Systematic Biology Advance Access published August 13, 2013

Running Head: GENE FLOW AND SPECIES TREE DISTORTION

The Influence of Gene Flow on Species Tree Estimation: A Simulation Study

ADAM D. LEACHÉ[1*], REBECCA B. HARRIS[1], BRUCE RANNALA[2,3], AND ZIHENG YANG[3,4]

[1]*Department of Biology and Burke Museum of Natural History and Culture, University of Washington, Seattle, WA 98195 USA*

[2]*Genome Center and Department of Evolution & Ecology, University of California, Davis, CA 95616, USA*

[3]*Center for Computational Genomics, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*

[4]*Department of Biology, University College London, Gower Street, London WC1E 6BT, UK*

**Correspondence to be sent to:*

Adam D. Leaché

Department of Biology, University of Washington, Seattle, WA 98195, USA.

Phone: 206 543 7622

Fax: 206 543 3041

Email: leache@uw.edu

Abstract.—Gene flow among populations or species and incomplete lineage sorting (ILS) are two evolutionary processes responsible for generating gene tree discordance and therefore hindering species tree estimation. Numerous studies have evaluated the impacts of ILS on species tree inference, yet the ramifications of gene flow on species trees remain less studied. Here, we simulate and analyze multilocus sequence data generated with ILS and gene flow to quantify their impacts on species tree inference. We characterize species tree estimation errors under various models of gene flow, such as the isolation-migration model, the *n*-island model, and gene flow between non-sister species or involving ancestral species, and species boundaries crossed by a single gene copy (allelic introgression) or by a single migrant individual. These patterns of gene flow are explored on species trees of different sizes (4 vs. 10 species), at different time scales (shallow vs. deep), and with different migration rates. Species trees are estimated with the multispecies coalescent model using Bayesian methods (BEST and *BEAST) and with a summary statistic approach (MPEST) that facilitates phylogenomic-scale analysis. Even in cases where the topology of the species tree is estimated with high accuracy, we find that gene flow can result in overestimates of population sizes (species tree dilation) and underestimates of species divergence times (species tree compression). Signatures of migration events remain present in the distribution of coalescent times for gene trees, and with sufficient data it is possible to identify those loci that have crossed species boundaries. These results highlight the need for careful sampling design in phylogeographic and species delimitation studies as gene flow, introgression, or incorrect sample assignments can bias the estimation of the species tree topology and of parameter estimates such as population sizes and divergence times. [*BEAST; BEST; coalescence; compression; dilation; introgression; MPEST; migration; simulation.]

Processes that generate gene tree discordance may hinder species tree estimation. One natural evolutionary process responsible for gene tree discordance across the entire tree of life is incomplete lineage sorting (ILS; Hudson 1983; Tajima 1983; Takahata 1995; Rannala and Yang 2008). Numerous studies have evaluated the impacts of ILS on species tree inference using simulated and empirical data and in doing so have provided practical advice for sampling design (Maddison and Knowles 2006; McCormack et al. 2009; Castillo-Ramírez et al. 2010; Heled and Drummond 2010; Camargo et al. 2011; Leaché and Rannala 2011).

Gene flow among populations and species is another evolutionary process that can generate gene tree discordance (Slatkin and Maddison 1989). The typical mode of species divergence whereby populations diverge under a model of strict allopatry is now being augmented with many empirical examples of divergence accompanied by gene flow (Pinho and Hey 2010), or allelic introgression across species boundaries (Wirtz 1999; Rheindt and Edwards 2011). However, the impacts of gene flow on species tree estimation and their ramifications on sampling design remain less studied (Eckert and Carstens 2008; Leaché 2009; Chung and Ané 2011; Heled et al. 2013).

Species tree inference methods that can effectively accommodate ILS are available; however, jointly considering ILS and gene flow remains a great challenge. Failing to account for gene flow during species tree estimation surely impacts parameter estimation, yet the resulting estimation errors are unclear. Bayesian methods for estimating species trees can accommodate population demographic parameters, such as population sizes and divergence times, but not migration (Liu et al. 2009; Heled and Drummond 2010). Choi and Hey (2011) recently proposed a method for the joint estimation of population demographic parameters, including gene flow, population assignments, and the species tree, but the method is currently applicable to only three species. Prior to this method, assuming a fixed species tree topology, known species assignments, and integrating across gene tree uncertainty was the only approach available for multilocus coalescent-based estimation of population sizes, divergence times, and gene flow (Nielsen and Wakeley 2001; Hey and Nielsen 2004; Kuhner 2009; Hey 2010).

Here, we quantify the impacts of gene flow on species tree inference by simulating multilocus data with varying levels of migration (Fig. 1). Several different models of gene flow are considered, including isolation-migration, paraphyletic gene flow between non-sister species, and ancestral gene flow occurring deeper in the species tree (Fig. 2). We also simulate data to mimic introgression of a single allele or migration of a single individual across a species boundary. We measure errors in estimates of the species tree topology, as well as divergence times and population sizes.

Materials and Methods

Data Simulations

Species trees are characterized by several parameters, including topology, depth (species divergence times) and width (population sizes for current and ancestral species). We measure species divergence time (τ) as the expected number of mutations per site, and population size as $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the mutation rate per nucleotide site per generation. In other words, θ is the average proportion of different sites between two sequences sampled at random from the population. The species trees used for gene tree simulations contain either 4 species or 10 species (Fig. 1). The rooted 4-species tree is sufficiently large to explore several different phylogenetic patterns of gene flow among species (Fig. 2), including 1) isolation-migration, modeled as gene flow between sister species, 2) paraphyletic gene flow, which involves gene flow between non-sister species, 3) ancestral gene flow, modeled as historical gene flow between sister lineages that ceases upon the divergence of a species, 4) a single migrant at $\tau = 0$ entering either a sister lineage or a non-sister lineage, which is equivalent to misclassification of a sample, and 5) allelic introgression at $\tau = 0$ where a species boundary is crossed by a single allele, which is similar to introgression of organellar DNA. The 10-species tree enables us to extend these scenarios to a larger phylogenetic context, as well as explore additional, more complex models, including an *n*-island model and models of gene flow that involve species with divergence times extending deeper into the tree (Fig. 2). The *n*-island model allows gene flow between all extant species and between ancestral species.

To introduce heterogeneity into the depth and width of the species tree, species divergence times (τ) and current and ancestral population sizes (θ) are drawn from separate prior probability distributions (Fig. 1). Simulation studies often generate test datasets using fixed parameter values; however, our Bayesian simulation strategy samples model parameters from a prior distribution, and we use the same prior distributions to analyze the simulated data (Huelsenbeck and Rannala 2004). Divergence times were simulated to produce relatively short trees to pose more challenging estimation problems (Maddison and Knowles 2006; Leaché and Rannala 2011). The species trees are ultrametric and the time gaps are independent exponential random variables, with mean $1/\lambda = 0.02$ expected mutations per site (Fig. 1b). This places the most recent species divergences (on average) at $\tau = 0.02$, and the mean root ages of $\tau = 0.06$ on the 4-species tree and $\tau = 0.12$ on the 10-species tree. Those parameter values reflect the low level of variation observed across nuclear loci in empirical species-level phylogenetic studies (Bell et al. 2010; Rowe et al. 2011). We use large values for the population size parameter θ to increase gene tree discordance, sampling from an inverse gamma distribution with parameters (α

= 3 and β = 0.003; Fig. 1c), with mean β/(α – 1) = 0.015. The values are chosen to reflect estimates obtained in studies of empirical data (Leaché, 2009; Castillo-Ramírez et al. 2010).

We simulated gene trees and multilocus nucleotide sequence data using the MCCOAL program in BPP v2.1a (Rannala and Yang 2003; Yang and Rannala 2010). Each simulation began with 100 species trees with species divergence times and population sizes sampled from the prior distributions described above. We sampled four sequences per species for all simulations, with the exception of the outgroup species, which only required one sequence for rooting purposes. MCCOAL generated gene trees for each species tree using the multispecies coalescent model (Rannala and Yang 2003), simulating the coalescent process in each population (Hudson, 2002). The mutation rates are assumed to be constant across loci (that is, the rate was fixed at 1). This is not a realistic assumption given that genes often evolve at different rates, but for our purposes we expect that including rate variation among loci for closely related species with low levels of divergence should only cause slight reductions in the effective number of variable sites for some loci. The gene trees were then used to simulate DNA sequences (1,000 bp per gene tree) along the branches of the genealogies using the Jukes-Cantor (JC) mutation model (Jukes and Cantor 1969). This simulation strategy produced average sequence divergences (uncorrected *p*-distances) of 1.2 – 1.6% within species and 4.8 – 5.7% between sister species.

The simulation program MCCOAL allows migration, even though Bayesian species tree inference programs assume no migration. Migration rates were assigned using the matrix $M = \{M_{ij}\}$, where the effective migration rate $M_{ij} = N_j m_{ij}$ is the expected number of migrants per generation from population *i* to population *j* (N_j is the population size of the receiving population *j*) and where m_{ij} is the migration rate from populations *i* to *j* defined as the proportion of individuals in population *j* that are immigrants from population *i*. One time unit is the expected time for one mutation to occur per site. The total coalescent rate in population i (with population size parameter θ_i) among a sample of n_i sequences is then $n_i(n_i - 1)/2 \times 2/\theta_i$. The coalescent rate between any pair of sequences in this population is $2/\theta_i$. The migration rate from population j into population i is then $4n_iM_{ji}/\theta_i$. The total migration rate for an individual in population i is the sum of the rates over all other populations. A full description of the MCCOAL migration simulation approach is given by Zhang et al. (2011).

We simulated data with no migration ($M_{ij} = 0$; ILS only) or up to four levels of migration ($M_{ij} > 0$; ILS plus migration), $M_{ij} = 0.001$, 0.01, 0.1, and 1.0 for the 4-species tree and $M_{ij} = 0.1$ and 1.0 for the 10-species tree. We restricted migration rates in the migration matrix M to reflect the phylogenetic patterns of migration outlined in Figure 2. Migration was assumed to be constant across the entire time interval. However, allelic introgression at $\tau = 0$ was simulated by replacing one allele from a locus with that of another species. A similar approach was used to generate datasets with a single migrant at $\tau = 0$, except here all sequences for a single individual were reassigned to a different species. Given that we sample four individuals per species, reassigning one sample produces an admixed population composed of 20% immigrants (i.e., 1/5). The introgression and single migrant datasets contained no other migration events in the migration matrix ($M_{ij} = 0$) to help clarify the impacts of recent gene flow at $\tau = 0$ versus on-going gene flow ($M_{ij} > 0$).

Bayesian Species Tree Estimation

We analyzed simulated datasets containing 10 loci with two Bayesian species tree estimation programs; *BEAST v1.6.2 (Heled and Drummond 2010) and BEST v2.3 (Liu 2008; Liu et al. 2008). These methods use the multispecies coalescent model to estimate species trees directly from the sequence data, calculating posterior probability distributions for gene trees, species trees, population sizes, and divergence times. BEST estimates the gene trees and then estimates the species tree using importance sampling (Liu et al. 2008). The gene trees and species tree are co-estimated in *BEAST (Heled and Drummond 2010).

We ran the MCMC algorithm for 10 million generations for four species and 100 million generations for 10 species sampling every 5,000 steps with a 25% burn-in. The run lengths were sufficient to generate effective sample sizes exceeding 200. Convergence was assessed in a subset of analyses by checking for stationarity in likelihood scores and tree lengths (using TRACER v1.5; Rambaut and Drummond 2007), and the posterior probability of clades (using AWTY; Nylander et al. 2008). Convergence problems prevented us from using BEST on the 10-species datasets. The JC model of nucleotide substitution was used for all loci to match the simulation conditions. The strict molecular clock was assumed and a Yule process prior is used for the divergence times in the species tree. For *BEAST, the population size model was set to “constant”. An inverse gamma prior ($\alpha = 3$, $\beta = 0.03$) was used for the population sizes (θ). For BEST, the mutation rate across loci was fixed at 1. The prior distributions for the effective population sizes and divergence times were the same as those used to simulate the data. The population size θ was modeled using an inverse gamma distribution ($\alpha = 3$, $\beta = 0.03$), corresponding to a prior mean for the population size of $\theta = 0.015$. Branch lengths were drawn from an exponential distribution ($\lambda = 50$), which results in an average branch length of 0.02 expected substitutions per site.

We calculated the posterior mean and variance of the divergence times and population sizes as well as the taxon bipartition probabilities. We also evaluated whether the true species tree was contained in the 95% credible set. The mean values of the posterior summaries are averaged across 100 replicate simulations.

Phylogenomic Simulations

We analyzed simulated datasets containing either 10 loci or 1,000 loci using the program MPEST v1.2 (Liu et al. 2010). The method estimates species trees from a set of gene trees by maximizing a pseudo-likelihood function. The fast computation times make the approach advantageous for large phylogenomic datasets, since full Bayesian methods such as *BEAST and BEST can only seem to handle small numbers of loci. We note that MPEST uses the estimated gene tree topologies only, and ignores information in the branch lengths and uncertainties in the estimated gene trees. As so little information is used, not all parameters in the multispecies coalescent model are identifiable. For closely-related species, the sequences may contain little phylogenetic information and the gene trees may be unresolved or highly uncertain, so that MPEST may not be expected to work well. We conducted MPEST analyses with 10 loci to provide a direct comparison with *BEAST and BEST. The 1,000 locus MPEST simulations help determine whether gene flow distortions identified with 10 loci are ameliorated with phylogenomic data. We used the simulated gene trees and gene trees estimated from the simulated sequence data as input into MPEST. In practice, we can expect gene tree estimation to introduce additional error into the species tree estimation procedure, and we start with DNA sequences to provide a more direct comparison between MPEST and *BEAST/BEST. Gene trees were estimated from DNA sequence data using RAxML-HPC v7.5.9 (Stamatakis 2006) with the proper species used to root the trees.

Distribution Of Coalescent Times

When post-divergence gene flow is absent, the coalescent times for alleles from different species must be greater than the species divergence time. With post-divergence gene flow, this expectation is not true anymore. Coalescent times between alleles from different species may thus be indicative of gene flow after species divergences (= shallow coalescence). We plot the estimated coalescent times for alleles from different species against the simulated species divergence time to distinguish between deep and shallow coalescence.

For the case of the 4-species tree, we contrast the true coalescent times in the simulated gene trees, which contain no inference error, with coalescent times estimated from the simulated sequence data. For both simulated and estimated gene trees we calculated the minimum coalescent time for alleles from different species using the R package phybase (Liu 2010). Polytomies were resolved using the multi2di function, which inserts branch lengths of 0 into unresolved nodes. A matrix of branch lengths corresponding to each node was constructed from each gene tree using the read.tree.nodes command, and the minimum coalescent times between alleles belonging to different species were found using the coaltime command.

The Bayesian simulation strategy used to simulate species trees introduces variability in the species divergence times. Here, we consider the deviation of the smallest coalescent time. Deviation is computed as the difference in the minimum gene tree coalescent time (t_{AB}) between species A and B from the true species divergence time (τ_{AB}):

$$D = \frac{t_{AB} - \tau_{AB}}{\tau_{AB}} \qquad (1)$$

$D = 0$ represents the inflection point separating deep coalescences (some of which will reflect ILS) and shallow coalescences (gene flow). With no post-divergence gene flow, D will be positive and near zero. The coalescent times can extend far back into the ancestral population, and therefore a large D indicates a large ancestral population size. With gene flow, D may be negative. Unlike the deep coalescences that reach far back into the species tree and produce large D values, shallow coalescences are bounded by $\tau = 0$, and therefore gene trees with $t_{AB} = 0$ will result in $D = -1$. Finally, D can also be negative in the absence of gene flow due to gene tree inference errors.

RESULTS

Effect of Gene Flow on Tree Probabilities

Summary statistics that describe the 95% credible sets of trees obtained from the *BEAST analyses of the simulated data, and the percentage of times the true species tree is contained in the 95% credible interval (coverage probability) are shown in Tables 1 and 2. BEST results for the 4-species analyses are largely similar to *BEAST and are provided in the Supplemental Materials (Dryad doi:10.5061/dryad.b7jh4). When $M = 0$ (ILS only) the coverage probability is 1.0 for the 4-species tree and 0.94 for the 10-species tree. Under the isolation-migration model, the 95% credible sets of trees indicate that the true tree is recovered more decisively, including a decrease in the size of the 95% credible set, a reduction of the maximum size observed across the 100 replicates, and an increase in coverage probability for the 10-species tree (Table 1). Paraphyletic gene flow results in similar reductions in the average and maximum sizes of the 95% credible sets; however, the sharp decrease in the coverage probability accompanying increasing migration rates indicates that the method is not recovering the true tree

(Table 1). These patterns are most pronounced under the model of deep paraphyly on the 10-species tree (coverage probability = 0; Table 1). Ancestral gene flow results in an increase in the average size of the 95% credible set of trees, and reduced coverage probability when $M = 1.0$ on the 4-species tree (Table 1).

Allelic introgression and migration of a single individual across species boundaries each reduce the average and maximum sizes of the 95% credible set (Table 2). However, their impacts on the coverage probability differ between sister species and non-sister species movement. Allelic introgression between sister species results in increased coverage probability, and for non-sister species the coverage probability is reduced (Table 2). These patterns are similar when a single individual migrates across a species boundary, but the increases (sister species) and reductions (non-sister species) in coverage probabilities are more extreme (Table 2).

Effect of Gene Flow on Posterior Clade Probabilities

The posterior probability for clades is a sensitive metric for measuring the effects of gene flow on inference of the species tree. Figures 3 - 6 present the mean values (over 100 replicates) for the posterior probabilities, divergence times, and population sizes. The full results including standard deviations for parameter estimates are included in the Supplemental Materials.

Under the isolation-migration model, migration increases the support for the true clade and essentially overcomes any uncertainty generated by ILS once $M \geq 0.01$; this pattern holds for both the 4-species tree (Fig. 3) and the 10-species tree (Fig. 4). Paraphyletic and ancestral gene flow both add gene tree discordance to the already present ILS, and under these models, whether gene flow is restricted to shallow or deep levels of the tree, the posterior probability for the true clade declines sharply and can result in strong support for an incorrect topology (Figs. 3 and 4).

When M = 0.1 and gene flow is paraphyletic, the posterior probability for the true clade is reduced to under 0.2 (Fig. 3 and 4). The posterior probability remains high at this point for ancestral gene flow, but decreases sharply at M =1.0 on the 4-species tree (Fig. 3). If we ignore random errors due to limited data, the posterior probability for the true clade should go to 1, 1/3, and 0 when M goes to infinity for these three gene flow scenarios (isolation-migration, ancestral, and paraphyletic, respectively).

Under the n-island model on the 10-species tree, the posterior probability for the clade undergoing gene flow (which includes 4 species) increases to 1.0 (Fig. 4). However, the posterior probability for the true relationships within this clade decreases with increasing migration rate until all 15 possible rooted topologies for the 4 species have nearly equal posterior probability (Table 3). Deep paraphyletic gene flow in the 10-species tree (Fig. 4) produces strong support (posterior probability = 0.99) for an incorrect topology uniting species H and I. Deep ancestral gene flow in the 10-species tree (Fig. 4) increases the posterior probability for the clades exchanging migrants, but reduces the posterior probability for the two clades stemming from the gene flow episode. These patterns all become more drastic with the increase of the migration rate (Fig. 4).

Allelic introgression and single migrants between sister species both increase the posterior probability for the true tree, while the same processes occurring between non-sister species produce support for incorrect species trees (Figs. 5 and 6).

Effect of Gene Flow on Estimation of Species Divergence Times

Divergence times are underestimated under all gene flow scenarios explored in our simulations (Figs. 3-6), and the divergence times approach τ = 0 under high migration rates (Fig.

3). The divergence times that are underestimated are restricted to the species and clades that exchanged migrants in the simulation, while the posterior estimates for divergence times for the remaining clades in the species tree match closely with those estimated with $M = 0$ (ILS only; Figs. 3-6; Supplemental Materials). For example, gene flow between sister species leads to underestimates of the divergence time between them, whereas the remaining clades in the trees are unaffected on the 4-species and the 10-species trees (Figs. 3-6). Paraphyletic gene flow results in underestimated species divergence times for the two species exchanging migrants, which form an inaccurate clade, as well as for the clade that represents the most recent common ancestor of the paraphyletic species exchanging migrants on the true tree (Figs. 3-6). When gene flow is restricted to an ancestral time episode, the divergence time for that clade is underestimated, but the divergence times for the clades stemming from the gene flow event are not underestimated (Figs. 3-4).

The estimation errors in species divergence times caused by the migration of a single individual are much greater compared to allelic introgression (Figs. 5 and 6). The divergence time for τ_{AB} on the 4-species tree (Fig. 5), or τ_{EF} on the 10-species tree (Fig. 6), are an order of magnitude lower than that found with $M = 0$. This strong bias is produced regardless of whether the single migrant crosses a sister species or non-sister species boundary.

Effect of Gene Flow on Estimates of Population Size Parameters

The population size parameter (θ) is overestimated when species exchange migrants, and these overestimates are restricted to the species and clades exchanging migrants (Figs. 3-6). The average estimated population sizes (over 100 replicates) are shown in Figures 3-6. The posterior mean for θ is a close match to the prior ($\theta = 0.015$). Under the isolation-migration and paraphyly models with $M = 1.0$ on the 4-species tree, the posterior estimates for θ_A and θ_B double (Fig. 3). A similar increase is also seen on the 10-species trees (Fig. 4). Ancestral gene flow does not result in any discernable impacts on population size estimates for contemporary species, but the ancestral population size is overestimated for the ancestral population exchanging migrants, and this pattern is seen on both the 4-species and 10-species trees (Figs. 3 and 4). The *n*-island model overestimates θ for the species exchanging migrants and the ancestral branch leading to the clade exchanging migrants (Fig. 4).

When a single migrant crosses a species boundary into a non-sister species, the posterior estimate of θ nearly doubles for the species receiving the migrant and the ancestral species (Figs. 5-6). Overestimation on a similar scale is apparent in the simulation of allelic introgression, yet here the overestimation of θ is restricted to the ancestral species (Figs. 5-6).

Effect of Gene Flow on Phylogenomic Estimates of Species Trees

The accuracy of the MPEST species trees estimated with 10 or 1,000 loci is shown in Tables 1 and 2. We calculate accuracy as the percentage of replicates (out of 100) that match the true species tree. Accuracy when $M = 0$ (ILS only) is 100% for the 4-species tree and 94% for the 10-species tree. These results are similar to the coverage probabilities calculated from the *BEAST analyses (Tables 1 and 2). Accuracy does not change substantially under the isolation-migration or ancestral gene flow models, but paraphyletic gene flow and the *n*-island models cause sharp reductions in accuracy. The deep paraphyly model results in the worst performance (as low as 0% accuracy), which is similar to the results from *BEAST. Using estimated gene trees as opposed to simulated gene trees reduces accuracy, but in some instances this pattern is not found when analyzing 1,000 loci (Tables 1 and 2). This may be due to our use of relatively high mutation rates (large values for θs and τs) in the simulation so that the sequences at each locus are fairly divergent and informative.

Single migrants between non-sister species do not reduce accuracy nearly as much in MPEST as it does in the *BEAST analyses (Table 2). With 1,000 loci in MPEST, the accuracy decreases to 91% on the 10-species tree, whereas with *BEAST the probability of finding the true tree in the 95% credible set is only 0.07 (Table 2). We did not investigate single locus introgression, but we predict that the impact of single locus introgression is likely to be minimal when estimating species trees with 1,000 loci.

Distribution of Coalescent Times

As expected, the distribution of true (simulated) coalescent times when $M = 0$ resembles an exponential distribution and produces positive D values (Fig. 7a). This same distribution remains detectable when M is increased to $M = 0.01$ (Figs. 7b-7c), but when $M = 0.1$ the simulated distribution becomes bimodal (Fig. 7d). Here, one peak tracks deep coalescence gene trees while the other records shallow divergence gene trees. The bimodal distribution of coalescent times largely disappears with increasing migration rates and is replaced with a single curve beginning at deviation = -1, which corresponds to gene tree $t_{\mathrm{AB}} = 0$ (Fig. 7e).

Gene tree coalescent times estimated by *BEAST from the sequences involve estimation errors (Fig. 7). Since gene flow was absent when $M = 0$, the negative D values from the estimated gene trees are the result of stochastic estimation errors and not true shallow divergences (Fig. 7a). Nevertheless, the prominent peak that corresponds to the deep divergence distribution found by the simulated gene trees is still present at $M = 0.001$ (Fig. 7b) and $M = 0.01$ (Fig. 7c). Whereas the true coalescent times produced a bimodal distribution for $M = 0.1$, the evidence for two peaks is absent with the estimates (Fig. 7d) and replaced with a relatively broad and flat distribution. When M = 1.0 the data produce a single curve beginning at deviation = -1, which corresponds to gene tree $t_{AB} = 0$ (Fig. 7e).

Discussion

Gene Flow and Species Tree Inference

Nonmonophyletic gene trees are common in empirical studies of well-established species (Carling and Brumfield 2008; Carstens and Dewey 2010; Lee et al. 2012). The large effective population size of nuclear genes and thus the large effect of ILS make it less likely that any single nuclear gene will recover the true species tree (Knowles and Carstens 2007). As a result, species tree estimation using multiple nuclear loci is quickly replacing single locus studies as a best practice in phylogenetics (Brito and Edwards 2009; Degnan and Rosenberg 2009) and species delimitation (Yang and Rannala 2010; Ence and Carstens 2011). The influence of ILS on phylogeny estimation has been studied quite extensively (Degnan and Salter 2005; Degnan and Rosenberg 2006; Maddison and Knowles 2006); however, the effect of gene flow on species tree inference has received far less attention. We have referred to the units of our simulations as species, but under some simulation conditions (i.e., high migration rates) they are probably more accurately described as populations belonging to the same species. However, empiricists often find themselves in the quandary of not knowing whether the units of analysis are populations or species. Thus, the results of our simulations are relevant to phylogeographic and species delimitation studies where we may, or may not, be dealing with different species.

Accounting for ILS in phylogenetic studies is imperative, since this process is intrinsically linked to all speciation events (Edwards 2009). On the other hand, gene flow among populations or species is not expected to accompany all speciation events, and therefore it is unnecessary to account for the process in every phylogenetic study. The impacts of gene flow on species tree estimation can be quite severe (Fig. 8), and our simulations show the ways in which failing to recognize gene flow can bias species tree estimates. We found that the phylogenetic pattern of gene flow plays a great role in determining the type of biases observed in the species tree topology, and that the migration rate then modulates the degree of parameter estimation error. Adding more loci is not likely to correct these errors, and our simulations with 1,000 loci demonstrate that species tree accuracy will suffer most under paraphyletic patterns of gene flow. Eckert and Carstens (2008) investigated species tree inference against four models of gene flow (*n*-island, stepping stone, parapatric, and allopatric) and found that the coalescent methods ESP-COAL (Carstens and Knowles 2007) and minimizing deep coalescences (Maddison 1997) typically worked better than concatenation at identifying the correct species tree. We investigated the influence of gene flow on species tree inference using Bayesian species tree inference (*BEAST and BEST) and MPEST. These Bayesian methods incorporates branch length information and genealogical uncertainty, and also provides posterior probability estimates of divergence times and population sizes, two important demographic patterns that are not necessarily estimated by species tree inference methods that do not utilize the multispecies coalescent model.

We found that gene flow between sister species increases the probability of estimating the correct species tree topology. In this case, gene flow is operating as a homogenizing force, which is acting to decrease the observed divergence between the sister species. Most phylogenetic methods should interpret this increase in similarity as strong evidence for shared ancestry; however, we only explored this under the context of species tree estimation using the multispecies coalescent model. Conversely, gene flow between species that are not sister taxa produces gene trees that are discordant with the species tree, which increases the difficulty of estimating the correct species tree. These findings seem intuitive from a phylogeny estimation perspective. From a species delimitation perspective, however, where sharp patterns of genealogical division help distinguish independent evolutionary lineages, it becomes difficult to distinguish the species boundary as it becomes blurred by gene flow (Zhang et al. 2011). Furthermore, the estimation errors that gene flow cause on divergence time and population size estimates that we identified are not as intuitive as the overall impact of gene flow on topology.

We leave a number of potential factors untested in our simulations. Avenues for expansion from our current simulations include: 1) additional gene flow scenarios, including models that enable pulses of migration through time, 2) mutation rate variation among loci, and modulating the sampling intensity of genes and individuals, 3) population size changes through time, including expansion-contraction models that mimic Pleistocene glacial cycles, 4) differential selection on subsets of loci, and 5) identifying the circumstances under which population subdivision could produce the same biases as gene flow (Slatkin and Pollack, 2008; Yu et al. 2011).

Species Tree Compression and Dilation

Tree topology is frequently used to assess the accuracy of phylogenetic tree reconstructions, yet species tree shapes convey other types of biologically relevant information in addition to the topology. The depths of branches indicate species divergence times, while the width of branches denote population sizes (Nichols 2001). From these dimensions, we can make inferences about the speciation history of a clade.

We characterize two types of distortions that gene flow causes on species trees in addition to changes in the topology and posterior probability for clades. The impact of gene flow on the overall shape of species trees is shown in Figure 8. The first type of distortion is species tree compression, which results from the underestimation of species divergence times. Compression causes the speciation times (τ) to appear more recent. We did not observe the opposite phenomenon, where the divergence times are overestimated and stretched deeper back in time. The inference model (multispecies coalescent model) assumes that all gene tree discordance is due to incomplete lineage sorting, and this forces the speciation times to delay until the gene flow event time. As a result, divergence times are underestimated. A similar result was found in a simulation study of horizontal gene transfer (Chung and Ané, 2011), which is similar to the paraphyly scenarios of gene flow investigated here. The second type of distortion is species tree dilation, or overestimation of the population size (θ). We did not observe any instances of branch attenuation. This overestimation of θ may be a consequence of underestimating τ, since the model has to account for the sequence diversity in the data. Alternatively, dilation could also be a consequence of the incoming migrants instantly increasing the effective population size of the sink population.

Our simulations suggest that species tree distortions due to gene flow are dependent on the phylogenetic locations of gene flow. For instance, gene flow between sister species causes them to experience compression and dilation while leaving other parts of the species tree unaffected (Fig. 8). Paraphyletic gene flow between non-sister species is more misleading in that it causes compression of all species divergence times subtending the gene flow event. In our simulations, the clade containing the non-sister species only included three species (on the 4-species tree) or up to eight species (on the 10-species tree), but we expect that a similar pattern of compression would occur if more species were included in the clade. Under paraphyletic gene flow, dilation appeared to be restricted to those species exchanging genes. This indicates that the parts of the species tree not affected by gene flow may remain comparatively easier to reconstruct. In a study of *Sceloporus* lizards, Leaché (2009) found that inaccurate species assignments produced posterior estimates of θ that were up to an order of magnitude higher than estimates obtained with correct species assignments, suggesting that deviations in θ could be useful for identifying rogue samples or identifying cryptic lineages.

Identifying Outlier Loci

The difficulty of distinguishing instances of gene tree incongruence stemming from ILS or gene flow has impeded the development of phylogenetic methods that can accommodate both processes simultaneously (but see Kubatko 2009; Meng and Kubatko 2009; Gerard et al. 2011; Yu et al. 2012). Under a standard phylogenetic model of no post-divergence gene flow, alleles from different species cannot coalesce until species divergence (i.e., until they are in the same ancestral population). Thus, the coalescent times for alleles can provide useful information for distinguishing deep coalescence from post-divergence gene flow. Joly et al. (2009) developed a posterior predictive approach for distinguishing hybridization from ILS based on the idea that minimum genetic distances between sequences from two species should be smaller for hybridization events than for ILS (Joly 2012). Yang (2010) developed a likelihood ratio test that compares variable species divergence times across loci under a model of allopatric speciation without gene flow against an alternative model of parapatric speciation with gene flow. The method requires hundreds of loci to achieve reasonable statistical power, a demand that is still difficult to meet with most empirical datasets for non-model organisms, but not insurmountable with current next-generation sequencing techniques (Glenn 2011).

Comparing the minimum coalescent times for alleles belonging to different species is a potential solution for identifying loci that may be crossing species boundaries (Sang and Zhong 2000; Holder et al. 2001). The simulation conditions used here produced a peak of gene tree coalescent times that corresponded to the species divergence time when $M = 0$, and increasing the migration rate induced a secondary peak corresponding to $\tau = 0$ (Fig. 7). Identifying a bimodal distribution in gene tree coalescent times with empirical data, which is suggestive of genetic exchange, will require more loci than are typically available, but this constraint is vanishing as more studies shift towards new sequencing technologies (Hohenlohe et al. 2010; vonHoldt et al. 2011). Although signatures of gene flow events were present under the simulation conditions used here, their presence in empirical data will depend on the level of divergence between species as well as the mutation rates of the sampled genes. Recent species divergence times and/or uninformative loci may result in a preponderance of gene tree coalescence times near $\tau = 0$, and this would make it difficult to distinguish ILS from gene flow. Alternative coalescent-based methods are available for estimating gene flow among populations under a variety of population models (Excoffier and Heckel 2006), and conducting these analyses alongside species tree inference is a logical way of identifying whether gene flow may be introducing biases into the species tree estimation procedure (Carling and Brumfield, 2008). However, this approach is rarely taken, since it is generally assumed that gene flow is either absent, has not occurred in the past, or is relatively unimportant compared to ILS in the context of estimating the species tree. A species tree approach that includes migration estimation (see

Choi and Hey 2011) would eliminate the need to conduct side-by-side population genetic and phylogenetic analyses to understand the divergence history of a clade.

Our simulations reveal some of the ways in which gene flow may bias species tree estimation, and that the estimation errors can impact different dimensions of the species tree. This highlights the need for careful sampling design in phylogenetic studies where gene flow, introgression, or incorrect sample assignments can potentially bias the estimation of the Bayesian species tree topology, population sizes, and divergence times.

Supplementary Material

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository at http://datadryad.org, doi:10.5061/dryad.b7jh4.

Funding

This work was funded by grants from the National Science Foundation (DBI-0805455, DBI-1144630) and the Royalty Research Fund (UW, A61649) awarded to A.D.L., and a Biotechnological and Biological Sciences Research Council grant (UK) and Royal Society/Wolfson Research Merit Award to Z.Y.

Acknowledgments

We thank the University of California Davis Genome Center Bioinformatics Core for providing computing infrastructure. We thank B. Carstens, S. Joly, and L. Kubatko for their comments, and J. Heled and A. Drummond for helpful advice on implementing species tree simulations with *BEAST.

REFERENCES

Bell R.C., Parra J.L., Tonione M., Hoskin C., MacKenzie J.B., Williams S.E., Moritz C. 2010. Patterns of persistence and isolation indicate resilience to climate change in montane rainforest lizards. Mol. Ecol. 19:2531–2544.

Brito P.H., Edwards S.V. 2009. Multilocus phylogeography and phylogenetics using sequence-based markers. Genetica 135:439–455.

Camargo A., Avila L.J,. Morando M., Sites Jr J.W. 2012. Accuracy and precision of species trees: effects of locus, individual, and base-pair sampling on inference of species trees in lizards of the *Liolaemus darwinii* group (Squamata, Liolaemidae). Syst. Biol. 61:272–288.

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 species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Malanoplus* grasshoppers. Syst. Biol. 56:400–411.

Carstens B.C., Dewey T.A. 2010. Species delimitation using a combined coalescent and information-theoretic approach: an example from North American *Myotis* bats. Syst. Biol. 59:400–414.

Castillo-Ramírez, S., Liu L., Pearl D., Edwards S.V. 2010. Bayesian estimation of species trees: a practical guide to optimal sampling and analysis. In: Knowles L.L., Kubatko L.S., editors. Estimating species trees: practical and theoretical aspects. Wiley-Blackwell: New Jersey. p. 15–33.

Choi S.C., Hey J. 2011. Joint inference of population assignment and demographic history. Genetics 189:561–577.

Chung Y., Ané C. 2011. Comparing two Bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer. Syst Biol. 60:261–275.

Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genetics. 2:762–768.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

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

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

Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.

Ence D.D., Carstens B.C. 2011. SpedeSTEM: a rapid and accurate method for species delimitation. Mol. Ecol. Res. 11:473–480.

Excoffier L., Heckel G. 2006. Computer programs for population genetic data analysis: a survival guide. Nat. Rev. Genetics 7:745–758.

Gerard D., Gibbs H.L., Kubatko L. 2011. Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. BMC Evol. Biol. 11:291.

Glenn T.C. 2011. Field guide to next-generation DNA sequencers. Mol. Ecol. Res. 11:759–769.

Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.

Heled J., Bryant D., Drummond A.J. 2013. Simulating gene trees under the multispecies coalescent and time-dependent migration. BMC Evol. Biol. 13:44.

Hey J. 2010. Isolation with migration models for more than two populations. Mol. Biol. Evol.27:905–920.

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.

Hohenlohe P.A., Bassham S., Etter P.D., Stiffler N., Johnson E.A., Cresko W.A. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genetics 6:e10000862.

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

Hudson R.R. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37:203–217.

Hudson R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Huelsenbeck J.P., Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. 53: 904–913.

Huelsenbeck J.P., Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Joly S., McLenachan P.A., Lockhart P.J. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. Am. Nat. 174:E54-E70.

Joly S. 2012. JML: testing hybridization from species trees. Mol. Ecol. Resour. 12:179–184.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian protein metabolism. Academic Press: New York. P. 21–123.

Knowles L.L., Carstens B.C. 2007. Delimiting species without monophyletic gene trees. Syst. Biol. 56:887–895.

Kubatko L.S. 2009. Identifying hybridization events in the presence of coalescence via model selection. Syst. Biol. 58:478–488.

Kuhner M.K. 2009. Coalescent genealogy samplers: windows into population history. Trends Ecol. Evol. 24:86–93.

Leaché A.D. 2009. Species tree discordance traces to phylogeographic clade boundaries in North American fence lizards (*Sceloporus*). Syst. Biol. 58:547–559.

Leaché A.D., Rannala B. 2011. The accuracy of species tree estimation under simulation: a comparison of methods.□Syst. Biol. 60:126–137.

Lee J.Y., Joseph L., Edwards S.V. 2012. A species tree for the Australo-Papuan Fairy-wrens and allies (Aves: Maluridae). Syst. Biol. 61:253–271.

Liu L. 2010. Phybase: Basic functions for phylogenetic analysis. R package version 1.1. http://CRAN.R-project.org/package=phybase

Liu L., Yu L., Kubatko L., Pearl D.K., Edwards S.V. 2009. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53:320–328.

Liu L., Yu L., Edwards S.V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10:302.

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.

McCormack J.E., Huang H., Knowles L.L. 2009. Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. Syst. Biol. 58:501–508.

Meng C., Kubatko L.S. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. Theor. Pop. Biol. 75:35–45.

Nichols R. 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16:358–364.

Nielsen R., Wakeley J. 2001. Distinguishing migration from isolation: A Markov chain Monte Carlo approach. Genetics 158:885–896.

Nylander J.A.A., Wilgenbusch J.C., Warren D.L., Swofford D.L. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24:581–583.

Pinho C., Hey J. 2010. Divergence with gene flow: models and data. Annu. Rev. Ecol. Evol. Syst. 41:215-230.

R Development Core Team 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rambaut A., Drummond A.J. 2007. Tracer. University of Oxford: Oxford.

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

Rannala B., Yang Z. 2008. Phylogenetic inference using whole genomes. Annu. Rev. Genomics Hum. Genet. 9:217–231.

Rannala B., Zhu T., Yang Z. 2012. Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. Mol. Biol. Evol. 29:325–335.

Rheindt F.E., Edwards S.V. 2011. Genetic introgression: an integral but neglected component of speciation in birds. Auk 128:620–632.

Rowe K.C., Aplin K.P., Baverstock P.R., Mortiz C. 2011. Recent and rapid speciation with limited morphological disparity in the genus *Rattus*. Syst. Biol. 60:188–203.

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

Slatkin M. 1985. Gene flow in natural populations. Annu. Rev. Ecol. Syst. 16:393–430.

Slatkin M., Maddison W.P. 1989. A cladistics measure of gene flow inferred from the phylogeny of alleles. Genetics 123:603–613.

Slatkin M., Pollack J.L. 2008. Subdivision in an ancestral species creates asymmetry in gene trees. Mol. Biol. Evol. 25:2241–2246.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analysis with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

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

Takahata N. 1995. A genetic perspective on the origin and history of humans. Annu. Rev. Ecol. Syst. 26:343–372.

vonHoldt B.M., Pollinger J.P., Earl D.A., Knowles J.C., Boyko A.R., Parker H., Geffen E., Pilot M., Jedrzejewski W., Jedrzejewska B., Sidorovich V., Greco C., Randi E., Musiana M., Kays R., Bustamante C.D., Ostrander E.A., Novembre J., Wayne RK. 2011. A genome-wide perspective on the evolutionary history of wolf-like canids. Genome Research 21:1294–1305.

Wirtz P. 1999. Mother species–father species: unidirectional hybridization in animals with female choice. Animal Behaviour 58:1–12.

Yang Z. 2010. A likelihood ratio test of speciation with gene flow using genomic sequence data. Genome Biol. Evol. 2:200–211.

Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. USA. 107:9264–9269.

Yu Y., Degnan J.H., Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genetics 8:e1002660.

Zhang C., Zhang D.X., Zhu T., Yang Z. 2011. Evaluation of a Bayesian coalescent method of species delimitation. Syst. Biol. 60:747–761.

TABLE 1. The coverage probability and average size of the 95% credible set of trees obtained from the *BEAST analyses. Species tree accuracy using MPEST is recorded as the percentage of correct topologies.

		*BEAST				MPEST	
		Coverage	95% credible	Min.	Max.	10 loci	1,000 loci
Model	*M*[a]	probability	set size	size	size	est./sim.	est./sim.
No migration (ILS[b] only)							
4	0	1.0	1.6	1	8	96/98	100/100
10	0	0.94	5.4	1	57	61/82	97/94
Isolation-migration							
4	0.001	0.99	1.5	1	7	94/97	100/100
4	0.01	1.0	1.3	1	3	95/99	99/100
4	0.1	1.0	1.4	1	3	99/100	100/100
4	1	1.0	1.4	1	3	100/100	100/100
10	0.1	0.99	4.3	1	33	64/84	97/94
10	1	0.99	4.7	1	28	71/85	98/93
n-island							
10	0.1	0.63	15.4	1	61	9/12	20/22
10	1	0.38	19.3	3	68	9/5	12/14
Paraphyly							
4	0.001	0.92	1.6	1	4	95/98	100/100
4	0.01	0.55	1.2	1	3	86/94	94/96
4	0.1	0.17	1.1	1	3	45/46	43/46
4	1	0.03	1.1	1	3	9/9	8/9
10	0.1	0.13	3.0	1	15	26/39	33/35
10	1	0.0	4.3	1	40	4/7	8/8
Deep paraphyly							
10	0.1	0.0	3.4	1	21	4/5	4/3
10	1	0.0	3.2	1	13	0/0	0/0
Ancestral							
4	0.001	0.98	1.6	1	7	97/98	100/100
4	0.01	1.0	1.4	1	12	92/98	100/100
4	0.1	1.0	1.5	1	8	95/96	100/100
4	1	0.98	2.7	1	12	54/60	97/100
10	0.1	0.99	6.2	1	43	66/77	98/94
10	1	0.98	7.6	1	59	29/51	96/94
Deep ancestral							
10	0.1	0.99	6.6	1	61	57/78	98/94
10	1	0.96	8.2	1	53	31/27	96/95

[a] *M*, migration rate.

[b] ILS, incomplete lineage sorting.

TABLE 2. The coverage probability and average size of the 95% credible set of trees obtained from the *BEAST analyses under simulations of single locus introgression and migration of a single individual at $\tau = 0$. Species tree accuracy using MPEST is recorded as the percentage of correct topologies (out of 100 replicates). Simulations marked “–“ were not conducted.

		*BEAST				MPEST	
Species	Model	Coverage probability	95% credible set size	Min. size	Max. size	10 loci est./sim.	1,000 loci est./sim.
No migration (ILS[a] only)							
4	ILS	1.0	1.6	1	8	96/98	100/100
10	ILS	0.94	5.4	1	57	61/82	97/94
Single migrant							
4	Sister species	1.0	1.4	1	3	99/100	100/100
4	Non-sister species	0.09	1.1	1	3	92/94	92/97
10	Sister species	0.98	4.6	1	56	66/83	97/94
10	Non-sister species	0.07	3.7	1	25	59/79	90/91
Deep single migrant							
10	Non-sister species	0.0	5.8	1	24	38/77	55/63
Single locus introgression							
4	Sister species	0.99	1.4	1	3	–	–
4	Non-sister species	0.37	1.2	1	3	–	–
10	Sister species	0.99	4.6	1	54	–	–
10	Non-sister species	0.28	3.8	1	24	–	–
Deep single locus introgression							
10	Non-sister species	0.0	6.5	1	28	–	–

[a] ILS, incomplete lineage sorting

TABLE 3. The posterior probabilities for the 15 possible rooted trees under the *n*-island model. The true tree contains clade (H,(G,(E,F))). Values are averages across 100 *BEAST analyses.

Tree	$M = 0$	$M = 0.1$	$M = 1$
(H,(G,(E,F)))	0.86	0.08	0.05
(H,(F,(E,G)))	0.01	0.07	0.05
(H,(E,(F,G)))	0.02	0.06	0.06
(G,(H,(E,F)))	0.02	0.07	0.06
(G,(F,(E,H)))	0.00	0.08	0.07
(G,(E,(F,H)))	0.00	0.09	0.06
(F,(H,(E,G)))	0.00	0.08	0.08
(F,(G,(E,H)))	0.00	0.05	0.07
(F,(E,(G,H)))	0.00	0.04	0.07
(E,(H,(F,G)))	0.00	0.05	0.06
(E,(G,(F,H)))	0.00	0.05	0.06
(E,(F,(G,H)))	0.00	0.05	0.07
((G,H),(E,F))	0.03	0.05	0.04
((F,H),(E,G))	0.00	0.03	0.05
((E,H),(F,G))	0.00	0.04	0.04

Figure Captions

FIGURE 1. Species trees used for simulating data for 10 species (*a*) and 4 species (*a'*) and prior probability distributions used for simulating species divergence times (*b*) and population sizes (*c*).

FIGURE 2. Gene flow patterns explored through simulation: (*a*) Isolation-migration; (*b*) paraphyly model of gene flow between non-sister species; (*c*) ancestral gene; and (*d*) a single migrant or single gene copy (e.g., allelic introgression) crossing a species boundary at $\tau = 0$. In each of the four cases we consider (I) 4 species, (II) 10 species, and (III) 10 species with deep introgression/gene flow.

FIGURE 3. The impacts of gene flow in the case of four species. Estimated θ values are plotted on the tree, and posterior probabilities are in bold. Gene flow patterns are indicated with colored boxes, and the order of species is fixed for all trees. Parameters are averages across 100 replicate runs. Plots show the parameters that are most heavily impacted by gene flow, including (a) the posterior probability for the true clade containing species A and B, (b) divergence times for the most recent common ancestor of the species experiencing gene flow, and (c) population size estimates for the most recent common ancestor of the species experiencing gene flow.

FIGURE 4. The impacts of gene flow in the case of 10 species. Estimated θ values are plotted on the tree, and posterior probabilities are in bold. Gene flow patterns are shown with colored boxes, and the order of species is fixed for all trees. The species trees depicted are for

simulations with $M = 0.1$; simulations with $M = 1.0$ produce more extreme results. Plots show the parameters that are most heavily impacted by gene flow for the *n*-island, deep paraphyly, and deep ancestral gene flow models, including (a) the posterior probability for the true clade containing species E, F, G, and H, (b) divergence times for the most recent common ancestor of the species experiencing gene flow, and (c) population size estimates for the most recent common ancestor of the species experiencing gene flow.

FIGURE 5. Single locus introgression and migration of a single individual on the 4-species tree. Estimated θ values are plotted on the tree, and posterior probabilities are in bold. The order of species is fixed for all trees. Species tree parameters are averages across 100 replicates.

FIGURE 6. Single locus introgression and migration of a single individual on the 10-species tree. Estimated θ values are plotted on the tree, and posterior probabilities are in bold. The order of species is fixed for all trees. Species tree parameters are averages across 100 replicates. Symbols indicate the locations of the clades shown in panels.

FIGURE 7. Distribution of gene tree coalescent times under a model with no migration (*a*) and with migration (*b*-*e*). Frequency histograms are shown for estimated gene trees (top; white) and simulated gene trees (bottom; gray). $D = 0$ (equation 1) represents the inflection point separating deep coalescences (positive values) and shallow coalescences (negative values). The *x*-axis is bounded by -1 for shallow coalescence and unbounded for deep coalescences (but truncated at +1.5 for clarity).

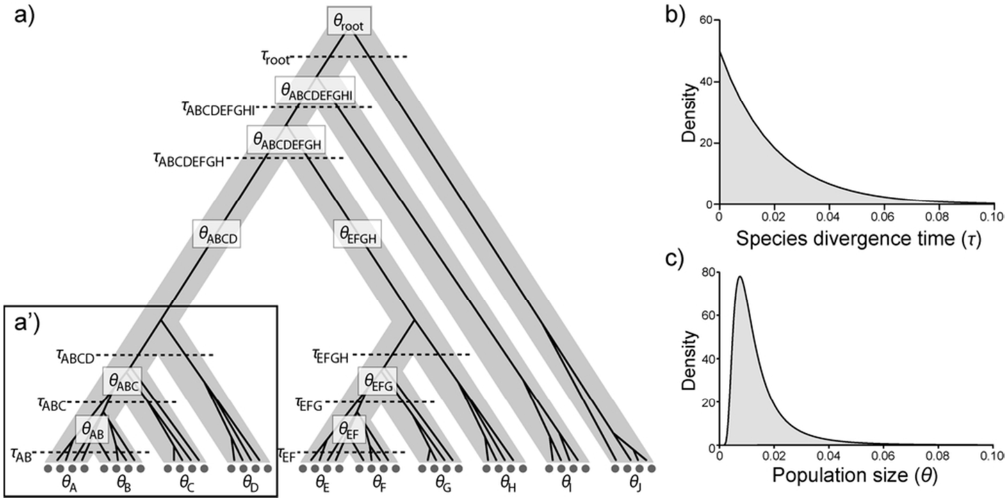
FIGURE 8. Species tree distortions caused by gene flow that can result from coalescent methods that only model ILS. Dashed lines illustrate species tree compression, and the widening of branches illustrates species tree dilation in relation to the starting species tree.

Data file(s):
Supplemental Materials

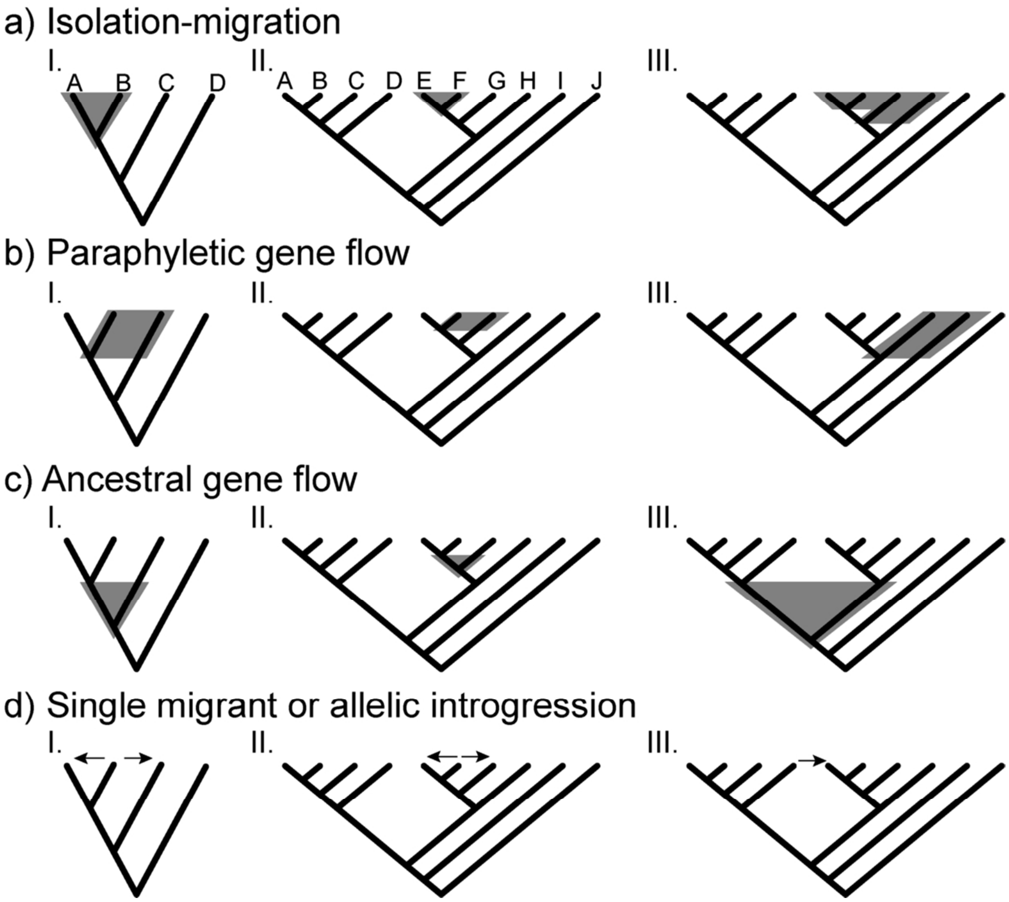
Dryad has assigned the following provisional DOI to the submission. This DOI may be included in the article manuscript. Although this DOI is not yet fully registered with the DOI system, it will be registered when the manuscript is ready for publication.

doi:10.5061/dryad.b7jh4

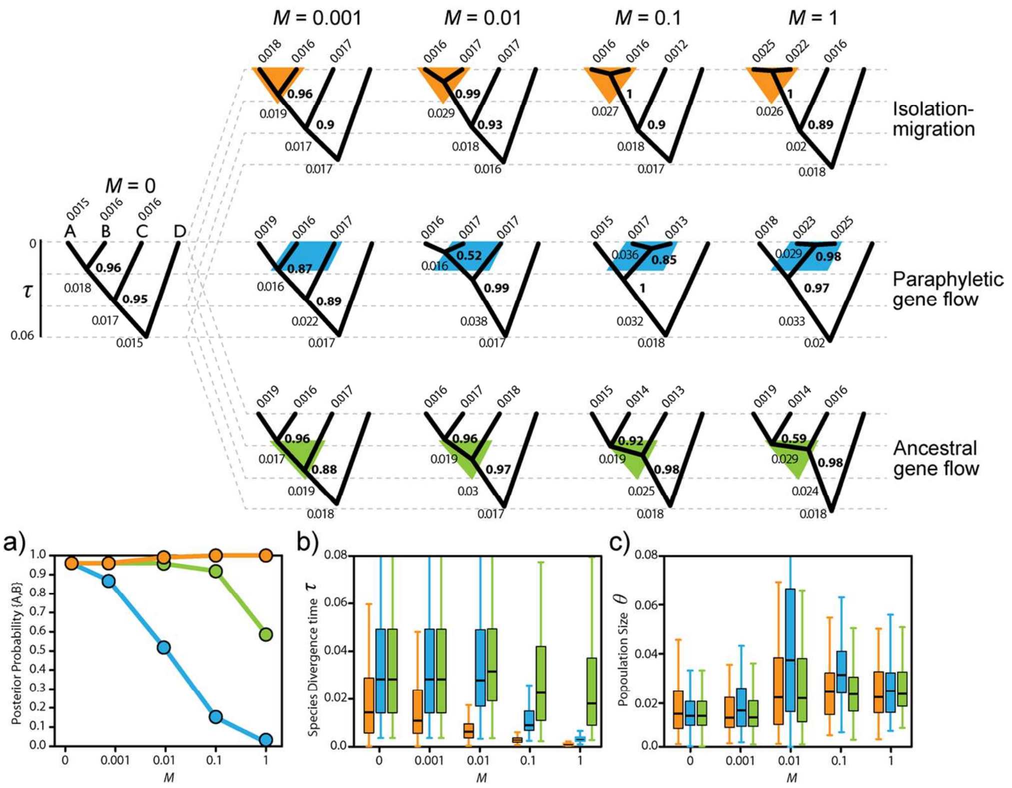
Journal editors and anonymous peer reviewers may view the submission for review purposes using the following url:
http://datadryad.org/review?wfID=16576&token=d22a6160-e5e4-400a-bf2c-d90bab2f89c0



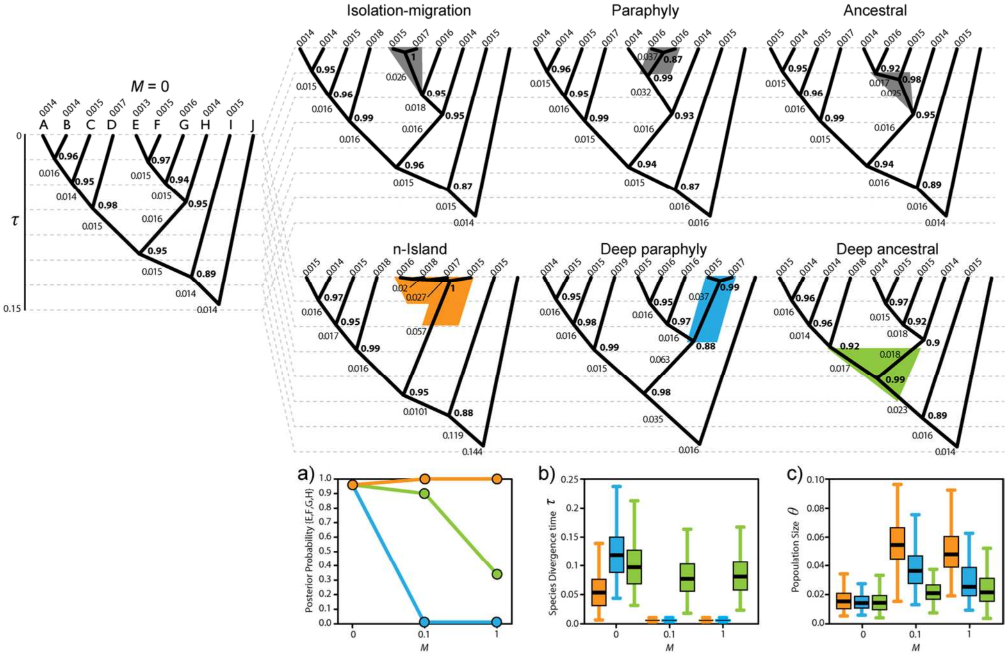
Species trees used for simulating data for 10 species (a) and 4 species (a') and prior probability distributions used for simulating species divergence times (b) and population sizes (c).
87x43mm (300 x 300 DPI)



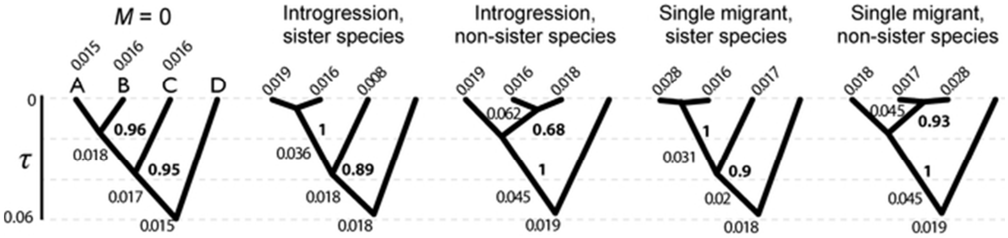
Gene flow patterns explored through simulation: (a) Isolation-migration; (b) paraphyly model of gene flow between non-sister species; (c) ancestral gene; and (d) a single migrant or single gene copy (e.g., allelic introgression) crossing a species boundary at τ = 0. In each of the four cases we consider (I) 4 species, (II) 10 species, and (III) 10 species with deep introgression/gene flow.
104x93mm (300 x 300 DPI)



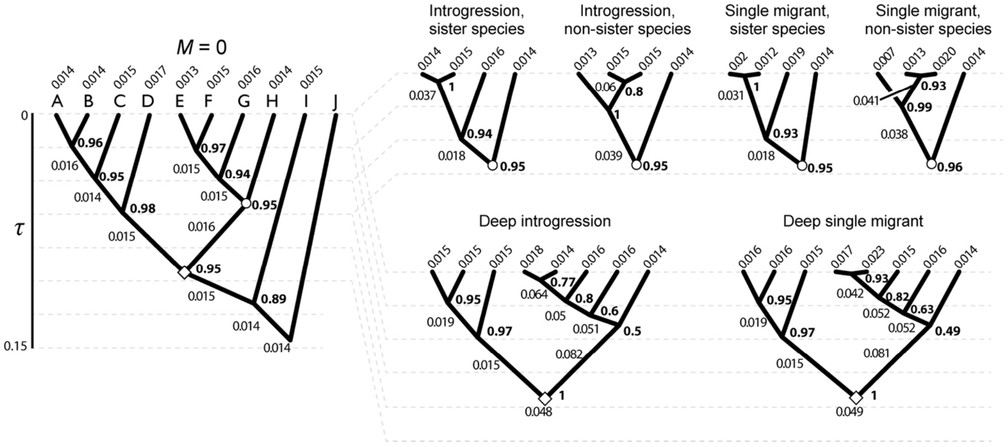
The impacts of gene flow in the case of four species. Estimated θ values are plotted on the tree, and posterior probabilities are in bold. Gene flow patterns are indicated with colored boxes, and the order of species is fixed for all trees. Parameters are averages across 100 replicate runs. Plots show the parameters that are most heavily impacted by gene flow, including (a) the posterior probability for the true clade containing species A and B, (b) divergence times for the most recent common ancestor of the species experiencing gene flow, and (c) population size estimates for the most recent common ancestor of the species experiencing gene flow.
122x95mm (300 x 300 DPI)



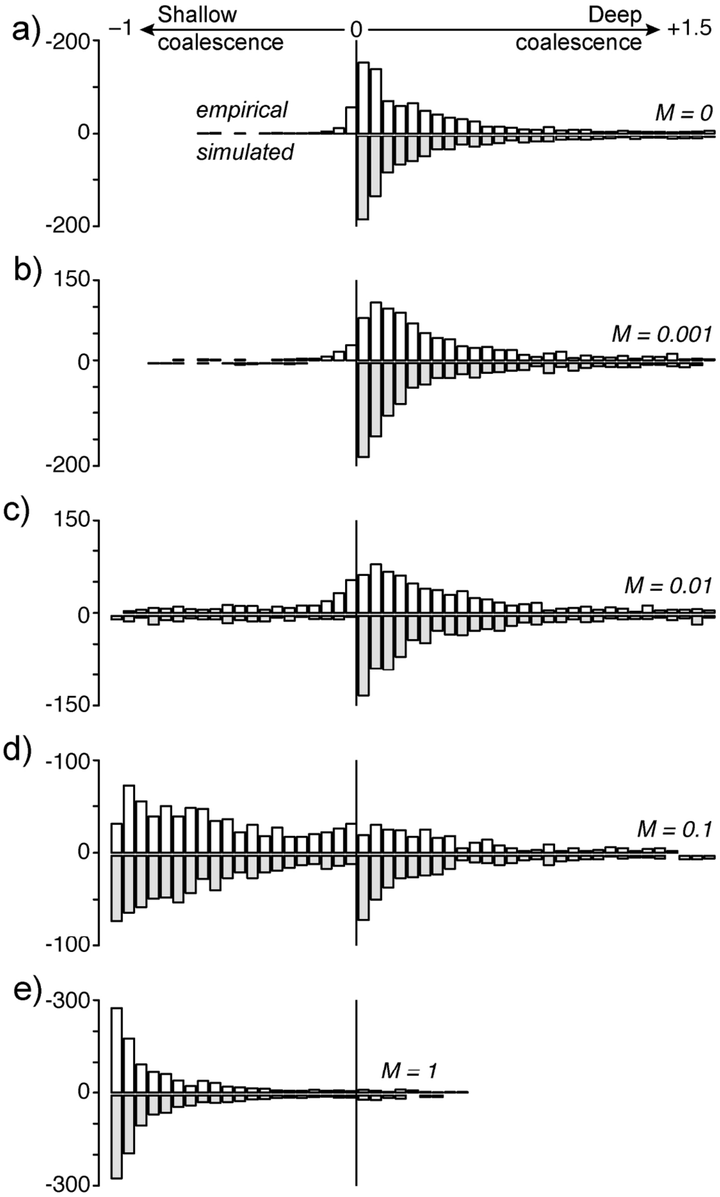
The impacts of gene flow in the case of 10 species. Estimated θ values are plotted on the tree, and posterior probabilities are in bold. Gene flow patterns are shown with colored boxes, and the order of species is fixed for all trees. The species trees depicted are for simulations with M = 0.1; simulations with M = 1.0 produce more extreme results. Plots show the parameters that are most heavily impacted by gene flow for the n-island, deep paraphyly, and deep ancestral gene flow models, including (a) the posterior probability for the true clade containing species E, F, G, and H, (b) divergence times for the most recent common ancestor of the species experiencing gene flow, and (c) population size estimates for the most recent common ancestor of the species experiencing gene flow.
135x88mm (300 x 300 DPI)



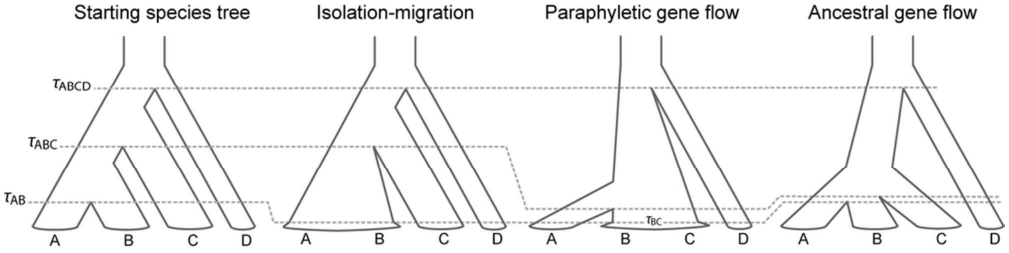
Single locus introgression and migration of a single individual on the 4-species tree. Estimated θ values are plotted on the tree, and posterior probabilities are in bold. The order of species is fixed for all trees. Species tree parameters are averages across 100 replicates.
28x6mm (600 x 600 DPI)



Single locus introgression and migration of a single individual on the 10-species tree. Estimated θ values are plotted on the tree, and posterior probabilities are in bold. The order of species is fixed for all trees. Species tree parameters are averages across 100 replicates. Symbols indicate the locations of the clades shown in panels.
68x30mm (600 x 600 DPI)



Distribution of gene tree coalescent times under a model with no migration (a) and with migration (b-e). Frequency histograms are shown for estimated gene trees (top; white) and simulated gene trees (bottom; gray). D = 0 (equation 1) represents the inflection point separating deep coalescences (positive values) and shallow coalescences (negative values). The x-axis is bounded by -1 for shallow coalescence and unbounded for deep coalescences (but truncated at +1.5 for clarity).
148x241mm (300 x 300 DPI)



Species tree distortions caused by gene flow that can result from coalescent methods that only model ILS. Dashed lines illustrate species tree compression, and the widening of branches illustrates species tree dilation in relation to the starting species tree.
45x11mm (600 x 600 DPI)