Syst. Biol. 67(6):1076–1090, 2018

DOI:10.1093/sysbio/syy034
Advance Access publication May 11, 2018

Assessing the Impacts of Positive Selection on Coalescent-Based Species Tree Estimation and Species Delimitation

Richard H. Adams, Drew R. Schield, Daren C. Card, and Todd A. Castoe*

Department of Biology, University of Texas at Arlington, 501 S. Nedderman Dr., Arlington, TX 76019, USA,
**Correspondence to be sent to: Department of Biology, University of Texas at Arlington, Arlington, TX 76010, USA;*
E-mail: todd.castoe@uta.edu.

Received 20 June 2017; reviews returned 3 May 2018; accepted 5 May 2018
Associate Editor: Richard Glor

Abstract.—The assumption of strictly neutral evolution is fundamental to the multispecies coalescent model and permits the derivation of gene tree distributions and coalescent times conditioned on a given species tree. In this study, we conduct computer simulations to explore the effects of violating this assumption in the form of species-specific positive selection when estimating species trees, species delimitations, and coalescent parameters under the model. We simulated data sets under an array of evolutionary scenarios that differ in both speciation parameters (i.e., divergence times, strength of selection) and experimental design (i.e., number of loci sampled) and incorporated species-specific positive selection occurring within branches of a species tree to identify the effects of selection on multispecies coalescent inferences. Our results highlight particular evolutionary scenarios and parameter combinations in which inferences may be more, or less, susceptible to the effects of positive selection. In some extreme cases, selection can decrease error in species delimitation and increase error in species tree estimation, yet these inferences appear to be largely robust to the effects of positive selection under many conditions likely to be encountered in empirical data sets. [Bayesian phylogenetics; natural selection; coalescent theory; speciation; population genetics; ecological speciation; systematics; species delimitation; phylogenomics.]

Multispecies coalescent models provide a valuable parameterization of the evolutionary processes that underlie neutral divergence between reproductively isolated lineages (Rannala and Yang 2003; Liu et al. 2009; Fujita et al. 2012; Edwards et al. 2016). Coalescent processes occurring within ancestral species can often yield genealogical discordance among loci as a result of incomplete lineage sorting (ILS). ILS is responsible for wide-spread phylogenetic heterogeneity observed across the Tree of Life, and when unaccounted for, ILS can have significant impacts on both species tree estimation and species delimitation (Heled and Drummond 2010; Huang et al. 2010; Camargo et al. 2012). Multispecies coalescent models account for ILS by parameterizing the width (population sizes) and depth (divergence times) of a given species tree, thereby providing a statistical framework for inferring evolutionary relationships despite genealogical conflicts (Degnan and Rosenberg 2009; Edwards 2009a; Yang and Rannala 2010).

Genetic variation, however, may be subject to a variety of evolutionary processes (in addition to neutral coalescence) occurring along branches of a species tree, several of which may violate assumptions of the multispecies coalescent model. For example, recent studies have documented the impacts of gene flow on coalescent species tree estimation and species delimitation in both simulated and empirical data sets (Zhang et al. 2011; Leaché et al. 2014; Burbrink and Guiher 2015). Under certain conditions (>0.1 migrant per generation), admixture occurring between lineages will bias species tree estimation and lead to false clustering of distantly related taxa, whereas species delimitation appears to be misled by the effects of gene flow only when migration rates are on the order of ~1 migrant per generation (Eckert and Carstens 2008; Zhang et al. 2011; Leaché et al. 2014). In contrast to the effects gene flow, simulation studies suggest that coalescent species tree estimation may be relatively robust to the effects of unrecognized recombination within loci (Lanier and Knowles 2012). The impacts of natural selection on species tree estimates and delimitation, however, are far less understood and have never been formally evaluated.

The multispecies coalescent model provides the probability distribution of coalescent times and gene tree topologies expected under neutral evolution on a given species tree. This assumption of neutrality is fundamental to all coalescent models used to infer population parameters and permits the mathematical treatment of the genealogical and mutational processes as independently modeled phenomena (Wakeley 2008). Natural selection, however, will favor the population trajectory of particular alleles such that the coalescent process of a selected locus will depend on its allelic state — this in turn may manipulate genealogical histories in complex and often unpredictable ways (Kaplan et al. 1989; Barton et al. 2004). Simulating coalescent genealogies with selection is often challenging, and only a single existing program allows the simulation of genetic data under evolutionary scenarios that incorporate both selection and complex demographic histories (Ewing and Hermisson 2010). Given the difficulties of modeling natural selection within a coalescent framework, no species tree estimation or species delimitation framework currently accounts for selection. Additionally, selection may further complicate phylogenetic inference by interacting with other aspects of the speciation, such as population sizes, divergence times, mutation rates, gene flow, and recombination

(Kaplan et al. 1989; Barton et al. 2004; Lanier and Knowles 2012).

The impacts of natural selection on species tree estimation and species delimitation have received little attention, and when discussed, opinions on the subject have varied widely among authors. Recent studies have disagreed over the relative importance of accounting for selection when conducting species tree estimation (Edwards et al. 2016; Springer and Gatesy 2016). At the gene tree level, particular patterns of selection are thought to have profound effects on phylogenetic inference when present (Edwards 2009b), and systematic errors have been documented in gene tree reconstruction in the presence of strong convergent selection (Stewart et al. 1987; Castoe et al. 2009). Given that selection is thought to occur in nature at a relatively small proportion of the nuclear genome, species tree estimation methods that analyze multiple unlinked loci are assumed to be relatively robust to the presence of selected loci—"misleading" signal generated by selected loci are assumed to be overwhelmed by the majority of neutral loci sampled (Edwards 2009b; Edwards et al. 2016). However, recent studies have suggested that both the direct and indirect effects of selection could be more pervasive across the genome than previously thought (Hahn 2008; McVicker et al. 2009; Scally et al. 2012; Corbett-Detig et al. 2015), and other studies have demonstrated that positive selection at even a small number of sites can indeed overwhelm gene tree inference (Castoe et al. 2009) and bias demographic estimates, such as reduced population sizes (Schrider et al. 2016).

Particular types or patterns of selection are thought to be less problematic for multispecies coalescent inferences (i.e., purifying selection), which may manifest primarily as reduced substitution rates and suppressed ILS at selected loci (Rannala and Yang 2003; Edwards 2009b; Zhu and Yang 2012; Edwards et al. 2016). Genes involved in speciation and adaptation are thought to provide better resolution of species histories (i.e., increased probability of monophyly), although it is unclear how this may directly translate to inferences under the multispecies coalescent model (Hey 1994; Ting et al. 2000; Rosenberg 2003). Recent studies have also shown that traits experiencing positive selection may provide better resolution of closely-related taxa when compared with neutral loci that exhibit minimal signal of reproductive isolation when species diverged recently (Solís-Lemus et al. 2015). Conversely, multiple studies have suggested that loci experiencing species-specific positive selection are not appropriate for coalescent species tree estimation and species delimitation analyses (Rannala and Yang 2003; Yang and Rannala 2010; Zhang et al. 2011; Springer and Gatesy 2016), primarily because they violate the core assumption of the model. Regardless, it is likely that large, multi-locus data sets may include some proportion of loci that have evolved under selection, and while it may be logical to filter away such loci from empirical data sets, the task of identifying targets of selection is not trivial. Accordingly, we see it as an urgent need to understand the potential consequences of positive selection on phylogenetic inference under models that assume strictly neutral evolution.

The question therefore remains: can species-specific positive selection influence coalescent species tree estimation and/or species delimitation? Herein, we address this question using coalescent simulations to evaluate the impacts of positive selection on multispecies coalescent inferences under a range of evolutionary scenarios and experimental conditions. We simulated genealogies and associated alignments both with and without selection occurring within a single taxon, and quantified differences between the simulated and inferred species models with respect to species tree topology, species delimitation, and demographic parameter estimates. Because these inferences are based on the assumption that gene trees are strictly a function of neutral coalescence occurring within species trees, we also characterized the effects of selection on gene tree distributions across our simulations. Our intentions were not to exhaustively explore all potential scenarios of selection and diversification histories, nor to evaluate the performance of different methods (Leaché and Rannala 2011; Sukumaran and Knowles 2017), but rather to provide a critical "first-step" perspective on the potential impacts of selection on coalescent inferences of evolutionary history. We evaluated the impacts of selection using the program Bayesian Phylogenetics and Phylogeography (BPP) (Yang and Rannala 2010) because it offers a general framework for both species tree estimation and species delimitation. Our analyses and interpretations were thus guided by three primary questions: 1) To what degree and in what direction can positive selection influence species tree estimation and delimitation? 2) What particular evolutionary scenarios and experimental conditions are most susceptible to the effects of selection? and 3) What practical concerns do positively-selected loci pose to analyses of empirical data sets?

Materials and Methods

Three-Species Simulation Model

We designed a multifactorial simulation experiment in which data were simulated under different evolutionary and experimental conditions that varied with respect to species divergence times, data set size (i.e., total number of loci), proportion of selected loci, selection strength, and sample size (i.e., number of haplotypes sampled per species). Our approach follows previous simulation-based studies of Bayesian species tree estimation and species delimitation methods, with several key differences (McCormack et al. 2009; Huang et al. 2010; Zhang et al. 2011; Lanier and Knowles 2012; Leaché et al. 2014). Briefly, our simulation framework consisted of 1) simulating genealogies (with and without selection) using the program MSMS (Ewing and Hermisson 2010), 2) simulating 1000 base DNA sequence alignments under the JC69 model

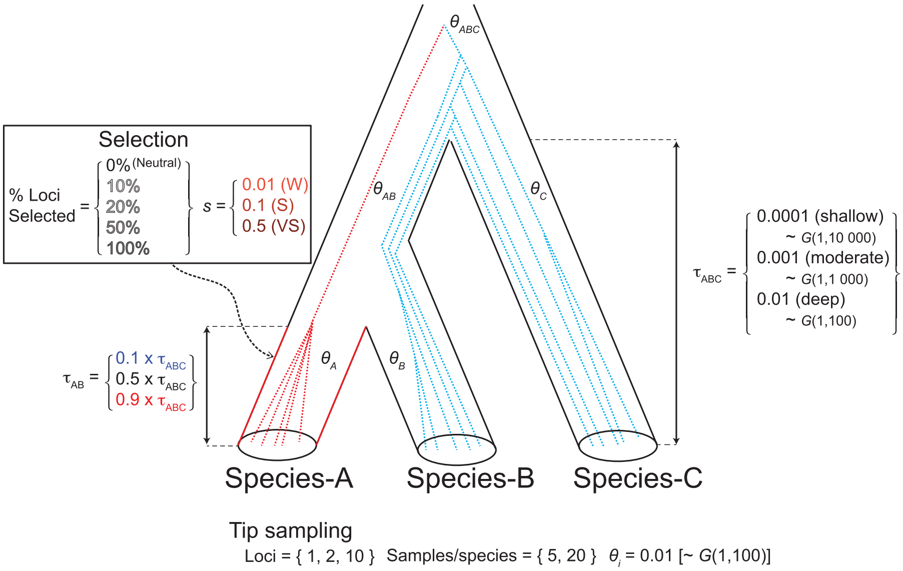
θ_ABC
Selection
% Loci Selected = { 0% (Neutral), 10%, 20%, 50%, 100% }
s = { 0.01 (W), 0.1 (S), 0.5 (VS) }
θ_AB
θ_C
θ_A
θ_B
τ_AB = { 0.1 x τ_ABC, 0.5 x τ_ABC, 0.9 x τ_ABC }
τ_ABC = { 0.0001 (shallow) ~ G(1,10 000), 0.001 (moderate) ~ G(1,1 000), 0.01 (deep) ~ G(1,100) }
Species-A
Species-B
Species-C
Tip sampling
Loci = { 1, 2, 10 } Samples/species = { 5, 20 } θ_i = 0.01 [~ G(1,100)]

FIGURE 1. Species tree and experimental parameters used for simulating genealogies under the multispecies coalescent model both with and without selection. Dotted lines within the species tree represent an example genealogy in which a selective sweep has occurred in Species-A lineages immediately after speciation, such that all Species-B and Species-C lineages coalesce in the root Species-ABC before reaching a common ancestor with any Species-A lineages.

(Jukes and Cantor 1969) on simulated genealogies, 3) conducting Bayesian species tree estimation and species delimitation using BPP (Yang and Rannala 2010) for each simulated data set, and 4) quantifying differences between the true species model (upon which simulations were made) and posterior models inferred via Markov Chain Monte Carlo (MCMC) sampling. We evaluated the effects of selection on inferences of a three-species model with parameters described by the multispecies coalescent: population size parameters ($\theta_A, \theta_B, \theta_C, \theta_{AB}, \theta_{ABC}$), divergence times ($\tau_{AB}, \tau_{ABC}$), and topology ((Species-A, Species-B), Species-C) (Fig. 1). We choose a three-species model so that we could tractably test a wide-range of experimental conditions, parameter values and combinations, and for comparative purposes with recent similar studies using a three-species model to study the effects of gene flow (Zhang et al. 2011).

We hypothesized that the impacts of positive selection would be most relevant when species are relatively closely related and population sizes are large, and thus we tailored our simulations to variations of these scenarios, which also represent more challenging problems for species delimitation and species tree estimation (Maddison and Knowles 2006; Leaché and Rannala 2011; Zhang et al. 2011). For all simulation experiments, we set a constant value of $\theta = 0.01$ for all ancestral and extant species in the model ($\theta_A = \theta_B = \theta_C = \theta_{AB} = \theta_{ABC} = 4N\mu = 0.01$) and a diploid population size $N_e =$ 100,000 individuals, which corresponds to a mutation rate $\mu = 2.5 \times 10^{-8}$ substitutions per site per generation. We chose this value of θ because it falls within the range of empirical estimates of θ (0.0005–0.02) for many animal and plant species (Zhang and Hewitt 2003), and the mutation rate of 2.5 × 10^{-8} has been suggested for a number of taxa, including humans (Nachman and Crowell 2000). This θ value is therefore likely representative of many species and has also been used in previous simulation-based studies (Zhang et al. 2011). For our simulations, we tested a total of nine different three-taxon models that differ in relative divergence times (τ_{AB}, τ_{ABC}; Fig. 1). We used three different simulation models that differed by three orders of magnitude for the root node depth of the species tree (τ_{ABC}): shallow ($\tau_{ABC} = 0.0001$), moderate-depth ($\tau_{ABC} = 0.001$), and deep species tree models ($\tau_{ABC} = 0.01$; Fig. 1). For each of these three different models of species tree depth, we also tested three values for the time at which Species-A and Species-B diverged from one another (τ_{AB}): recent ($\tau_{AB} = \tau_{ABC} \times 0.1$), medium ($\tau_{AB} = \tau_{ABC} \times 0.5$), and ancient divergence ($\tau_{AB} = \tau_{ABC} \times 0.9$; Fig. 1). This has the effect of shortening or elongating the internode distance (i.e., length of the ancestral Species-AB branch) in relation to the species tree height (τ_{AB}); parameters that have been shown to significantly impact both species tree estimation and delimitation (Maddison and Knowles 2006; Leaché and Rannala 2011; Zhang et al. 2011).

Simulating Selection on Multispecies Coalescent Models

We used the program MSMS (Ewing and Hermisson 2010) to simulate both neutral and selected genealogies under each three-species model. Selection coefficients are specified in units of $2N_es$ and $waa = 1 + \frac{saa}{2Ne}$, where N_e is the diploid population size, w_{aa} is the Malthusian fitness for the aa genotype, and s_{aa} is the selection coefficient against the homozygous aa genotype. For

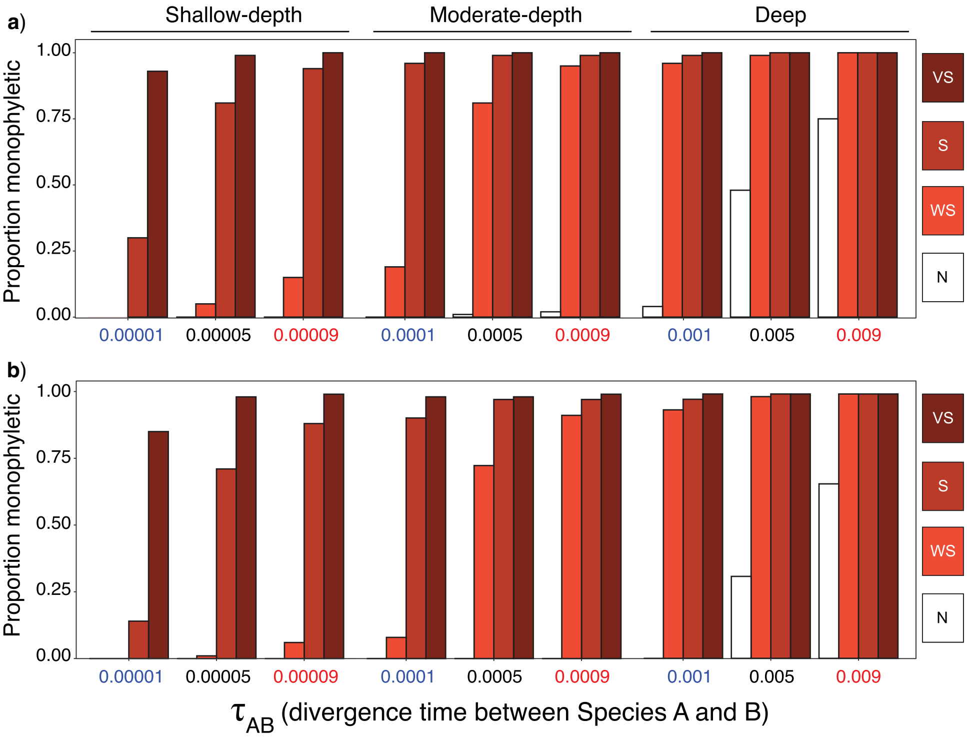

FIGURE 2. The impact of species-specific positive selection on gene tree distributions and the probability of monophyly. Barplots indicate the percentage of simulated genealogies with monophyletic relationships for all Species-A lineages (i.e., all Species-A lineages reach a common ancestor before coalescing with any outgroup lineages, see example genealogy shown in Fig. 1). Results are shown from left to right for the shallow ($\tau_{ABC} = 0.0001$), moderate-depth ($\tau_{ABC} = 0.001$), and deep ($\tau_{ABC} = 0.01$) species tree models and for each respective Species-AB divergence: $\tau_{ABC} \times 0.10$, $\tau_{ABC} \times 0.50$, and $\tau_{ABC} \times 0.90$. For each simulation model and associated parameters, we simulated 10^4 genealogies under neutral evolution ("N"), as well as weak ("W", s = 0.01), strong ("S", s = 0.10), and very strong ("VS", s = 0.50) selection coefficients, represented by a gradient from light to dark red for increasing selection strength. Results are shown for the simulations with 5 (a) and 10 (b) haplotypes sampled per species.

example, with a diploid population size of $N_e = 100{,}000$ and $w_{aa} = 0.90$ (aa homozygotes produce 10% fewer offspring), we would specify $s_{aa} = -20{,}000$ to simulate data in which strong positive selection is driving the A allele toward fixation with complete dominance. Our goal was to tractably evaluate the effects of selection across a variety of conditions using three different selection strengths for each species model and parameter combination: weak ($s_{aa} = -2000$), strong ($s_{aa} = -20{,}000$), and very strong ($s_{aa} = -100{,}000$) selection pressure against the recessive genotype within a single species (Species-A). For brevity, we refer to these three selection strengths in terms of the absolute difference in fitness between the homozygous AA and aa genotypes $\left(\mathrm{s} = \frac{1-\mathrm{Waa}}{1}\right)$: "weak" ("W", s = 0.01), "strong" ("S", s = 0.10), and "very strong" ("VS", s = 0.50) selection. We also specified the forward and backward mutation rate at the selected site equal to 2.5×10^{-8}. We set the starting time of selection to occur immediately after the divergence of Species-A and Species-B (τ_{AB}), and set the starting allele frequency to 0.000005; these scenarios effectively represent a novel, beneficial mutation within a single individual within Species-A that arises immediately after its ancestral divergence from Species-B. We tested different sampling schemes (number of total loci, number of selected loci, and number of haplotypes sampled per species) to evaluate how different experimental designs may be more or less susceptible to the effects of selection (Fig. 1). We used three different data set sizes (1-locus, 2-loci, and 10-loci) and varied the proportion of selected loci within these data sets: 0% (neutral), 10%, 20%, 50%, and 100% (Fig. 1). We also explored how two different sample sizes interacted with the amount of selection present in the data sets (5 or 20 haplotypes sampled per species; Fig. 1).

In addition to our BPP analyses, we simulated 10^4 genealogies and alignments for each species tree model (9 total divergence models) and experimental condition (neutral and 3 selection coefficients, 2 sample sizes: $10^4 \times 9 \times 4 \times 2 = 720{,}000$) that were used to quantify the effects of selection on gene tree distributions and to provide a population genetic perspective to our findings (Figs. 2 and 3). Based on these data, we quantified the percentage of gene trees that exhibit complete monophyly for all Species-A lineages (i.e., the example genealogy shown in Fig. 1) for each set of simulated genealogies. Next, we simulated alignments

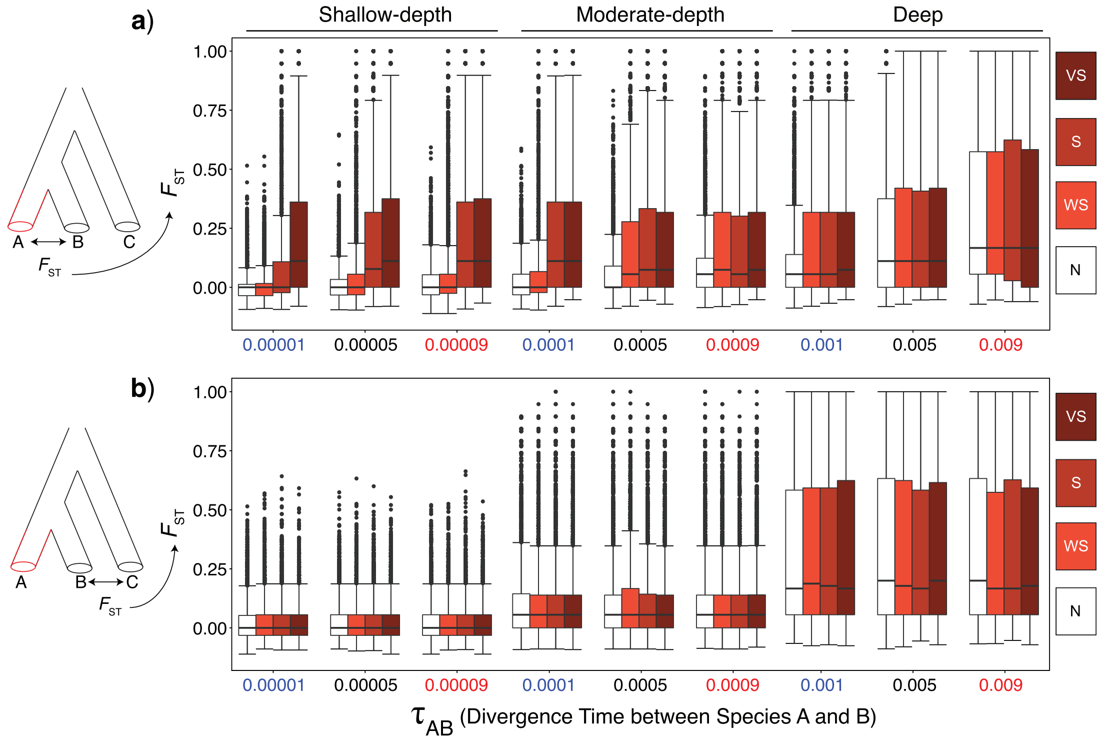

FIGURE 3. The effects of species-specific positive selection on measures of population differentiation. Boxplots represent the distribution of F_{ST} estimates between Species-A and Species-B (a) and Species-B and Species-C (b) across 10^4 simulated alignments with 20 haplotypes sampled per species. Results are shown from left to right for the shallow ($\tau_{ABC}=0.0001$), moderate-depth ($\tau_{ABC}=0.001$), and deep ($\tau_{ABC}=0.01$) species tree models and for each respective Species-AB divergence (τ_{AB}): $\tau_{ABC}\times 0.10$, $\tau_{ABC}\times 0.50$, and $\tau_{ABC}\times 0.90$. For each simulation model and associated parameters, we simulated 10^4 genealogies under neutral evolution ("N"), as well as weak ("W", s = 0.01), strong ("S", s = 0.10), and very strong ("VS", s = 0.50) selection coefficients, represented by a gradient from light to dark red for increasing selection strength.

and calculated F_{ST} by sampling a single SNP from each of simulated locus to obtain a distribution of F_{ST} for each simulation condition. We also conducted two pairwise lineage comparisons: Species-A versus Species-B and Species-B versus Species-C using scripts provided in the R package *GppFst* (Adams et al. 2016).

Simulation of Sequence Alignments

DNA sequence data were simulated using the program Seq-Gen (Rambaut and Grassly 1997) for each genealogy simulated by MSMS. We evolved 1000 bp alignments under the JC69 model (Jukes and Cantor 1969) for all simulated data sets. For the genealogies that experience positive selection, our approach effectively models a 1000 bp sequence that is genetically linked to a single positively-selected site (i.e., the selected site is not included in the alignment).

Running the rjMCMC Algorithms

We simulated 200 replicate data sets for each parameter and sampling combination. We conducted Bayesian species tree estimation (algorithm 01), unguided species delimitation (algorithm 11), and parameter estimation (algorithm 00) using the program BPP (Yang and Rannala 2010). For all BPP analyses, we used gamma prior distributions with expectations at the true simulated value for the root node depth (τ_{ABC}) and population parameters θ (Fig. 1), and we set the species model prior to the default "Prior 1" setting, which assigns equal probabilities on the three rooted topologies; similar prior settings have been used in other recent simulation studies (Yang and Rannala 2010; Zhang et al. 2011). We used the true simulated species topology ((A, B), C) as the starting topology for all analyses. We ran the MCMC algorithms implemented in BPP for a total of 110,000 iterations (sampling every 10) and designated the first 10,000 iterations to be discarded as burn-in. We calculated the mean and standard deviation of posterior probabilities of the three possible rooted topologies and of species delimitation hypotheses across all 200 replicates. We used the mean value of the posterior distribution for each of the 200 replicates for θ and τ parameters and plotted the total mean and standard deviation for these estimates under each set of experimental parameters. Our entire simulation study comprised 86,400 uniquely simulated data sets, which were analyzed independently for species tree estimation, species delimitation and parameter estimation for a total of 86,400 × 3 = 259,200 BPP analyses (Fig. 1).

Results

The Effect of Positive Selection on Gene Tree Distributions and Population Genetic Statistics

We find that species-specific positive selection can bias gene trees toward topologies in which all Species-A lineages coalesce before coalescing with Species-B or Species-C lineages (Fig. 2). In other words, genealogies simulated under selection show an increased propensity for Species-A monophyly when compared with neutral loci (i.e., the example genealogy shown in Fig. 1). As would be predicted, our simulations demonstrate that the degree to which selection influences lineage sorting is a function of the selection coefficient and divergence times (both τ_{ABC} and τ_{AB}). This effect scales with the strength of selection, and in all cases we found that >85% of genealogies exhibited monophyly of Species-A lineages even when species diverged very recently. We also observed a strong inverse relationship between tree depth (τ_{ABC}) and the strength of selection required to influence genealogical distributions. For example, even weak selection can result in major shifts in gene tree distributions in our deep species tree models, whereas only stronger selection coefficients are able to substantially influence the sorting of Species-A lineages in our shallow species simulations (Fig. 2). For the shallow species simulations ($\tau_{ABC} = 0.0001$, $\tau_{AB} = 0.00001$) with five samples per species, 0% of genealogies are completely sorted within Species-A under both neutral evolution and weak selection (s = 0.01), while 29.54% and 93.20% are completely sorted with strong (s = 0.10) and very strong selection (s = 0.50), respectively (Fig. 2a). We observed a similar trend between weaker selection and the relative divergence time between Species-A and Species-B (τ_{AB}).

We find that selection increases estimates of F_{ST} compared with neutral loci, yielding patterns of differentiation that are incorrectly interpreted as greater lineage divergence when compared with neutral loci (Fig. 3). Importantly, F_{ST} between Species-A and Species-B often exceeded that between Species-B and the more distantly related outgroup Species-C, when loci are under selection in Species-A (Fig. 3a vs. b). For example, although the divergence time between Species-A and Species-B ($\tau_{AB} =$ 0.00001) was two orders of magnitude lower than the divergence with Species-C ($\tau_{ABC} = 0.001$), average F_{ST} between Species-A and Species-B under very strong selection is over twice (0.227) that measured between Species-B and Species-C (0.093; Fig. 3a vs. b).

The Effects of Selection on Estimates of Species Divergence Time and Population Size Parameters

Evaluation of the effects of selection on four parameters (τ_{ABC}, τ_{AB}, θ_A, θ_B) confirm that selection can bias parameter estimates toward larger estimates of species divergence times (τ_{ABC}, τ_{AB}) and smaller estimates of population size parameters for the species under selection (θ_A) compared with the true simulated values and neutral estimates (Fig. 4 and Supplementary Figs. S1 and S2 available on Dryad at https://doi.org/10.5061/dryad.5v3b5). We also observe a slight increase in population size parameter estimates of the sister taxon, Species-B (θ_B) in some analyses (Supplementary Fig. S1 available on Dryad). Biases in parameter estimates appear to be largely a function of the proportion of loci under selection, and the strength of selection, and in many scenarios, increasing the number of individuals sampled per taxa also increases the severity of bias. Because the relative severity of the impacts of selection on these parameter estimates depends largely on the species tree depth (τ_{ABC}) and Species-AB divergence time (τ_{AB}), we discuss our results separately in the context of each of the three-species depth models below.

Shallow species trees.—Our simulation analyses indicate that selection can bias estimates of divergence times (τ_{ABC}, τ_{AB}) and population size parameters (θ_A, θ_B) on shallow species trees ($\tau_{ABC} = 0.0001$; Fig. 4 and Supplementary Fig. S1 available on Dryad). Selection can bias estimates of τ_{ABC} and τ_{AB} toward larger values, meaning that data sets including loci under selection lead to incorrectly older estimates of speciation times when compared with neutral data sets. While strong selection at multiple loci can substantially bias parameters inferred from 2- and 10-locus data sets, θ and τ estimates appear robust to the presence of weak selection in many cases (Fig. 4). We also find that selection can bias estimates of the population size parameters θ_A and θ_B under certain conditions (Fig. 4a–c and Supplementary Fig. S1d–f available on Dryad). Under the most extreme conditions explored in which 100% of loci in 10-locus data sets evolved under very strong selection, θ_A is decreased by 97.9% (0.00021; Fig. 4c). Using the simulated mutation rate ($\mu = 2.5 \times 10^{-8}$), this corresponds to an N_e estimate of only 2100 individuals, while the true population size simulated was 100,000. These biases are substantially reduced under more realistic conditions, as when only 10% of loci are under selection (Fig. 4c, light gray).

Moderate-depth species trees.—Positive selection can also bias parameter estimates under our models of moderate species trees ($\tau_{ABC} =$ 0.001, Fig. 4), but these biases are less pronounced compared with our shallow species tree analyses. Estimates of τ_{ABC}, τ_{AB}, and θ_B are often inflated as the number and strength of selection increases, while θ_A estimates are biased toward smaller values (Fig. 4 and Supplementary Fig. S1 available on Dryad). These effects are most prominent when Species-A and Species-B diverged recently under this moderate-depth species tree model (i.e., $\tau_{AB} = 0.0001$; Fig. 4, blue lines) and are less pronounced with greater relative divergence. Parameter estimates appear relatively robust to larger

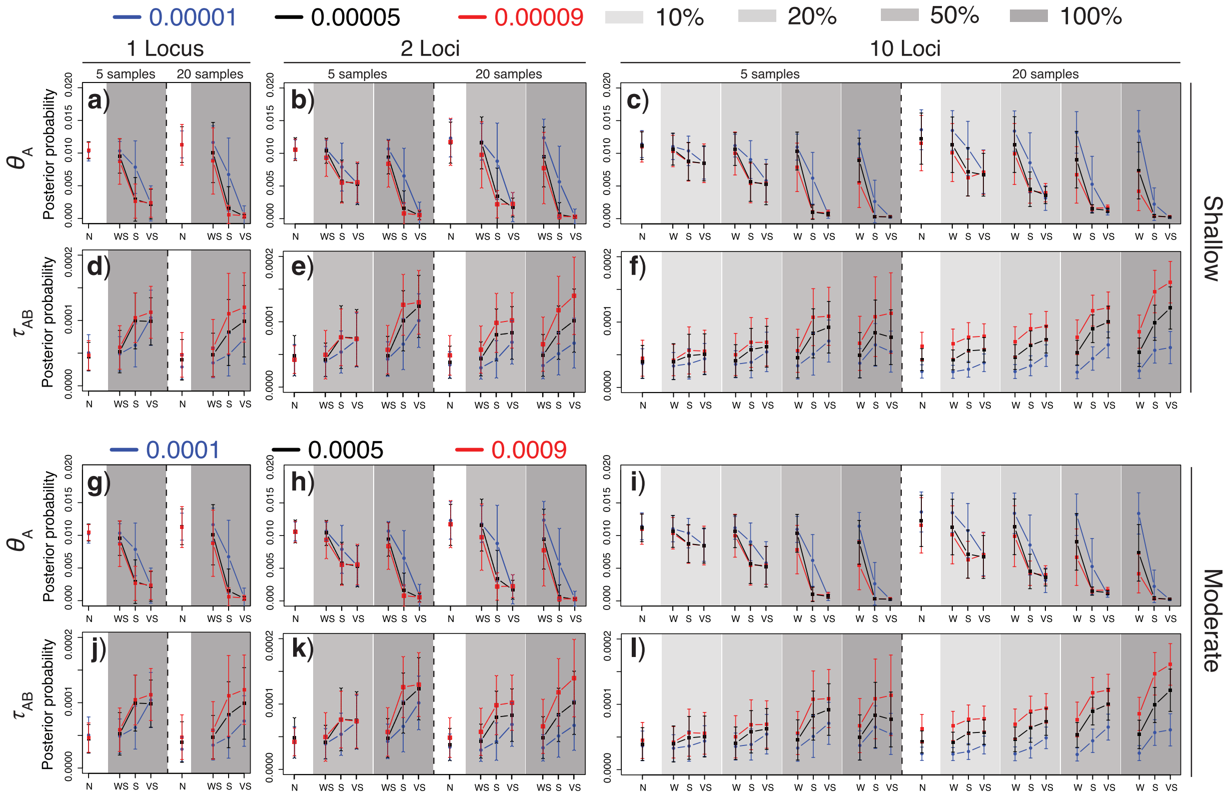

FIGURE 4. Selection can decrease estimates of θ_A and inflate divergence time estimates (τ_{AB}) for both the shallow (top) and moderate-depth (bottom) species model. Results are shown for θ_A (a–c and g–i) and τ_{AB} (d–f and j–l) for simulated data sets consisting of 1-locus (a, d, g, j), 2-loci (b, e, h, k), and 10-loci (c, f, i, l). The mean (points) and standard deviation (error bars) of parameter estimates based on 200 replicates are shown for three different Species-AB divergence times: $\tau_{AB}=0.00001$, 0.00005, and 0.00009 for the shallow species model (top) and $\tau_{AB}=0.0001$, 0.0005, and 0.0009 for the moderate-depth species model (bottom). Each panel is split into two subpanels representing 5 (left of dotted line) or 20 (right of dotted line) haplotypes sampled per species. A color gradient ranging from white to dark gray is used to indicate the different percentages of loci under selection: 0% (neutral, white), 10%, 20%, 50%, and 100% (dark gray). For simulations with selection, we varied the strength of selection: weak ("W", s = 0.01), strong ("S", s = 0.10), and very strong ("VS", s = 0.5) selection coefficients.

data sets that include only a small proportion of loci (10–20%) that have evolved under even strong selection, as well as weak selection even at 100% of loci (Fig. 4 and Supplementary Fig. S1 available on Dryad; light gray shading).

Deep species trees.—Our results suggest that selection has little influence over parameter estimates for deep species tree models ($\tau_{ABC}=0.01$, Supplementary Fig. S2 available on Dryad), except when $\tau_{AB}=0.001$ and only θ_A appears to be susceptible to strong selection (Supplementary Fig. S2a–c, blue lines available on Dryad). In all other scenarios, estimates of τ_{ABC}, τ_{AB}, θ_A, and θ_B under scenarios of selection are nearly equivalent to neutral inferences, regardless of the strength or prevalence of selection (i.e., proportion of selected loci), and regardless of sample sizes (5 vs. 20). Even under the most extreme scenarios of positive selection in 10-locus data sets (100% of loci under very strong selection), the parameter estimates are nearly identical to neutral inferences when $\tau_{AB}\geqslant 0.005$ (Supplementary Fig. S2, black and red lines available on Dryad).

The Effects of Selection on Species Delimitation

We evaluated the effects of positive selection on Bayesian coalescent species delimitation by comparing the average posterior probability (across the 200 replicates) of a species model consisting of three-species (P_3), the posterior probabilities of Species-A (P_A) and Species-B (P_B), and the posterior support for an incorrect inference of Species-B and Species-C being a single species (Species-BC; P_{BC}). Herein, increasing P_3, P_A, and P_B due to the presence of selection in the data represents increased confidence in the true simulation model. Conversely, an increase in P_{BC} due to selection represents a statistical bias toward an incorrect inference, because Species-B and Species-C were simulated as true, genetically isolated species. In general, we find that the effects of selection on posterior probabilities of species hypotheses are strongest in our shallow simulation models and when Species-AB diverged relatively recently (i.e., shorter τ_{AB}). Additionally, our simulation analyses indicate that the effects of selection on posterior probabilities increase with larger sample sizes (i.e., 5 vs. 20). In most cases, weak selection appears

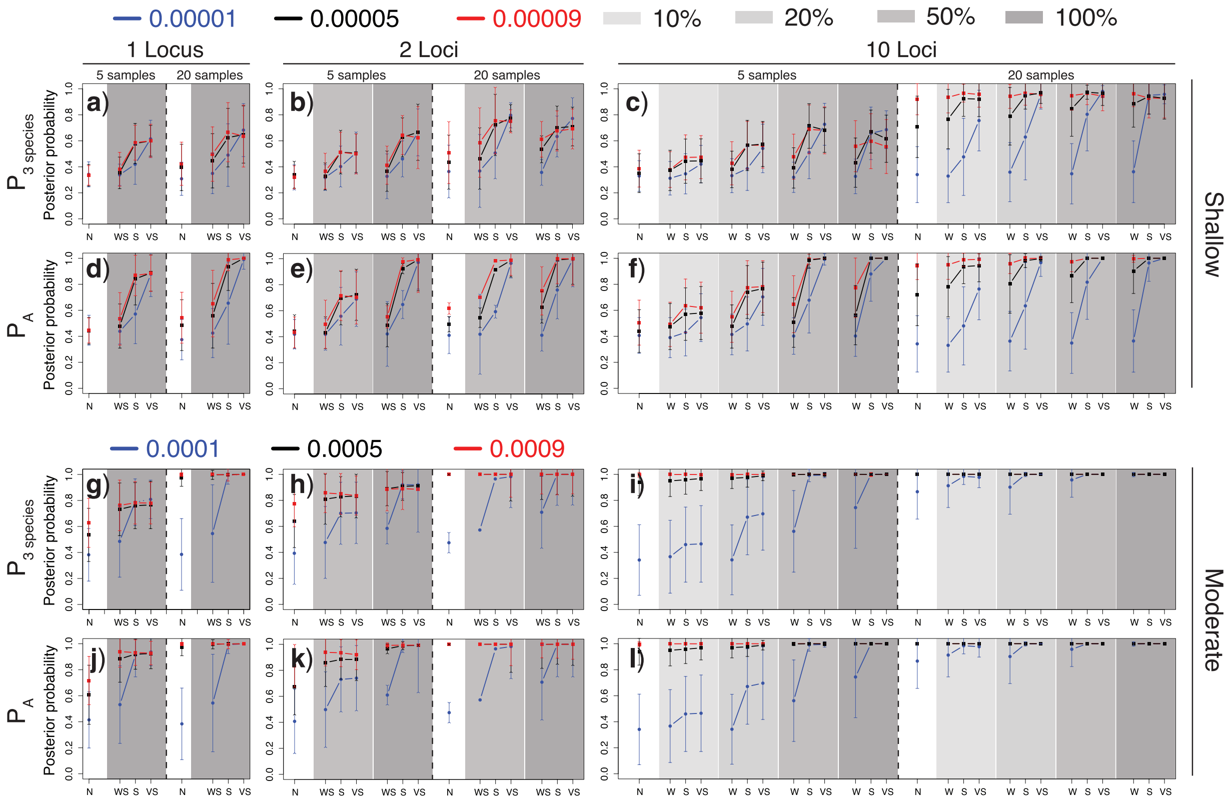

FIGURE 5. Selection can increase posterior probabilities of species hypotheses for the shallow (top) and moderate-depth (bottom) species tree model. Results are shown for the probability of three species (P_3; a–c and g–i) and the probability of Species-A (P_A; d–f and j–l).

to have minimal influence over posterior probabilities (Fig. 5, light gray), and we often find little difference between estimates obtained from strictly neutral data sets and those inferred from data sets comprising fewer loci under selection (i.e., 10–20%), but not always.

Shallow species trees.—We find that positive selection can influence Bayesian coalescent species delimitation on shallow species trees (τ_{ABC} = 0.0001), particularly when multiple loci have evolved under strong selection (s = 0.10, 0.50; Fig. 5 and Supplementary Fig. S3 available on Dryad). The effects of selection on posterior probabilities of species hypotheses increases with the strength of selection and the number of selected loci included in the analyses. For example, selection inflates estimates of P_3, P_A, P_B, to varying degrees depending on the percentage of loci under selection and the particular selection coefficient. We also find that the relative divergence times between Species-A and Species-B (τ_{AB}), and larger sample sizes, have substantial synergistic effects that determine the degree that selection influences posterior probabilities, which appear most susceptible to the effects of selection when Species-A and Species-B are more closely related (Fig. 5, blue lines) and 20 individuals are sampled (Fig. 5c and f).

We find that selection increased posterior probabilities of single-locus based species delimitation when a single neutral locus did not appear to provide strong resolution of species (Fig. 5a and d). We also identified similar trends in species probabilities as the proportion of loci under strong selection increased for the 2- and 10-locus analyses. When 10% of loci are under selection in 10-locus data sets (i.e., a single selected locus) the effects of selection on P_3, P_A, and P_B are relatively weak, but are greater when selection is strong, 20 haplotypes are sampled per species, and Species-A and Species-B diverged recently (τ_{AB}=0.00001, Fig. 5c and f and Supplementary Fig. S3a–c available on Dryad). We also find a small, but measurable increase in P_{BC} in several analyses (Supplementary Fig. S3d–f available on Dryad). In the most extreme scenarios where all 10 loci evolved under strong selection and 5 haplotypes were sampled per species, P_{BC} (0.314) is over four times that of 10 neutral loci (P_{BC}=0.077). However, this bias appears largely restricted to scenarios of strong selection, and is reduced under even slightly more realistic conditions (Supplementary Fig. S3, light gray vs. dark gray available on Dryad).

Moderate-depth species trees.—Our results indicate that selection can also influence species delimitation on moderate-depth species trees under some conditions (τ_{ABC}=0.001; Fig. 5), but far less than we observed with shallow species model. In other words, selection has less influence over estimates of more distantly related taxa when compared with more recently-diverged species

(Fig. 5, top vs. bottom panels). Similar to our analyses of the shallow simulation models, selection yields higher P_3, P_A, P_B, and P_{BC} estimates compared with neutral locus data sets. These effects are largely limited to scenarios in which Species-A and Species-B are recently diverged ($\tau_{AB} = 0.0001$; Fig. 5, blue line), and are far less pronounced or unobserved when τ_{AB} is older (Fig. 5, black and red lines). In general, moderate-depth species tree simulations have far less sensitivity to the varying strengths of selection and reduced sensitivity to the number of haplotypes sampled.

Deep species trees.—As with our moderate-depth simulation analyses, we find that selection only impacts species delimitation on the deep species model ($\tau_{ABC} = 0.01$) when Species-A and Species-B diverge relatively recently, and only in 1- and 2-locus analyses ($\tau_{AB} = 0.001$, blue line; Supplementary Fig. S4 available on Dryad). In these scenarios, we find that even weak selection can increase P_3, P_A, and P_B when compared with neutral loci. Outside of these special conditions, we otherwise find that the posterior probabilities of the true simulation model (P_3, P_A, and P_B) approach 1.0 and $P_{BC} = 0.0$ under nearly all other simulated scenarios, regardless of the selection strength, the proportion of selected loci, and the number of individuals sampled (Supplementary Fig. S4 available on Dryad).

The Effects of Selection on Species Tree Estimation

We quantified the effects of species-specific positive selection on coalescent species tree estimation by measuring the posterior probability of two competing rooted topologies: the true species topology ((Species-A, Species-B), Species-C) indicated by P_{ABC}, and an incorrect topology ((Species-B, Species-C), Species-A) indicated by P_{BCA}. We find that decreases in P_{ABC} always coincide with increases in P_{BCA}, and that the probability of the third possible rooted topology ((Species-A, Species-C), Species-B) is largely unaffected by selection and remains consistently low (results not shown). If selection increases P_{ABC} compared with neutral conditions, then selection reduces error in species tree estimation. Conversely, if selection increases $P_{BCA,}$ selection increases error in species tree estimation and biases inferences toward the incorrect rooted topology (i.e., selection is positively misleading). In general, we find that positive selection can influence species tree estimation particularly when species are more closely related, the ancestral Species-AB branch is shorter, more individuals are sampled per taxa, and strong selection is present at multiple loci.

Shallow species trees.—Our simulations suggest that species-specific positive selection can influence species topology estimates in the context of our shallow species model ($\tau_{ABC} = 0.0001$), particularly when selection is strong, the proportion of selected loci is high, and more than a single selected locus is sampled (Fig. 6a and Supplementary Fig. S5a available on Dryad). Additionally, the effects of selection on posterior probabilities of species topologies increase with larger sample sizes (i.e., 5 vs. 20) and when Species-A and Species-B diverged more anciently (i.e., larger τ_{AB}). Under specific scenarios of selection, our simulations demonstrate that selection can mislead species tree estimation by increasing P_{BCA} and simultaneously decreasing P_{ABC} to varying degrees as a function of experimental parameters (i.e., sample sizes) and evolutionary conditions (i.e., selection coefficient). For example, positive selection at a single locus slightly increases the probability of the wrong species topology from $P_{BCA} = 0.324$ under neutral conditions to 0.358, and 0.447 for strong and very strong selection, respectively, when five haplotypes are sampled and $\tau_{AB} = 0.00001$ (Supplementary Fig. S5a available on Dryad). When sampling is increased to 20, P_{BCA} is further increased to over 2.5 × (0.608) that of neutral inferences (0.242) for data sets consisting of a single locus under very strong selection. Tracking increases in P_{BCA} in the presence of selection, the posterior probability of the true tree (P_{ABC}) decreases from 0.523 under neutral estimates to 0.493, 0.395, and 0.207 under weak, strong, and very strong selection coefficients, respectively ($\tau_{AB} = 0.00001$; Supplementary Fig. S5a available on Dryad).

We observed similar effects of selection for species trees inferred from 2-locus data sets (Supplementary Fig. S5a available on Dryad). As expected, the statistical bias introduced by selection increases as the number of selected loci, the strength of selection, number of individuals sampled, and τ_{AB} increases. For example, when $\tau_{AB} = 0.00009$, P_{BCA} (0.752) is over twice that inferred from strictly neutral loci and P_{ABC} is less than half that of neutral inferences (0.128) when both loci are under strong selection (neutral $P_{BCA} = 0.332$, $P_{ABC} = 0.351$). Similarly, species trees estimated from 10-locus data sets may also be biased toward the wrong topology as the number of loci under selection, strength of selection, number of individuals, and τ_{AB} increase. In the most extreme conditions in which all 10 loci are under very strong selection, 20 haplotypes are sampled per species, and $\tau_{AB} = 0.00001$, P_{BCA} increases to over 60-fold (0.629) that inferred from 10 neutral loci (0.020) and P_{ABC} decreases to 0.224 (neutral $P_{ABC} = 0.960$; Fig. 6a). When $\tau_{AB} = 0.00009$, P_{BCA} increases to 0.472, 0.906, and 0.962, while P_{ABC} decreases to 0.273, 0.048, and 0.019 under weak, strong, and very strong selection, respectively (neutral $P_{BCA} = 0.308$ and $P_{ABC} = 0.4000$). However, these biases are substantially reduced under even slightly more realistic conditions, as observed when 10% of loci experienced positive selection (i.e., a single locus; Fig. 6a).

Moderate-depth species trees.—Selected loci can also bias species tree estimation toward the incorrect topology for moderate-depth species trees ($\tau_{ABC} = 0.001$; Fig. 6b and Supplementary S5b available on Dryad). As with shallow species models, the effects of selection on P_{BCA} and P_{ABC} increase with the strength of selection, number

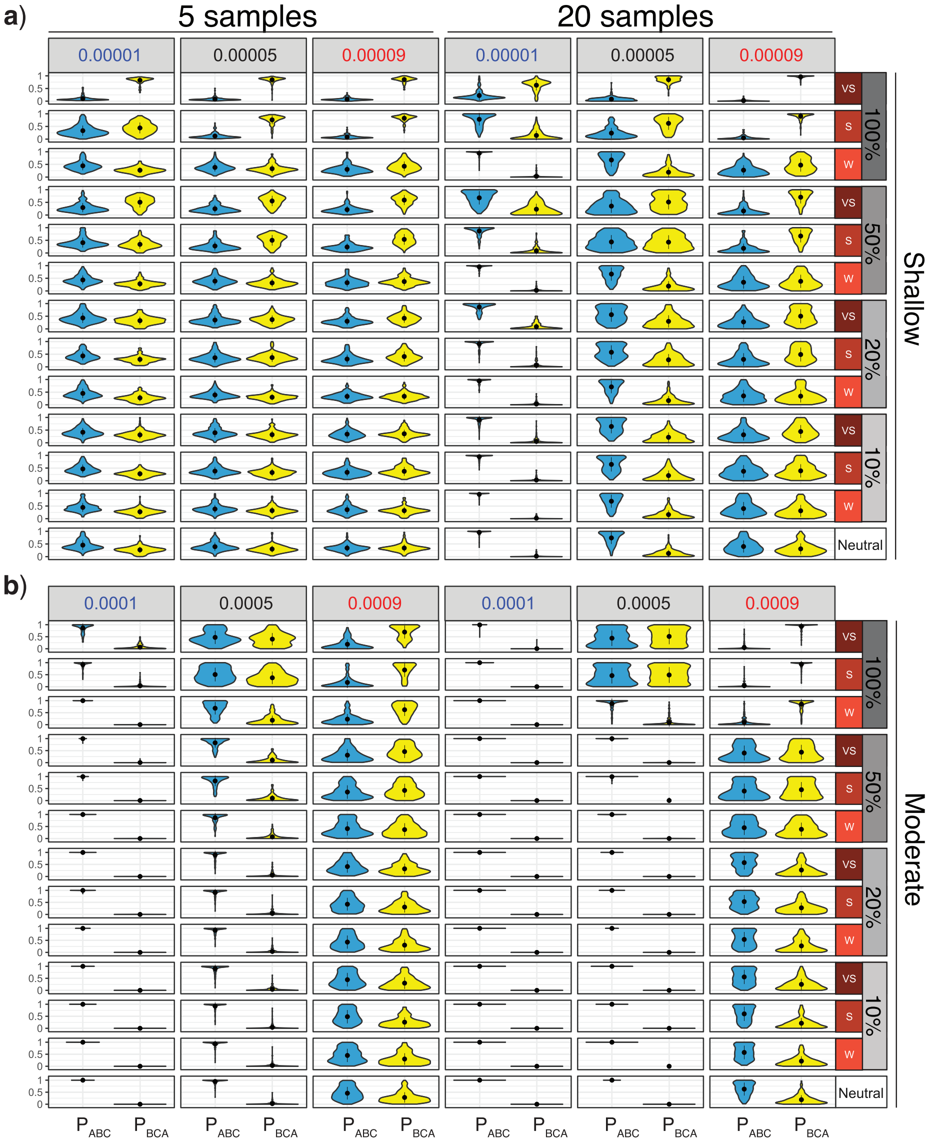

FIGURE 6. Species-specific positive selection can bias species tree estimates of shallow (a) and moderate-depth (b) species models. Violin plots show the distribution of posterior probabilities of the correct rooted species topology (P_{ABC}) and incorrect topology (P_{BCA}) across 200 replicates (mean shown in black) for data sets consisting of 1-locus (bottom), 2-loci (middle), and 10-loci (top) that were simulated with either 5 (left) or 20 samples per species (right) under three different Species-AB divergence times (from left to right): $\tau_{AB} = 0.00001$, 0.00005, and 0.00009. A gradient ranging from white to dark gray shading indicates the different percentages of loci under selection: 0% (neutral, white), 10%, 20%, 50%, and 100% (dark gray). For simulations with selection, we varied the strength of selection: weak ("W", s = 0.01), strong ("S", s = 0.10), and very strong ("VS", s = 0.5) selection coefficients.

of selected loci, number of individuals sampled per taxa, and τ_{AB}. Generally, we find that biases introduced by selection are less pronounced on moderate-depth species trees when compared with shallow species trees (Fig. 6a vs. b). We find that in all cases, reducing the number of individuals sampled also reduces the statistical biases observed in the analysis of selected loci.

In our single-locus estimates of species topologies, P_{BCA} increases to 0.533, 0.557, and 0.583, while P_{ABC} decreases to 0.236, 0.239, and 0.212 under weak, strong, and very strong selection coefficients, respectively, when $\tau_{AB} = 0.0009$ and 20 haplotypes are sampled per taxa (neutral $P_{BCA} = 0.299$ and $P_{ABC} = 0.404$). Analyses of the more recently diverged simulations ($\tau_{AB} = 0.0001, 0.0005$) show similar trends and the overall effects of selection on P_{BCA} and P_{ABC} are reduced when only five haplotypes are sampled per species (Supplementary Fig. S5b available on Dryad). We observed similar trends for 2-locus data sets, as biases introduced by selection are also most prominent when $\tau_{AB} = 0.0009$ and 20 haplotypes are sampled per taxa. However, when only one of the two loci are under selection, P_{BCA} and P_{ABC} are closer to those based on neutral inferences (i.e., inferences are less biased with the addition of even a single neutral locus). For 10-locus data sets, our results suggest that species tree estimates can be strongly biased toward the wrong topology in the presence of weak selection at all 10 loci: P_{BCA} increases to 0.856, 0.934, and 0.932, while P_{ABC} decreases to 0.096, 0.042, 0.049, for weak, strong, and very strong selection coefficients, respectively, when $\tau_{AB} = 0.0009$ and 20 haplotypes are sampled per species (neutral $P_{BCA} = 0.189$, $P_{ABC} = 0.628$, Fig. 6b). Similar biases are observed in our simulated data sets when $\tau_{AB} = 0.0005$, but are less pronounced. We find that species tree probabilities are largely unaffected by selection when $\tau_{AB} = 0.0001$, even when all 10 loci evolved under very strong selection (Fig. 6b).

Deep trees.—Our simulation analyses indicate that selection does not appear to measurably influence species tree estimation on deep species trees (Supplementary Fig. S6 available on Dryad). Regardless of the number of selected loci, selection strength, sample sizes (5 vs. 20), and Species-AB divergence time (τ_{AB}), P_{BCA} and P_{ABC} are equivalent between results for neutral and selected loci (0 and 1.0, respectively).

Discussion

The multispecies coalescent model has become a cornerstone of molecular systematics, yet major questions remain about the impacts of model violations, such as selection. Previous studies have relied on intuition to formulate arguments for the robustness (or lack of) of coalescent inferences to the presence of selection. Our study thus represents a "first-step" perspective into the effects of model violations in the form of species-specific positive selection, which we have shown to bias gene tree distributions and influence downstream estimates of evolutionary history under certain conditions explored in our simulations. Importantly, our simulations suggest that the efficacy of natural selection to influence species tree and species delimitation estimates is highly dependent on particular evolutionary scenarios and experimental conditions, and these factors are relevant when considering the practical implications of our study. In general, we find that selection often acts synergistically with other parameters, such that the effects of selection are greatest when sample sizes are large, strong selection is present at multiple loci, and species are recently diverged. In agreement with opinions discussed in previous studies (Edwards 2009a; Edwards et al. 2016), we find that species tree estimates and delimitations are relatively robust to the effects of selection under more realistic conditions explored in our simulations that are most likely to be encountered in empirical studies. Nonetheless, we documented both expected and unexpected trends in the presence and absence of selection, which should serve as an initial benchmark for understanding the effects of positive selection on coalescent inferences of phylogeny.

Selection can Bias Estimates of Population Size and Divergence Time

We find that loci under positive selection tend to provide misleading evidence of smaller effective population sizes and deeper divergence times for the taxa experiencing positive selection (Fig. 4). These biases are most exaggerated when species are relatively closely related (i.e., shallow- and moderate-depth simulations) and are minor when lineages are deeply diverged. Selection increases the rate of coalescence (O'Fallon et al. 2010), thus resembling a decrease in θ that yields genealogies characterized by short coalescent times and monophyletic topologies for taxa under selection (Charlesworth 2009). Interestingly, we also identified a slight increase in the effective population size estimates for the sister taxon not experiencing selection (θ_B) when selection is present in a closely related, yet genetically-distinct species (herein, Species-A). The number of individuals sampled per species also increases these biases, particularly when species are recently-diverged. In many scenarios, we find that relatively weak selection had little effect on parameter estimates, and that increasing the proportion of neutral loci substantially reduced (i.e., diluted) biases introduced by selection.

Selected Loci Increase Posterior Probabilities of Species Hypotheses

Selected loci may have substantial effects on coalescent species delimitation under some scenarios. In our simulations, the inclusion of selected loci tended to increase the statistical resolution of species because selected genealogies exhibit an increased propensity for monophyly. Loci under selection also bias estimates of θ and τ, providing stronger evidence of genetic isolation of lineages when compared with neutral loci. These findings are intuitive from a population genetic perspective: positive selection will drive more rapid changes in allele frequency, providing stronger signal of population differentiation, compared with loci evolving under neutral processes (i.e., Fig. 3).

As observed in previous studies (e.g., Zhang et al. 2011; Yang and Rannala 2010b), we find that statistical resolution of species increases with the number of loci, number of individuals sampled per taxa, and divergence times. Our simulations demonstrate that selected loci can further increase posterior probabilities of species hypotheses (P_3, P_A, and P_B) when compared with inferences based solely on neutral loci. We also observed a slight increase in P_{BC} in some cases; this effect was relatively weak compared with increases in P_3, P_A, and P_B, and was largely restricted to specific scenarios. Our findings therefore imply that the type of selection we simulated (directional selection within single species) primarily acts to decrease error in species delimitation inferences. However, these effects are substantially reduced as the proportion of neutral loci is increased. Indeed, inferences from neutral data sets and data sets containing 10-20% selected loci were often comparable. Selection also had little influence over estimates of deeply-diverged taxa because neutral loci alone exhibit sufficient evidence of evolutionary independence when species are distantly related (i.e., $P_3 = P_A = P_B = 1.0$).

Phylogenetic resolution of closely-related species complexes is notoriously challenging, and thus we based our simulations on recently diverged taxa to understand the effects of selection in such scenarios (Maddison and Knowles 2006; Shaffer and Thomson 2007; Leavitt et al. 2011; Zhang et al. 2011; Liu et al. 2012; Pepper et al. 2013). Inferences based on a single locus or on few loci often suffer from considerable uncertainty due to ILS and gene tree estimation error (i.e., lack of phylogenetic signal) that may be prevalent when species are recently-diverged. In many cases, we find that posterior inferences derived from neutral data sets were the same or nearly the same as the prior probabilities (i.e., $P_3 = 1/3$), generally highlighting the need for larger data sets and sample sizes to resolve species limits using neutral loci alone. Conversely, we find increased statistical support for the true species model even for single-locus inferences in the presence of relatively weak selection. Non-model based approaches, such as reciprocal monophyly, will also likely "benefit" from increased resolution afforded by selected genealogies that are more likely to exhibit monophyly (i.e., Fig. 2).

Species Tree Inferences can Be Biased Under Some Conditions of Positive Selection

While selection largely reduced error in species delimitation, our simulations revealed an opposite effect on species topology estimates under some conditions. Our analyses of population differentiation under selection provide insight into these behaviors, where we find F_{ST} estimates are often higher between sister taxa Species-A and Species-B than between Species-B and the outgroup (Species-C; Fig. 3a vs. 3b). Simulated genealogies with selection tended to have an overrepresentation of coalescent events between neutrally evolving, non-sister lineages (Species-B and Species-C), and an under-representation of coalescent events between the closely-related sister lineages (Species-A and Species-B, see e.g., genealogy in Fig. 1). Therefore, selection can bias species tree inferences toward an incorrect topology because gene tree distributions simulated under some scenarios of selection do represent those expected under the multispecies coalescent model (i.e., coalescent events between more distantly-related taxa are more probable; Fig. 2) and because selection tends to inflate divergence time estimates that are used to root the topology at the longest branch length with BPP (Fig. 4).

The effects of selection on species tree estimation are also highly sensitive to the particular simulation conditions—we observe the strongest biases when selection is strong and present at multiple loci, when sample sizes are large, and when lineages diverged recently. Under the most extreme conditions in which 100% of loci were under strong selection, we find that the true rooted species topology is nearly absent from the posterior distribution, such that the incorrect topology is inferred with nearly 100% probability (Fig. 6 and Supplementary Fig. S5 available on Dryad). This misleading effect of selection was largely limited to extreme scenarios of strong selection occurring at multiple loci in our shallow and moderate-depth species models, although we also observed increases in P_{BCA} and decreases in P_{ABC} for single and 2-locus data sets in some cases. Importantly, increasing the number of neutral loci appears to effectively overcome this bias, such that topology estimates are relatively robust in many scenarios. For example, species tree estimates are largely unaffected by even strong selection present at 100% of loci for our deep species models (Supplementary Fig. S6 available on Dryad).

Does Species-Specific Positive Selection Pose Risks for Empirical Studies?

The relevance of selection-driven biases in empirical studies is largely contingent on the loci sampled for analyses, and the probability that such loci are under selection. Accommodating ILS as a source of gene tree conflict is imperative for species tree estimation and species delimitation because ILS is inherently linked to the process of speciation and acts on a genome-wide scale (Edwards 2009a). Unlike ILS, the "genomic footprint" of positive selection is thought to comprise only a small proportion of the genome containing alleles that increase the fitness of certain individuals, and surrounding regions that are genetically linked to such loci. It is unclear whether speciation is commonly accompanied by positive selection or not. Debates on this subject have continued over the past century, with some authors suggesting speciation-with-selection is widespread in nature (Mayr 1949; Panhuis et al. 2001; Rundle and Nosil 2005; Schluter 2009), and others arguing the opposite (Nei 1976; Nei et al. 1983; Orr and Orr 1996). Thus, the persistent question of how pervasive the genomic effects of selection are in nature has major bearing on how relevant biases due to selection are for empirical analyses.

Our simulations demonstrate that the impacts of selection can be quite severe, yet the effects of selection were largely limited to specific scenarios of strong selection occurring at multiple loci in the analysis of closely-related species. Importantly, we found that both species tree estimation and delimitation were fairly robust to the presence of even strong selection when as much as 10–20% loci were under selection in 10-locus data sets. Although our simulations revealed strong biases in gene tree distributions simulated under some scenarios of selection, the impacts of selection on downstream inferences appear to behave in a "dosage-dependent" manner, such that any effects are diminished by increasing the proportion of neutral loci. Empirical data sets commonly include hundreds to thousands of loci, such that the presence of a small number of positively-selected loci is likely of little consequence for genome-scale analyses, based on our simulations. The proportion of loci that have experienced positive selection likely differs greatly from species to species, but most empirical studies support the idea that only a relatively small fraction of the genome is likely under direct positive selection (i.e., <10% of genomic loci; Voight et al. 2006; Hohenlohe et al. 2010). For example, comparison of human and chimp genomes revealed that ~1.7% and ~1.1% of loci have undergone direct positive selection in each lineage, respectively (Bakewell et al. 2007). However, some empirical studies have documented evidence of widespread positive selection in nature: >90% of genomic loci are thought to have undergone species-specific positive selection in *Campylobacter* (Lefébure and Stanhope 2009), 30-94% of loci in *Drosophilia* (Fay et al. 2002), and 60% of amino acid substitutions in *Orychtolagus* (Carneiro et al. 2012). In light of these findings, several authors have proposed a shift toward a selection model of molecular evolution that may better explain these patterns (Hahn 2008; Corbett-Detig et al. 2015). These topics have remained a subject of intense debate among evolutionary biologists and are beyond the scope of this study. Until more examples of widespread positive selection emerges, we expect that coalescent inferences of phylogenetic relationships are relatively robust to the effects of positive selection under most conditions likely to be found in nature.

Our study represents a “first-step” analysis of scenarios of speciation-with-selection in the context of the multispecies coalescent model, and although we have explored a variety of scenarios and conditions, there are many other factors that we were not able to evaluate. Specifically, we restricted our simulations to the study of three-species models to explore the effects of selection across a range of conditions in a tractable manner. Given the relatively short divergence times used in our simulations, our species models may be interpreted as closely-related populations or incipient species in which a single taxon has experienced positive selection following speciation. Because both selection and ILS act in relation to population sizes and divergence times, we expect the impacts of selection will vary with different population sizes and trajectories (i.e., bottlenecks), as well as divergence times. Balancing selection, unlike positive selection simulated in our study, is predicted to have substantially different effects on gene trees (i.e., deeper coalescent times, which may be important considerations for future studies (Takahata and Nei 1990). We also restricted analyses to a single program (BPP) for computational feasibility and for direct comparisons across simulations. While we expect similar results with other programs, it is notable that there are now a variety of coalescent frameworks that differ in key model assumptions, such as gene tree estimation error, among-locus rate variation, and heterotachy (i.e., substitution rates differ among branches), as well as statistical approaches (i.e., Bayesian vs. maximum likelihood). For example, methods that only use minima of gene tree parameters (i.e., minimized coalescent times) to reconstruct species trees, such as BEST (Liu 2008), would be predicted to be more heavily influenced by locus-specific effects of selection. Further evaluation of the impacts of selection in such expanded contexts would be valuable because results may differ from what we have found using BPP.

Important avenues for future research include evaluating potential interactions of selection with other evolutionary processes, such as recombination, gene flow, and impacts of other types of selection (i.e., disruptive, convergent, and balancing selection). For example, adaptive convergent evolution at even a small proportion of sites has been shown to mislead gene tree inference (Castoe et al. 2009), yet we do not know how biases introduced by these sites may percolate from gene tree to species tree inferences. Although our simulations suggest that neutral loci are largely capable of overcoming signal from positive selection, empirical evidence suggests that information provided by a small number of sites or genes may dominate phylogenomic inferences (Shen et al. 2017); these and other concerns are important for understanding how genomic-scale inferences may be influenced by model violations at both site-specific and genealogical levels.

We focused our study on the analyses of sequences linked to a single, positively-selected site whereby the selective pressure is applied immediately after speciation and occurs continuously until the present within a single taxa. Selection, however, often acts to increase genetic linkage among sites and may also involve distant, coevolving loci via epistasis, which could entail further model violations to the assumption of independence among loci required by coalescent methods such as BPP. Genetic linkage and epistasis may therefore lead to a more substantial portion of the genome being effected by selection, and thus increase the effects of selection beyond that observed in our study. Recent analyses of primate genomes illustrate this point, as they suggest that most regions of the hominid genome have been influenced by selection either directly or indirectly (i.e., because of genetic linkage) throughout primate evolution (McVicker et al. 2009; Hobolth et al. 2011; Scally et al. 2012). Finally, while gene flow can mislead species tree estimation and delimitation (Leaché et al. 2014; Burbrink and Guiher 2015; Solís-Lemus et al. 2016), we expect that selected loci will provide increased resolution of species histories under scenarios of migration when neutral loci may fail to provide accurate inferences. Although computationally expensive, realistic whole-genome simulations that incorporate selection, recombination, gene flow, and other processes will be necessary to fully evaluate whether species tree estimates and delimitations are robust to more complex—yet perhaps more realistic—scenarios of speciation.

Conclusion

Questions remain about how pervasive positive selection is in nature, and how many loci it may impact throughout the genome—addressing these questions are of broad relevance for understanding speciation and the evolutionary process, and are also of central importance for predicting the practical relevance of selection-driven effects observed in our study. Our results suggest that coalescent species tree estimation and delimitation can be susceptible to selection-driven biases under certain circumstances, including when lineages are recently diverged, and when selection is more pervasive. However, if selection and its effects are relatively rare on the scale of genomes, empirical inferences are likely to be fairly robust to these violations of the multispecies coalescent model. While larger genomic sampling should overcome biases in species tree estimation due to selection, it would also be feasible to identify and remove loci with evidence of species-specific positive selection prior to analyses—although identifying selected loci can be difficult in practice. Although filtering of data to avoid model violations is one logical approach, counter-arguments to include neutral and selected loci are also logical, at least for species delimitation. For example, it is notable that recent selection tended to reduce error in species delineation in closely-related lineages, leading to higher probabilities of delimiting recently-diverged (presumably locally-adapted) species when selection is occurring. Further, an indirect observation arising from our study is that coalescent species delimitation approaches might be useful for identifying positive selection in multi-locus data sets: one might conduct species delimitation independently for each locus, and loci that provide higher posterior probabilities of species hypotheses may represent targets of selection (as demonstrated in Fig. 5). Such an approach would be attractive because it would effectively account for ILS while conducting genomic scans of selection, which is important because measures of population differentiation between lineages are inherently a function of these processes.

Supplementary Material

Data are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.5v3b5.

Funding

This work was supported by the University of Texas at Arlington to T.A.C., NSF [DEB-1655571 to T.A.C.; DEB-1501747 to T.A.C. and D.C.C.; DEB-1501886 to T.A.C. and D.R.S.], and Phi Sigma Support to R.H.A.

Acknowledgments

We thank Matthew Fujita and Adam Leaché for valuable insights and helpful discussions. Additionally, both the Lonestar and Stampede compute systems of the Texas Advanced Computing Center (TACC) were utilized for these analyses.

References

Adams R.H., Schield D.R., Card D.C., Blackmon H., Castoe T.A. 2016. *GppFst*: Genomic posterior predictive simulations of F_{ST} and d_{XY} for identifying outlier loci from population genomic data. Bioinformatics 33(9):1414–1415.

Bakewell M.A., Shi P., Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc. Natl. Acad. Sci. U.S.A. 104:7489–7494.

Barton N.H., Etheridge A.M., Sturm A.K. 2004. Coalescence in a random background. Ann. Appl. Probab. 14:754–785.

Burbrink F.T., Guiher T.J. 2015. Considering gene flow when using coalescent methods to delimit lineages of North American pitvipers of the genus *Agkistrodon*. Zool. J. Linn. Soc. 173:505–526.

Camargo A., Avila L.J., Morando M., Sites 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.

Carneiro M., Albert F.W., Melo-Ferreira J., Galtier N., Gayral P., Blanco-Aguiar J.A., Villafuerte R., Nachman M.W., Ferrand N. 2012. Evidence for widespread positive and purifying selection across the European rabbit (*Oryctolagus cuniculus*) genome. Mol. Biol. Evol. 29:1837–1849.

Castoe T.A., de Koning A.P.J., Kim H.-M., Gu W., Noonan B.P., Naylor G., Jiang Z.J., Parkinson C.L., Pollock D.D. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc. Natl. Acad. Sci. U.S.A. 106:8986–91.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. 10:195–205.

Corbett-Detig R.B., Hartl D.L., Sackton T.B. 2015. Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. 13.

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

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

Edwards S. V. 2009b. Natural selection and phylogenetic analysis. Proc. Natl. Acad. Sci. U.S.A. 106:8799–8800.

Edwards S. V., Xi Z., Janke A., Faircloth B.C., McCormack J.E., Glenn T.C., Zhong B., Wu S., Lemmon E.M., Lemmon A.R., Leaché A.D., Liu L., Davis C.C. 2016. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Mol. Phylogenet. Evol. 94:447–462.

Ewing G., Hermisson J. 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26:2064–2065.

Fay J.C., Wyckoff G.J., Wu C.-I. 2002. Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature 415:1024–6.

Fujita M.K., Leaché A.D., Burbrink F.T., McGuire J.A., Moritz C. 2012. Coalescent-based species delimitation in an integrative taxonomy. Trends Ecol. Evol. 27:480–488.

Hahn M.W. 2008. Toward a selection theory of molecular evolution. Evolution (N.Y.) 62:255–265.

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

Hey, J. 1994. Bridging phylogenetics and population genetics with gene tree models. In: Schierwater, B., Streit, B., Wagner, G.P., Desalle, R. editors. *Molecular Ecology and Evolution: Approaches and Applications.* Basel: Birkhäuser 435–449.

Hobolth A., Dutheil J.Y., Hawks J., Schierup M.H., Mailund T. 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Res. 21:349–356.

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 Genet. 6:e1000862.

Huang H., He Q., Kubatko L.S., Knowles L.L. 2010. Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Syst. Biol. 59:573–583.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. Mamm. Protein Metab. 3:21–123.

Kaplan N.L., Hudson R.R., Langley C.H. 1989. The "hitchhiking effect" revisited. Genetics 123:887–899.

Lanier H.C., Knowles L.L. 2012. Is recombination a problem for species-tree analyses? Syst. Biol. 61:691–701.

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

Leaché A.D., Harris R.B., Rannala B., Yang Z. 2014. The influence of gene flow on species tree estimation: a simulation study. Syst. Biol. 63:17–30.

Leavitt S.D., Fankhauser J.D., Leavitt D.H., Porter L.D., Johnson L.A., St. Clair L.L. 2011. Complex patterns of speciation in cosmopolitan "rock posy" lichens—Discovering and delimiting cryptic fungal species in the lichen-forming *Rhizoplaca melanophthalma* species-complex (Lecanoraceae, Ascomycota). Mol. Phylogenet. Evol. 59:587–602.

Lefébure T., Stanhope M.J. 2009. Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus *Campylobacter*. Genome Res. 19:1224–1232.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24:2542–2543.

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 S., Colvin J., De Barro P.J. 2012. Species concepts as applied to the whitefly *Bemisia tabaci* systematics: how many species are there? J. Integr. Agric. 11:176–186.

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

Mayr E. 1949. Speciation and selection. Proc. Am. Philos. Soc. 93:514–519.

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.

McVicker G., Gordon D., Davis C., Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 5.

Nachman M.W., Crowell S.L. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156:297–304.

Nei, M. 1976. Mathematical models of speciation and genetic distance. In: Samuel Karlin, editor. *Population genetics and ecology*. New York: Academic Press 723–766.

Nei M., Maruyama T., Wu C.I. 1983. Models of evolution of reproductive isolation. Genetics 103:557–579.

O'Fallon B.D., Seger J., Adler F.R. 2010. A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. Mol. Biol. Evol. 27:1162–1172.

Orr H.A., Orr L.H. 1996. Waiting for speciation: the effect of population subdivision on the time to speciation. Evolution (N.Y.) 50:1742.

Panhuis T.M., Butlin R., Zuk M., Tregenza T. 2001. Sexual selection and speciation. Trends Ecol. Evol. 16:364–371.

Pepper M., Doughty P., Fujita M.K., Moritz C., Keogh J.S. 2013. Speciation on the rocks: integrated systematics of the *Heteronotia spelea* species complex (Gekkota; Reptilia) from western and central Australia. PLoS One 8.

Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

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.

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

Rundle H.D., Nosil P. 2005. Ecological speciation. Ecol. Lett. 8:336–352.

Scally A., Dutheil J.Y., Hillier L.W., Jordan G.E., Goodhead I., Herrero J., Hobolth A., Lappalainen T., Mailund T., Marques-Bonet T., McCarthy S., Montgomery S.H., Schwalie P.C., Tang Y.A., Ward M.C., Xue Y., Yngvadottir B., Alkan C., Andersen L.N., Ayub Q., Ball E. V., Beal K., Bradley B.J., Chen Y., Clee C.M., Fitzgerald S., Graves T.A., Gu Y., Heath P., Heger A., Karakoc E., Kolb-Kokocinski A., Laird G.K., Lunter G., Meader S., Mort M., Mullikin J.C., Munch K., O'Connor T.D., Phillips A.D., Prado-Martinez J., Rogers A.S., Sajjadian S., Schmidt D., Shaw K., Simpson J.T., Stenson P.D., Turner D.J., Vigilant L., Vilella A.J., Whitener W., Zhu B., Cooper D.N., de Jong P., Dermitzakis E.T., Eichler E.E., Flicek P., Goldman N., Mundy N.I., Ning Z., Odom D.T., Ponting C.P., Quail M.A., Ryder O.A., Searle S.M., Warren W.C., Wilson R.K., Schierup M.H., Rogers J., Tyler-Smith C., Durbin R. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483:169–175.

Schluter D. 2009. Evidence for ecological speciation and its alternative. Science 323:737–741.

Schrider D., Shanku A.G., Kern A.D. 2016. Effects of linked selective sweeps on demographic inference and model selection. Genetics 204(3):1207–1223

Shaffer H.B., Thomson R. 2007. Delimiting species in recent radiations. Syst. Biol. 56:896–906.

Shen X.-X., Hittinger C.T., Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1:126.

Solís-Lemus C., Knowles L.L., Ané C. 2015. Bayesian species delimitation combining multiple genes and traits in a unified framework. Evolution 69:492–507.

Solís-Lemus C., Yang M., Ané C. 2016. Inconsistency of species-tree methods under gene flow. Syst. Biol. 65:843–851.

Springer M.S., Gatesy J. 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94:1–33.

Stewart C.B., Schilling J.W., Wilson A.C. 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature 330:401–404.

Sukumaran J., Knowles L.L. 2017. Multispecies coalescent delimits structure, not species. Proc. Natl. Acad. Sci. U.S.A. 114:1607–1612.

Takahata N., Nei M. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics 124:967–978.

Ting C.-T., Tsaur S.-C., Wu C.-I. 2000. The phylogeny of closely related species as revealed by the genealogy of a speciation gene, Odysseus. Proc. Natl. Acad. Sci. U.S.A. 97:5313–5316.

Voight B.F., Kudaravalli S., Wen X., Pritchard J.K. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:e72.

Wakeley J. 2008. Coalescent Theory: An Introduction. Greenwood Village (CO): Roberts and Company Publishers.

Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. U.S.A. 107:9264–9.

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

Zhang D.-X., Hewitt G.M. 2003. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Mol. Ecol. 12:563–584.

Zhu T., Yang Z. 2012. Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. Mol. Biol. Evol. 29:3131–3142.