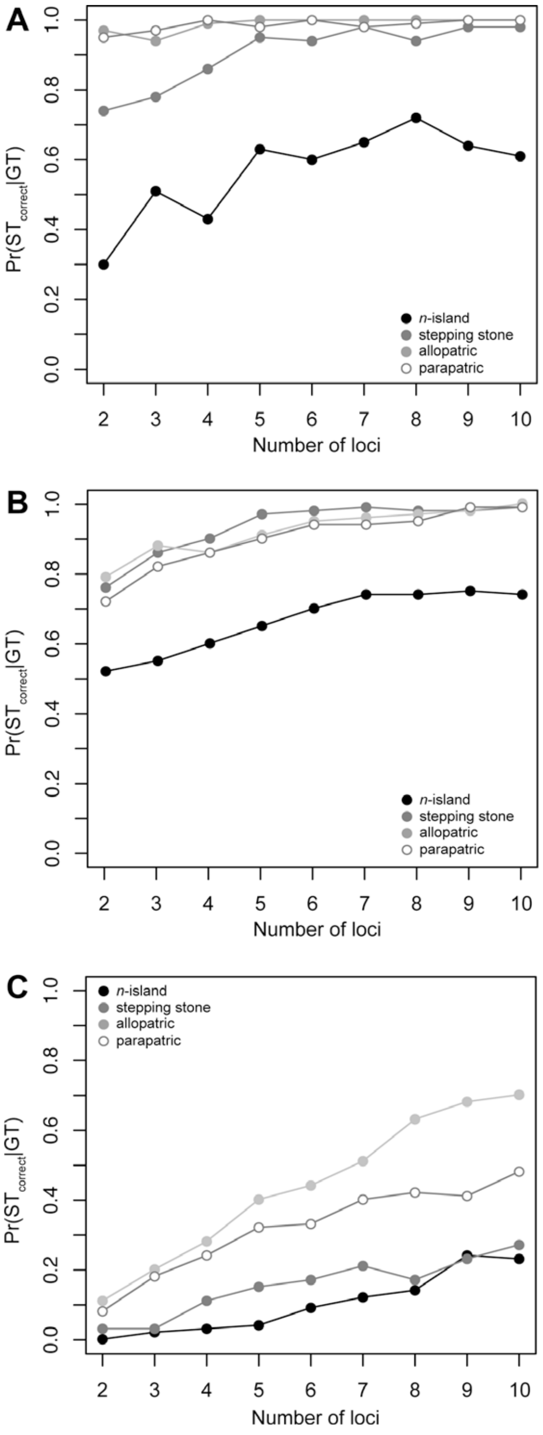

**Fig. 2.** Performance of (A) ESP-COAL, (B) MDC, and (C) concatenation for the four models of gene flow as shown in Fig. 1. All models had $N_em$ = 0.10, $N_e$ = 10,000, $d$ = 0.10, and a ST depth of 2$N_e$ generations.

probability of identifying the correct ST never dipped below 0.70 for any parameter combination for either the parapatric or allopatric models. In contrast, phylogenetic accuracy was low under the *n*-island and stepping stone models, with the magnitude of the performance drop depending more on the magnitude of gene flow than on the number of loci examined. For example, the two locus data sets for the *n*-island model at a magnitude of $N_em = 1.00$ only identified the correct ST with probabilities of 0.00–0.03 depending upon the value of $N_e$, but the phylogenetic accuracy did not improve when 10 loci were used (Table S1; online Supplemental data). However, at lower magnitudes of gene flow (e.g. $N_em = 0.01$), ESP-COAL retained reasonable power ($\Pr[ST_{correct}|GT] > 0.72$) to identify the correct species topology, so long as more than five loci were sampled (Fig. 2A and Tables S1–S3; online Supplemental data).

### 3.2. Performance of MDC

With one exception, phylogenetic accuracy was high in the MDC analyses across all models of gene flow and parameter combinations (Tables S4–S6, online Supplemental data). The lowest probability of estimating the true species tree for the stepping stone, parapatric, and allopatric models of gene flow was 0.69, and values were typically much higher so long as four loci were used. The performance of MDC was also correlated with the assumed value of $N_e$, the depth of the ST, and the number of sampled loci, with higher probabilities associated with larger values of each of those quantities. As in the ESP-COAL analysis, the *n*-island model presented the greatest difficulty to the analysis, although this was correlated with the strength of migration (Table S4, online Supplemental data). With $N_em = 0.01$, MDC was reasonably accurate (e.g. accuracy greater than 0.70) regardless of the numbers of loci. Accuracy decreased with $N_em = 0.10$, but still trended upwards as loci were added. However, when $N_em = 1.00$, the probability of identifying the correct ST was never greater than 0.34 (Fig. 2B).

### 3.3. Performance of concatenation

The identification of the correct species topology when data were concatenated was severely affected by both the magnitude and type of gene flow. In general, concatenation performed poorly under the *n*-island and stepping stone models even when the number of sampled loci increased (Fig. 2C). As the magnitude of gene flow increased under these models, probabilities of identifying the correct ST reached zero for all data sets (Tables S7–S9, online Supplemental data). For the models of historical gene flow, concatenation achieved reasonable power only when the number of loci sampled was greater than five, when gene flow was limited in magnitude ($N_em < 1.00$) and duration ($d = 0.10$), or when the depth of the ST increased.

### 3.4. A comparison among methods

Noticeable differences existed among the three methods considered for the inference of the species phylogeny (Table 1 and Fig. 2). Both ESP-COAL and MDC outperformed concatenation, especially for the *n*-island and stepping stone models (Fig. 2). Moreover, MDC performed better than ESP-COAL for those models, while ESP-COAL performed better for the allopatric and parapatric models. This difference was asymmetric, with MDC exhibiting much larger increases ($\Delta\Pr[ST_{correct}|GT] = 0.10–0.35$) relative to ESP-COAL under the *n*-island model than for ESP-COAL relative to MDC under the allopatric and parapatric models ($\Delta\Pr[ST_{correct}|GT] = 0.05–0.15$). In both cases, the differences between ESP-COAL and MDC narrowed as the number of sampled loci increased.

**Table 1**
Comparison of phylogenetic accuracy across four models of gene flow for $N_em = 0.10$, ST depth = $2N_e$, $N_e = 100{,}000$, and $d = 0.10$

| | Number of loci | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| *n-Island* | | | | | | | | | |
| ESP-COAL | 0.30 | 0.51 | 0.43 | 0.63 | 0.60 | 0.65 | 0.72 | 0.64 | 0.61 |
| MDC | 0.52 | 0.55 | 0.60 | 0.65 | 0.70 | 0.74 | 0.74 | 0.75 | 0.74 |
| CONC | 0.00 | 0.02 | 0.03 | 0.04 | 0.09 | 0.12 | 0.14 | 0.24 | 0.23 |
| *Stepping stone* | | | | | | | | | |
| ESP-COAL | 0.74 | 0.78 | 0.86 | 0.95 | 0.94 | 0.98 | 0.94 | 0.98 | 0.98 |
| MDC | 0.76 | 0.86 | 0.90 | 0.97 | 0.98 | 0.99 | 0.98 | 0.98 | 0.99 |
| CONC | 0.03 | 0.03 | 0.11 | 0.15 | 0.17 | 0.21 | 0.17 | 0.23 | 0.27 |
| *Allopatric* | | | | | | | | | |
| ESP-COAL | 0.97 | 0.94 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MDC | 0.79 | 0.88 | 0.86 | 0.91 | 0.95 | 0.96 | 0.97 | 0.98 | 1.00 |
| CONC | 0.11 | 0.20 | 0.28 | 0.40 | 0.44 | 0.51 | 0.63 | 0.68 | 0.70 |
| *Parapatric* | | | | | | | | | |
| ESP-COAL | 0.95 | 0.97 | 1.00 | 0.98 | 1.00 | 0.98 | 0.99 | 1.00 | 1.00 |
| MDC | 0.72 | 0.82 | 0.86 | 0.90 | 0.94 | 0.94 | 0.95 | 0.99 | 0.99 |
| CONC | 0.08 | 0.18 | 0.24 | 0.32 | 0.33 | 0.40 | 0.42 | 0.41 | 0.48 |

CONC, concatenated data approach.

The greatest degradation of phylogenetic signal occurred when gene flow was simulated using the *n*-island model, and as such this scenario provides the clearest illustration of the effects of varying parameters such as $N_em$, $N_e$, and ST depth. For example, phylogenetic accuracy decreases at a much faster rate as the amount of gene flow increases when ESP-COAL is used as opposed to MDC (Fig. 3A). While effective population size does not appear to have a large effect on phylogenetic accuracy (Fig. 3B), both ESP-COAL and MDC perform better under the *n*-island model when the species tree depth is shallow ($2N_e$) rather than deeper ($6N_e$; Fig. 3C). This finding is in contrast to the general trend demonstrated for MDC (Maddison and Knowles, 2006) and ESP-COAL (Carstens and Knowles, 2007), where accuracy increases with increasing species tree depth.

### 3.5. An empirical example from Rhododendron

Non-monophyly among the four *Rhododendron* species was apparent for both RNA polymerase data sets (Fig. 4A–C), despite differences in the number of polymorphisms, length of the aligned sequences, and sample size among sampled loci (Table 2). Estimated genealogies for each locus were fully resolved and differed in topology, with samples from *R. macrophyllum* often paraphyletic (Fig. 4). Analyses using ESP-COAL and MDC identified different species tree topologies as the most likely (Fig. 5A and B). However, the number of deep coalescences between the topology identified by ESP-COAL and that identified using MDC differed by only a single event (17 vs. 18). For the ESP-COAL analyses, the optimal tree was significantly better than the second best tree when compared using an approximate likelihood ratio test ($-2\Delta\ln L = 6.03$, df = 1, $P = 0.014$). Concatenation of the three data sets still resulted in a tree where the four species were not monophyletic (Fig. 5C).

Do the phylogenetic accuracy values reported here (Supplementary Tables 1–6) change appreciably when genealogies are estimated from sequence data, as opposed to using the actual genealogies? We reanalyzed the case thought to most closely approximate the *Rhododendron* data ($N_e = 10{,}000$, $N_em = 0.10$, $d = 0.10$, ST depth = $2N_e$) using gene trees estimated from sequence data simulated on our actual genealogies. Using this

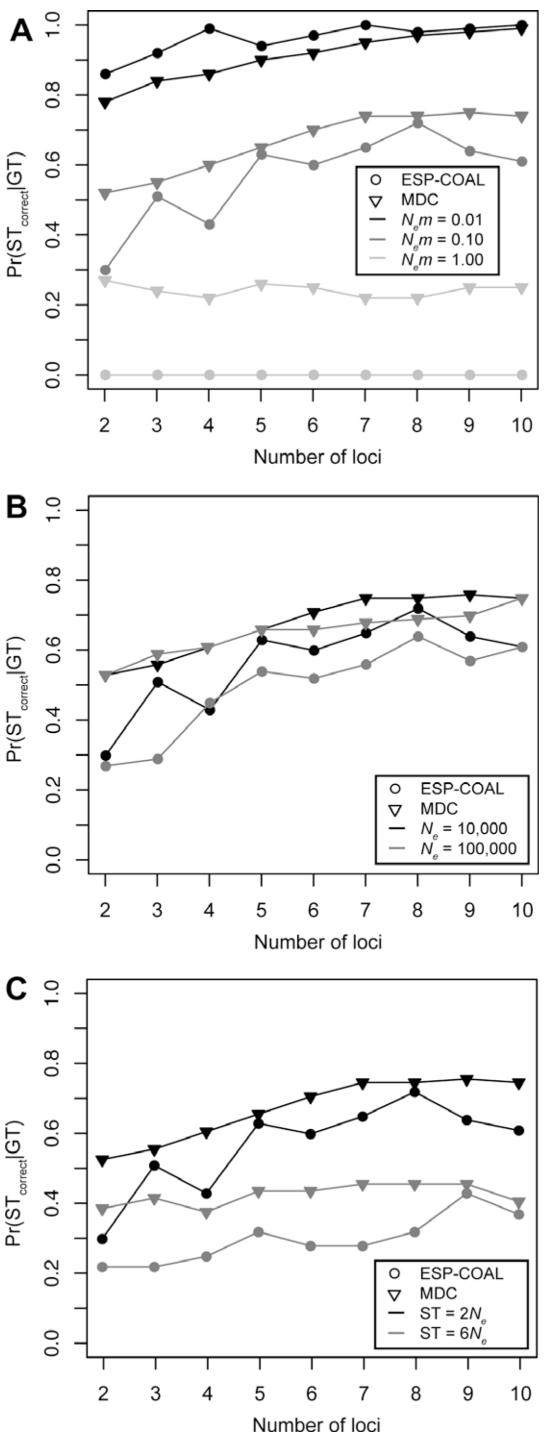

**Fig. 3.** Comparison of ESP-COAL and MDC assuming the *n*-island model of gene flow when (A) $N_e$ is 10,000, the depth of the species tree is equal to 2$N_e$, and migration rates vary from $N_em$ = 0.01–1.00, (B) $N_em$ is 0.10, the depth of the species tree is equal to 2$N_e$, and $N_e$ varies from 10,000 to 100,000, and (C) $N_e$ is 10,000, $N_em$ is 0.10, and the depth of the species tree varies from 2$N_e$ to 6$N_e$.

approach, accuracy decreased by slightly less than 10% in the ESP-COAL analyses and by slightly less than 2% in the MDC analyses (Fig. 6). We expect this decrease to be influenced by such factors as the number of variable sites in each locus as well as the topology of the species tree. This example illustrates one possible application of power analyses to phylogenetic investigations in empirical systems.

## 4. Discussion

### 4.1. Explanation of results

Incomplete lineage sorting has emerged as a common problem for phylogenetic inference at the species level. Given the volume of mathematical theory predicting this phenomenon (cf. Pamilo and Nei, 1988; Rosenberg, 2002, 2003), this may not be surprising. Several methods of inferring species phylogenies from gene trees have incorporated the stochastic process of incomplete lineage sorting (Maddison, 1997; Degnan and Salter, 2005; Liu and Pearl, 2007). While these methods are clearly at early stages of development, they appear to perform well when incomplete lineage sorting is the only process contributing to the discord between gene trees and species trees (Maddison and Knowles, 2006; Carstens and Knowles, 2007; Edwards et al., 2007). Here, we explore the further confounding effects of gene flow on the ability to identify the correct species tree topology using three methods: ESP-COAL, MDC, and concatenation. The results presented here suggest that methods derived from the coalescent are robust to the confounding effects of various models of gene flow and particularly to gene flow of lower magnitudes.

The greatest degradation of phylogenetic signal occurred under the *n*-island model with moderate to high levels of gene flow. In the *n*-island model all extant lineages have the same probabilities of sharing migrants per generation, and at sufficiently high levels of gene flow the notion of an underlying species phylogeny probably loses validity. The net effect of this process is to scramble the gene copies among lineages to such a magnitude that the gene tree topology in reference to the species topology becomes highly reticulate. However, at low magnitudes of gene flow ($N_em$ = 0.01), there remains some phylogenetic signal that can be recovered. Surprisingly, MDC performed quite well even at moderate levels of gene flow under the *n*-island model. This observation may partially be explained by the argument that the process of gene copy exchange among lineages may mirror the sorting of ancestral polymorphism when gene flow is equal among demes and large in magnitude (as in our simulations). Since deep coalescences become uncommon (Nielsen, 2005) as gene flow increases in magnitude under the *n*-island model, unequal migration among demes may present significant difficulties to MDC that are not predicted by our simulation results. In the case where gene flow is low, deep coalescences are more common, and gene copies have to wait to coalesce until they are in the same species lineage.

The stepping stone model presented more difficulties for ESP-COAL than MDC at high levels of gene flow ($N_em$ = 1.00). ESP-COAL can be misled when stepping stone migration mimics a spurious pattern of gene coalescence, especially between non-sister species, in multiple loci because the probability of these gene trees would be maximized under an incorrect species tree. Because MDC counts the number of deep coalescent events, and uses this as an optimality criterion across all gene trees, it may be less severely impacted by biological processes (like stepping stone migration) that increase the discord among loci. Of course, this explanation is valid only when divergence is not extremely recent or characterized by rapid radiations (cf. Degnan and Rosenberg, 2006).

Both ESP-COAL and MDC performed exceptionally well under models of historical gene flow (Tables S2–S3, S5–S6, online Supplemental data). In these models, factors such as the magnitude of the gene flow, the depth of divergence, and the effective population size exert less influence on phylogenetic accuracy than they do in the *n*-island or stepping stone models. In taxa where divergence occurs in parapatry, or where secondary contact has occurred, this suggests that methods for estimating species phylogeny from gene trees will be able to accurately estimate the species phylogeny

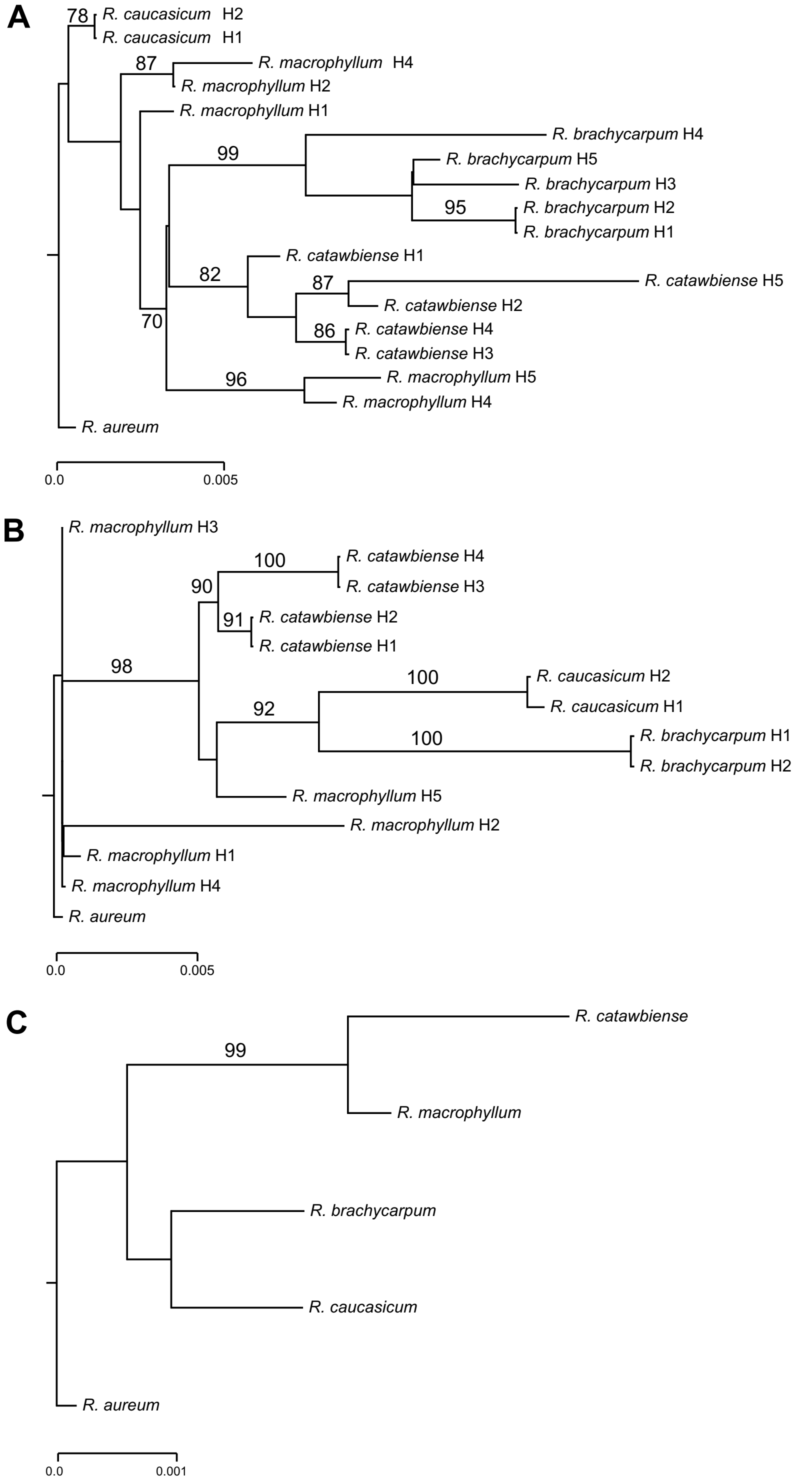

**Fig. 4.** An empirical example from *Rhododendron.* Maximum-likelihood genealogies for (A) *RPB*2, (B) *RPC*1, and (C) the chloroplast *trn*K and *trn*L-*trn*F gene regions. Numbers above or below branches are bootstrap support values (%). Only support values ⩾70% are shown.

even at the initial stages of divergence and at relatively high levels of migration. Concatenation, on the other hand, performed poorly for most models, even as the number of sampled loci increased. This may partially be an artifact of the simulation framework since concatenation analyses were confounded with the estimation of the true genealogy, as well as the fact that gene flow is a genome-wide phenomenon so that additional loci do not necessarily clarify phylogenetic relationships.

The optimism gleamed from these results, however, needs to be hedged with the realization that recent theoretical and simulation studies have shown that certain gene and species tree characteristics can lead to positively misleading results. For example, Degnan and Rosenberg (2006) prove that for species trees with ⩾5 tips there always exists a region of the branch length parameter space which guarantees that the most likely gene tree will not match the true underlying species tree. These optimal gene trees that differ

**Table 2**
Description of sample sizes, sequence diversity, and GenBank Accession Nos. for the data used to infer relationships among *R. brachycarpum*, *R. catawbiense*, *R. caucasicum*, and *R. macrophyllum*

| | *RPB*2 | *RPC*1 | Chloroplast (*trn*K, *trn*L-*trn*F) |
|---|---|---|---|
| *Sample size* | | | |
| *R. brachycarpum* | 5 | 2 | 1 |
| *R. catawbiense* | 5 | 4 | 1 |
| *R. caucasicum* | 2 | 2 | 1 |
| *R. macrophyllum* | 5 | 5 | 1 |
| *R. aureum* | 1 | 1 | 1 |
| Total | 18 | 14 | 5 |
| Sequence diversity | 37 (1301) | 71 (1596) | 18 (2718) |
| GenBank Accessions | EU822187–EU822204 | EU822173–EU822186 | AY494173–AY494176<br>AY496914–AY496917 |

All inferred trees were rooted using *R. aureum* as an outgroup. Sequence diversity is reported as the total number of polymorphic sites followed by the alignment length in parentheses. Insertion–deletion polymorphisms were ignored for all analyses.

from the underlying species tree were coined as anomalous gene trees (AGTs) and were shown to occur most often when deep internal branches within the true species tree were extremely short. For species trees with four tips and asymmetric topologies, as presented here (cf. Fig. 1), AGTs occur when the two internal branch lengths are approximately less than 0.156 coalescent time units or in the case of populations consistent with Wright-Fisher mating, $0.156N_e$ generations. The relative branch lengths used in the true species tree for our simulations, however, did not fall below this threshold (Fig. 1). Similarly, Kubatko and Degnan (2007) show that concatenated data sets suffer from similar problems of convergence to the incorrect species topology when single individuals are sampled and internal branches are short relative to external branches on the underlying species topology. The direct ramifications of the aforementioned studies is that methodologies relying on gene-species tree discordance (e.g. MDC) can quickly become misleading when divergence is recent or species radiations are rapid. It would be interesting in future work, therefore, to examine the effect of gene flow within this anomaly zone, because it is precisely those populations which diverged recently that are likely to share migrants.

### 4.2. Implications for empirical studies

The results presented here are of interest to two broad groups of molecular systematists; those who construct large phylogenies that include clades of closely related species, and those who conduct phylogeographic investigations. For the former, concatenation across loci may lead to statistical error in tip clades where the species from which the exemplars are sampled have a history of gene flow. While the degree to which error in the tip clades contributes to error at deeper nodes is unknown, it may be prudent to estimate species phylogeny for the tip clades using an approach such as MDC or ESP-COAL, and then to conduct the broader analysis using this estimate of species phylogeny as a constraint. Alternately, systematists could first conduct the broad analysis, and then double-check the results for poorly supported tip clades using one of these approaches.

Phylogeography studies have long assumed that the population structure implied by a single locus (commonly mitochondrial or chloroplast DNA in animals or plants, respectively) can be used as a proxy for the actual population structure. Since Avise (2000) and Knowles and Maddison (2002), an influx of methodological approaches derived from the coalescent have been incorporated into phylogeographic investigations (Kuhner et al., 1998; Beerli and Felsenstein, 1999, 2001; Pritchard et al., 2000; Nielsen and Wakeley, 2001; Hey and Nielsen, 2004, 2007; Kuhner, 2006; Hickerson et al., 2007). While the main objective of those methodologies is to estimate the parameters of interest using gene trees as a nuisance parameter, many phylogeographers would benefit directly from the ability to estimate robust species or population phylogenies. Our results suggest that, even at low levels of divergence, direct estimates the population phylogeny can be accurate given a modest number of loci (but see Degnan and Rosenberg (2006)). A model of the population phylogeny can be used as the basis for other analyses, particularly those that aim to test hypotheses pertaining to population demography using parametric bootstrapping (Knowles and Carstens, 2007).

Extending the utility of these methods will also aid in the inference of species relationships for many systematic and evolutionary questions across a broad set of taxa. This is true especially for organisms with large generation times, effective population sizes, and recent divergences where incomplete lineage sorting has been shown to be common (Syring et al., 2007). By using tree inference methods that incorporate incomplete lineage sorting into their formulation, rather than trying to diagnose the fact that incomplete lineage sorting has occurred using standard phylogenetic methods, (e.g. maximum parsimony) which inherently assume an unknown bi- or multi-furcating tree without reticulations, systematists can formulate and test biogeographic and evolutionary hypotheses previously unable to be addressed due to the lack of a reliable species topology.

We illustrate this change in focus by analyzing intraspecific DNA sequence data from two RNA polymerase and two chloroplast gene regions obtained from four *Rhododendron* species. The relationships among these species changes depending upon the gene region analyzed and whether the data are concatenated or collapsed into one consensus sequence per species (cf. Milne, 2004; Goetsch et al., 2005). In the best species tree identified using ESP-COAL, *R. caucasicum* is sister to the remaining three taxa, with *R. cawtabiense* being placed as sister to the *R. macrophyllum* and *R. brachycarpum* clade (Fig. 5A). Alternatively, the topology requiring the fewest number of deep coalescences places *R. macrophyllum* as sister to *R. brachycarpum* which together are sister to a clade composed of *R. catawbiense* and *R. caucasicum*. As pointed out previously, however, this topology required only one less deep coalescence as opposed to the topology identified by ESP-COAL (Fig. 5). It is important to note, moreover, that both of these topologies differ from those estimated by Milne (2004) and Goetsch et al. (2005) suggesting that incomplete lineage sorting and gene flow may complicate inference of shallow phylogenetic structure within this clade. These topologies also provide scaffolds from which to formulate and test biogeographic (Ree and Smith, 2008) and divergence time

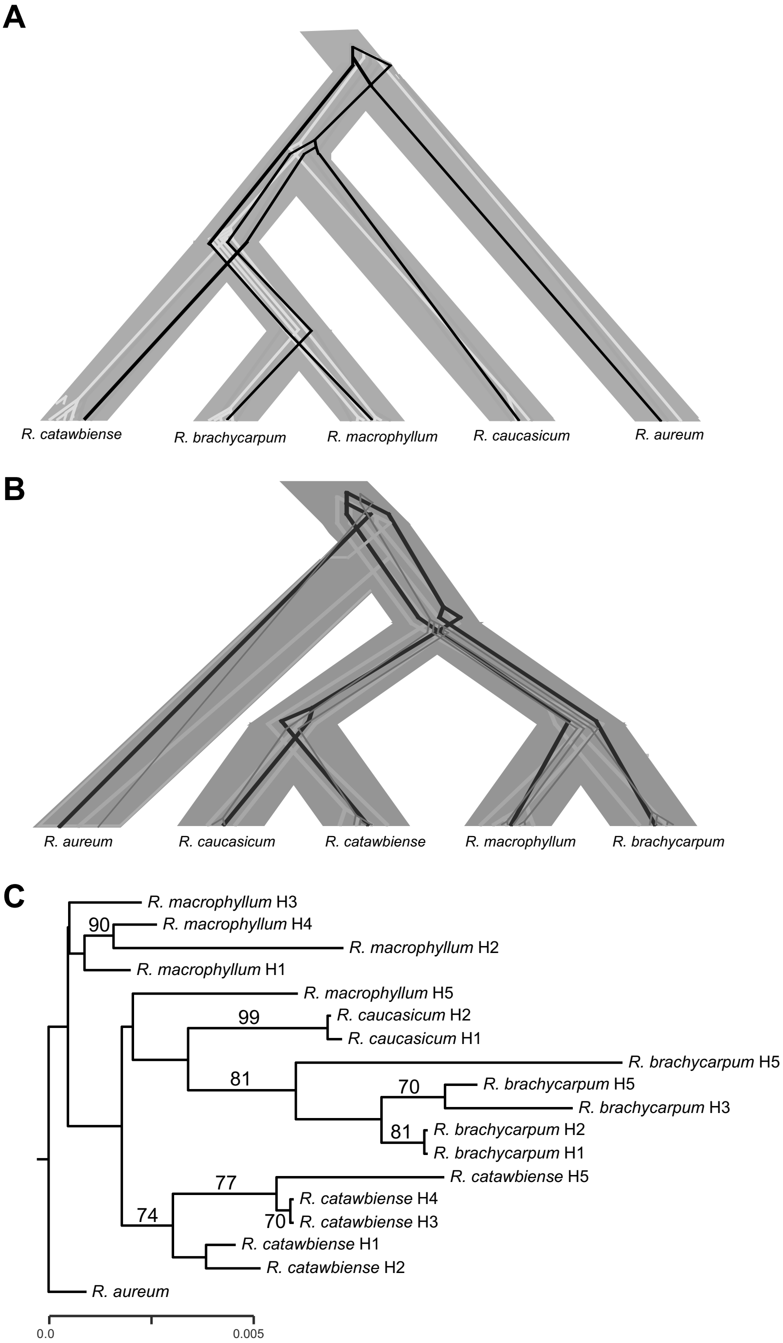

**Fig. 5.** An illustration of the three phylogenetic hypotheses corresponding to the evolutionary relationships among four *Rhododendron* species inferred using (A) ESP-COAL, (B) MDC, and (C) concatenation. Numbers above or below branches are bootstrap support values (%). Only support values ⩾70% are shown. Maximum-likelihood gene trees (cf. Fig. 4) for each locus are contained within the species trees.

(Sanderson, 2002) hypotheses previously hampered by a lack of monophyly.

The results presented here suggest that ESP-COAL and MDC perform well with respect to inference of a species topology when gene flow and incomplete lineage sorting are occurring among closely related species. Our simulations, however, did not encompass all possible configurations and magnitudes of gene flow or sampling designs. For example, unequal rates of gene flow among lineages could result in a reduction of the statistical power. Historical demographic events may also change our results and the effect of such events on the power of these methods remains to be investigated. Furthermore, the effect of incomplete species sampling remains unknown relative to the power of all three methods examined here. The magnitudes and models of gene flow investigated here, however, mirror those commonly used in population genetic and phylogeographic inference and our results provide an initial step forward in understanding the effects of common population processes on the ability to infer phylogenetic trees for closely related species and populations. Understanding the effects of such processes will not only aid in the inference of robust species topologies for taxonomic and conservation studies, but will also contribute to the emerging field of statistical phylogeography.

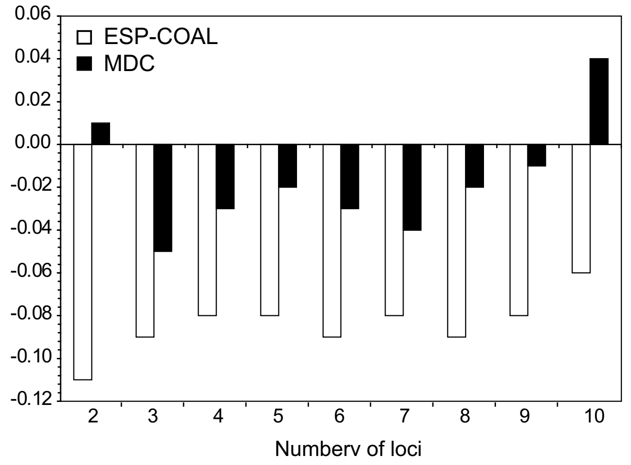

**Fig. 6.** An illustration exploring the effect that error in the estimation of gene trees has on the performance of ESP-COAL and MDC. Shown are results from conditions that correspond to the empirical *Rhododendron* data ($\theta = 20$, $N_e = 10{,}000$, $N_e m = 0.10$, $d = 0.10$), as well as a parapatric model of gene flow and a species tree depth equal to $2N_e$.

## Acknowledgments

We thank Benjamin Hall and Wennie Chou for providing the *Rhododendron* DNA sequences. Special thanks to Gabriel Rosa and John Liechty for assistance with the Department of Plant Sciences computing cluster located at the University of California, Davis and with PERL scripting. We thank Amy Litt, Jeffrey Oliver, and one anonymous reviewer for providing insightful comments that significantly improved this manuscript.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2008.09.008.

## References

Anisimova, M., Gascuel, O., 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. 55, 539–552.

Avise, J.C., 2000. Phylogeography: The History and Formation of Species. Harvard University Press, Cambridge, MA.

Beerli, P., Felsenstein, J., 1999. Maximum likelihood estimation of migration rates and population numbers of two populations using a coalescent approach. Genetics 152, 763–773.

Beerli, P., Felsenstein, J., 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc. Nat. Acad. Sci. USA 98, 4563–4568.

Brumfield, R.T., Liu, L., Lum, D.E., Edwards, S.V., 2008. Comparison of species tree methods for reconstructing the phylogeny of bearded manakins (Aves: Pipridae: Manacus) from multilocus sequence data. Syst. Biol. 57, 719–731.

Carling, M.D., Brumfield, R.T., 2008. Integrating phylogenetic and population genetic analyses of multiple loci to test species divergence hypotheses in *Passerina* buntings. Genetics 178, 363–377.

Carstens, B.C., Knowles, L.L., 2007. Estimating phylogeny from gene tree probabilities in *Melanoplus* grasshoppers despite incomplete lineage sorting. Syst. Biol. 56, 400–411.

Cox, P.A., Cox, K.N.E., 1997. The *Encyclopedia* of Rhododendron Species. Glendoick Publishing, Perth, Scotland.

Cummings, M.P., 1994. Transmission patterns of eukaryotic transposable elements: arguments for and against horizontal transfer. Trends Ecol. Evol. 9, 141–145.

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

Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 3, e68.

Edwards, S.V., Liu, L., Pearl, D.K., 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA 104, 5936–5941.

Excoffier, L., Novembre, J., Schneider, S., 2000. SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. J. Hered. 91, 506–509.

Felsenstein, J., 1985. Confidence limits on phylogenies. An approach using the bootstrap. Evolution 39, 783–789.

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.

Fitch, W.M., 1970. Distinguishing homologous from analogous proteins. Syst. Zool. 19, 99–113.

Goetsch, L., Eckert, A.J., Hall, B.D., 2005. The molecular systematics of Rhododendron (Ericaceae): a phylogeny based upon RPB2 gene sequences. Syst. Bot. 30, 616–626.

Hickerson, M.J., Stahl, E., Takebayashi, N., 2007. MsBayes: a flexible pipeline for comparative phylogeographic inference using approximate Bayesian computation (ABC). BMC Bioinformatics 8, e268.

Hey, J., 2006. Recent advances in assessing gene flow between diverging populations and species. Curr. Opin. Genet. Dev. 16, 592–596.

Hey, J., Nielsen, R., 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., Nielsen, R., 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., 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37, 203–217.

Hudson, R.R., 1990. Gene genealogies and the coalescent process. In: Futuyma, D.J., Antonovics, J. (Eds.), Oxford Survey Evolutionary Biology. Oxford University Press, New York, pp. 1–44.

Hudson, R.R., 1992. Gene trees, species trees and the segregation of ancestral alleles. Genetics 131, 509–512.

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

Knowles, L.L., Carstens, B.C., 2007. Estimating a geographically explicit model of population divergence for statistical phylogeography. Evolution 61, 477–493.

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

Kron, K.A., Judd, W.S., Stevens, P.F., Crayn, D.M., Anderberg, A.A., Gadek, P.A., Quinn, C.J., Luteyn, J.L., 2002. Phylogenetic classification of Ericaceae: molecular and morphological evidence. Bot. Rev. 68, 335–423.

Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56, 17–24.

Kuhner, M.K., 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22, 768–770.

Kuhner, M.K., Yamato, J., Felsenstein, J., 1998. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics 149, 429–434.

Liu, L., Pearl, D. K. 2006. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Technical report #53. Ohio State University.

Liu, L., Pearl, D.K., 2007. Species trees from gene trees: reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Bio. 56, 504–514.

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

Maddison, W.P., Knowles, L.L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21–30.

Maddison, W.P., Maddison, D.R., 2004. MESQUITE: a modular system for evolutionary analysis. Version 1.01. Available from: <http://mesquiteproject.org>.

Minin, V., Abdo, Z., Joyce, P., Sullivan, J., 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52, 674–683.

Nielsen, R., 2005. Molecular signatures of natural selection. Ann. Rev. Genet. 39, 197–218.

Milne, R.I., 2004. Phylogeny and biogeography of Rhododendron subsection Pontica, a group with a tertiary relict distribution. Mol. Phylogenet. Evol. 33, 389–401.

Nielsen, R., Wakeley, J.W., 2001. Distinguishing Migration from Isolation: an MCMC approach. Genetics 158, 885–896.

Page, R.D.M., Charleston, M., 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Mol. Phylogenet. Evol. 7, 231–240.

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

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Ree, R.H., Smith, S.A., 2008. Maximum-likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. Syst. Biol. 57, 4–414.

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

Rosenberg, N.A., 2003. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. Evolution 57, 1465–1477.

Sanderson, M.J., 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19, 101–109.

Slatkin, M., Maddison, W.P., 1989. A cladistic measure of gene flow inferred from phylogenies of alleles. Genetics 123, 603–613.

Slowinski, J.B., Page, R.D.M., 1999. How should species phylogenies be inferred from sequence data? Syst. Biol. 48, 814–825.

Swofford, D.L., 2002. PAUP∗. Phylogenetic Analysis Using Parsimony (and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Syring, J., Farrell, K., Businsky, R., Cronn, R., Liston, A., 2007. Widespread genealogical nonmonophyly in species of the *Pinus* subgenus *Strobus*. Syst. Biol. 56, 163–181.

Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105, 437–460.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.

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

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

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

Yang, Z., 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 162, 1811–1823.