WHO CARES ABOUT MODELS WHEN YOU HAVE GENOMES?

by

RICHARD HUNTER ADAMS

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2019

Acknowledgements

I have many people to thank for their support and encouragement over the past few years. Above all, I am grateful to my advisor, Todd Castoe, for being an outstanding mentor, colleague, and friend. Five years ago, I was working in a dog food factory with zero experience in computational biology. You nonetheless accepted me into your lab and have since been one of the most important contributors to my success and academic career and this dissertation is a testament to that fact. I would also like to thank my friends and fellow lab mates Drew Schield, Daren Card, Andrew Corbin, Blair Perry, Giulia Pasquesi, Nicky Hales, Andrea Westfall, Zach Nikolakis, and Ricky Orton – It's been a long, strange trip! I am also very grateful to my committee members and their labs: Esther Betran, Jeff Demuth, Matt Fujita, and Matt Walsh. Thank you to all my friend and colleagues in and outside the UTA biology department, including Jill Castoe, Dr. Steve Mackessy, Dr. Tereza Jezkova, Dr. Paul Chippindale, Dr. Jacobo Reyes-Velasco, James Titus MacQuillan, Shannon Beston, Contessa Ricci, Eli Wostl, and Rachel Wostl. Finally, I would like to thank Eric Smith for the many collaborations and chats about coralsnakes population genetics and phylogenetics over the years, as well as Heath Blackmon for being a friend, colleague, and role model.

April 10th 2019

Dedication

My dissertation is dedicated to my wife, Emily, and my parents. Not only have you suffered my insanity for the past 5 years, but you have actively supported it. To my mom – my guidepost in not only academia, but in my life. I hope you always know how much I admire your dedication to science and your family above all else. You represented a woman scientist in a time when the cards were stacked against you in this field and I am proud to call you mom. Thank you for fostering in me a love and respect of all things living (from cleaning up pet stores to paying our our respects to the reptile cemetery). To my dad – thank you for supporting me every step of the way, and I am proud to call you my dad. Emily – you have been my partner through it all. Since that first day in Folklore class, you've been a constant source of support and encouragement for my pursuits in and outside academia.

Abstract

WHO NEEDS MODELS WHEN YOU HAVE GENOMES?

Richard H. Adams, PhD

The University of Texas at Arlington, 2018

Supervising Professor: Todd A. Castoe, PhD

There is no question about it: genomic data are revolutionizing biology. This is certainly evident in the fields of population genetics and phylogenetics for which genome-scale analyses have been used to study a myriad of evolutionary processes and organismal relationships across the Tree of Life. While genomic data have unquestionably advanced our understanding of biology by incredible leaps and bounds, the ease and affordability of generating such large and complex data has unfortunately, in some circumstances, led to the idea that simply "throwing more data" at a particular evolutionary question is likely to be sufficient. This notion has led to an emphasis on obtaining larger datasets with the hope that one can overcome most any obstacle by simply increasing the sample size without considering the fit of these large, complex datasets to the highly oversimplified models that we often use to analyze these types of data. The title of my dissertation represents a rhetorical sarcastic question that my research has addressed.

Table of Contents

Chapter 1

Introduction

Much of evolutionary research is inferential in nature, as we seldom – if ever – know the "true" evolutionary history of organisms or molecules. On a fundamental level, genomic data offer unprecedented opportunity to increase the accuracy and resolution of evolutionary inferences by reducing sampling error that results simply from insufficient sample size. For example, one of the primary goals of population genetics is to understand how different evolutionary processes shape patterns of genetic variation in nature, and studies now routinely leverage population genomic data to scan along entire chromosomes and dissect locus-specific evidence of selection, migration, recombination, and other processes (Folla & Gaggiotti, 2008; Narum & Hess, 2011; O'Reilly, Birney, & Balding, 2008; Oleksyk, Smith, & O'Brien, 2010; Vitti, Grossman, & Sabeti, 2013). Additionally, the field of phylogenetics has largely transitioned into phylogenomics, whereby species-level relationships are commonly reconstructed with high resolution from thousands to millions of base pairs (Edwards et al. 2007; Degnan and Rosenberg 2009a; Edwards 2009a; Fujita et al. 2012). Indeed, genome-scale data are well-poised to solve many longstanding questions in evolutionary biology with high precision.

However, there is another source of statistical error that is not a result of small sample size and may not be alleviated by even genome-scale data: systematic error. Systematic error arises from the failure of a model to adequately describe the important statistical properties of a dataset, which in turn may mislead inferences towards incorrect conclusions with high confidence – even with infinite data (Felsenstein 1978; Bollback 2002; Huang et al. 2010; Warnow 2015). In the context of population genomic inference, systematic error may mislead researchers to incorrectly

conclude that certain loci are important for adaptation that may have – in fact – evolved via drift (Teshima et al. 2006; O'Reilly et al. 2008; Narum and Hess 2011; Pavlidis et al. 2012). Phylogenomic analyses can be biased by systematic error towards incorrect estimates of evolutionary history with high statistical support when the model of molecular evolution is inadequate (Sullivan and Swofford 1997; Buckley 2002; Brown and Lemmon 2007; Huang et al. 2010; Brown 2014; Leache et al. 2014; Roch and Warnow 2015). For example, recent studies have demonstrated that poor model fit at even a small handful of sites can overwhelm genome-scale inferences (Castoe et al. 2009a; Shen et al. 2017a). Thus, simply "throwing more data" at a particular evolutionary question does not guarantee that conclusions will be any more accurate.

The overarching goal uniting the five chapters of my dissertation is to illuminate both strengths and weaknesses of contemporary models and statistical methods for evolutionary inference from molecular data. Although it is well-acknowledged that poor model fit and its accompanied systematic error can mislead evolutionary inferences, we seldom understand how *well* current models describe our data, and even cursory examinations of model adequacy are rare in the literature (Goldman 1993; Sullivan and Swofford 1997; Bollback 2002; Lemmon and Moriarty 2004; Kelchner and Thomas 2007; Waddell et al. 2009; Reid et al. 2014). As an effort to more fully understand evolutionary model adequacy, my dissertation is designed to address questions of model-based inference and selection in four different areas of evolutionary research: population genomics (Chapter 2), species tree estimation and delimitation (Chapter 2), heuristic phylogenomic approaches (Chapter 3), and model-based measures of tree distance (Chapter 4).

Chapter 2

Bayesian inference of natural selection and demography from population genomic data

Richard H. Adams[a], Drew R. Schield[a], Daren C. Card[a], and Todd A. Castoe[a]

[a]Department of Biology & Amphibian and Reptile Diversity Research Center, 501 S. Nedderman Drive, University of Texas at Arlington, Arlington, TX 76019 USA

Abstract

Genome scans of population differentiation can provide powerful insight into the evolutionary processes at work across the genome. Perhaps the most popular application for these approaches is to infer natural selection based on patterns of genetic variation at different genomic loci. Signatures of selection can manifest as extreme measures of population differentiation at specific genomic regions, which are often referred to as "outlier" loci when evaluating population genomic data. In practice, ad-hoc techniques and heuristic thresholds are typically used to determine if these outlier loci represent targets of selection (or not). Lacking from most of these methods is an explicit, statistical framework that accounts for the demographic history of organisms (and uncertainty in demographic history) when determining if putative selected loci are under selection. Importantly, neutral processes alone can occasionally yield "outlier" loci due solely to random change, and it is likely that – particularly when sampling large genome-scale datasets – some small fraction of loci may exhibit "outlier" patterns of divergence, even when these loci evolve under drift alone. Here we have developed two Bayesian approaches for inferring selection from genome scans while explicitly considering the particular demographic history of the populations under study. These two methods present theoretical improvements to the rigor of genome scans of selection and highlight the importance of specifying appropriate null models when inferring specific evolutionary processes.

GppFst: Genomic posterior predictive simulations of FST and DXY for identifying outlier loci from population genomic data

Richard H. Adams[1], Drew R. Schield[1], Daren C. Card[1], Heath Blackmon[2], and Todd A. Castoe[1, §]

[1]Department of Biology, 501 S. Nedderman Dr., The University of Texas at Arlington, Arlington, TX 76010, USA

[2]Department of Biology, Texas A&M University College Station, TX 77845, USA

Introduction

Genomic distributions of genetic differentiation provide a powerful framework for inferring evolutionary processes that have impacted regions of the genome. Two commonly used measures of genetic differentiation are F_{ST} (Wright 1949), and d_{XY} (Takahata and Nei 1985). These metrics have been applied extensively to characterize genome-wide patterns of genetic variation and differentiation across a wide range of populations and species (Jensen et al. 2016).

In nature, most genomic variation is thought to derive from genetic drift occurring within structured populations. This expectation serves as a null model for identifying loci with patterns of genetic variation that differ significantly from the rest of the genome (so-called 'outlier' loci). Numerous studies have applied this principle to identify loci with extreme patterns of genetic differentiation that are poorly-explained by neutral processes alone, and thus may indicate selection (Jensen et al. 2016). Genetic differentiation can, however, be influenced by multiple factors; for example, small population sizes and deep divergence may shift neutral genomic distributions towards larger values of F_{ST} and d_{XY}, which can confound inferences of selection. Furthermore, most F_{ST}-based models assume equal rates of drift within the populations under study (Weir and Cockerham 1984). Currently, no methods use an explicit probabilistic population model that incorporates demographic parameters to predict the distribution of neutral variation in F_{ST} and d_{XY}. For example, *pFst* employs a likelihood ratio test of allele frequency differences between populations (Shapiro et al. 2013), while *BayeScan*, uses logistic regression to determine locus-specific departure from neutrality (Foll and Gaggiotti 2008).

Here we describe a posterior predictive simulation (PPS) framework to generate theoretical distributions of F_{ST} and d_{XY} under the neutral coalescent model for two populations that accounts

for demographic parameters in a probabilistic framework. Importantly, our method allows users to explicitly test the null hypothesis of genetic drift when conducting genomic scans. PPS is a popular method for evaluating model fit within a Bayesian framework that has been used to test a variety of evolutionary models (Gelman et al. 2004; Reid et al. 2014). Unlike other F_{ST} outlier tests, our PPS approach explicitly accounts for the demographic history of two genetically-isolated species, including multiple demographic and experimental parameters (and uncertainty in those parameters), such as sample sizes, demographic parameters ($\theta = 4N_e\mu$), unequal rates of genetic drift within populations (unequal θs), and divergence time (τ). Additionally, other genomic F_{ST} outlier tests assume free recombination among SNPs. Our method allows users to simulate theoretical distributions that are conditioned on sampling multiple linked SNPs per locus – allowing users to take full advantage of large genomic datasets. We provide our PPS model in the package *GppFst* (Genomic Posterior Predictive distributions of F_{ST}), which offers a user-friendly, open-source framework to generate theoretical distributions of F_{ST} and d_{XY} under the neutral coalescent model.

Implementation

The R package *GppFst* was written in R 3.2.2 (R Core Team, 2013), and requires two other R packages, phybase (Liu and Yu 2010) and Geneland (Guillot et al. 2005) for simulating genealogies and computing Weir and Cockerham's F_{ST} (Weir and Cockerham, 1984). The functions *GppFst* and *GppDxy* require a posterior distribution of coalescent parameters (θ, τ) for a two-population model inferred via Markov Chain Monte Carlo (MCMC) sampling. This posterior distribution can be obtained using any program that implements a two-population coalescent model (see tutorial for examples). For each step in the MCMC, *GppFst* simulates

coalescent genealogies and sequence alignments using a modified version of the function *simSeqfromSp* provided from phybase. F_{ST} and d_{XY} values are then computed for each simulated alignment, with the number of alignments to simulate per step specified by the user. Users can account for several experimental parameters, including variation in missing data per population and locus, locus lengths, and particular SNP-subsampling schemes (SNPs sampled per linked genomic region). Both locus length and number of individuals per population are sampled from their empirical distributions, and users specify the number of SNPs to retain per simulated locus, which can be fixed at empirical values (i.e., 1 SNP per locus) or set to use all SNPs per locus. After generating a theoretical distribution of F_{ST} or d_{XY}, users can compare empirical and simulated distributions to assign significance to outlier loci poorly explained by the neutral coalescent model.

Biological Application

As a demonstration, we applied our *GppFst* model to a published RADseq SNP dataset (NCBI SRP051070) from two rattlesnake populations (Schield et al. 2015). We inferred demographic parameters from 7,031 unlinked nuclear SNPs with SNAPP (Bryant et al. 2012). Using *GppFst,* we generated a PPS distribution of F_{ST} to identify loci that are poorly explained by neutral processes alone. Comparisons of the relative frequencies of simulated and empirical loci within F_{ST} intervals highlight extreme F_{ST} intervals that exhibit an excess of empirical loci when compared to the PPS distribution (Fig. 1). To calculate the empirical *P*-value, we use the PPS distribution to determine the probability of observing a given proportion of empirical loci within a specified F_{ST} interval. For example, the proportion of loci with $F_{ST} = 1$ in the empirical distribution (0.0014) is more than ~10-fold greater than the proportion observed in the PPS

distribution (0.00012). Thus, observing 10 loci with $F_{ST} = 1$ is extremely unlikely under the neutral model ($P < 0.0001$). Comparisons between our method and others that do not incorporate probabilistic model-based approaches suggest that *GppFst* provides more conservative estimates of outlier F_{ST} loci. For example, *GppFst* incorrectly identified a significant excess of SNPs with $F_{ST} = 1$ in 4 of 100 simulated datasets (1,000 neutral SNPs each), while the program *Arlequin* (Excoffier *et al.*, 2005) incorrectly assigned significance to every locus with an $F_{ST} = 1$ in all 100 datasets (see tutorial). *GppFst* allows users to identify F_{ST} intervals with an excess of loci than expected under a neutral model. Our PPS framework employs the coalescent model of allopatric divergence between populations, which assumes free recombination between loci, no recombination within loci, and no gene flow. Because gene flow, recombination, and other factors may influence genomic variation, we recommend that users test all assumptions prior to using *GppFst*.

Acknowledgments and Funding

This work has been supported by a University of Texas at Arlington Phi Sigma Society Grant to R.H.A, startup funds to T.A.C., and NSF DDIG grants to D.R.S & T.A.C (NSF DEB-1501886) and D.C.C. & T.A.C (NSF DEB-1501747).

Figures

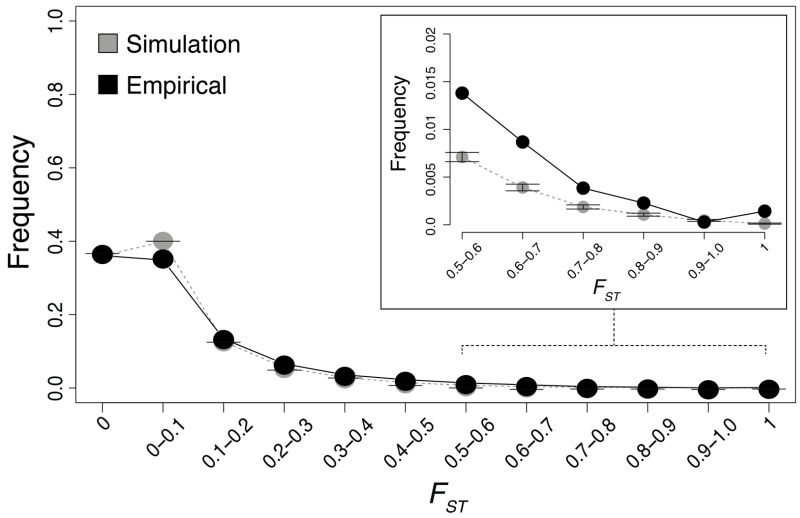


Figure 1. Empirical and posterior predictive simulated (PPS) distributions of F_{ST} for example data, with standard deviations. The mean proportion of loci from the 100 replicate PPS runs (gray) and proportion of loci in the empirical data (black) are shown. Inset (top-right) shows upper limit of the F_{ST} distribution, highlighting the difference between simulated and empirical distributions at extreme F_{ST} values.

Appendix

Genomic distributions of population genetic differentiation

F_{ST} and d_{XY} are two commonly used measures of genetic differentiation between two (or more) populations. When population genomic data are available, empirical distributions of these two measurements can be leveraged to identify loci with observed values that differ significantly from the rest of the genome. Under the assumptions of neutral theory, the majority of genomic loci are thought to evolve by genetic drift, such that these so-called 'outlier' loci are inferred to be targets of various forms of natural selection. For example, population genetic processes that homogenize genetic variation between populations (i.e., gene flow, convergent selection) will lead to lower F_{ST} and d_{XY} values, while divergent selection will yield stronger genetic differentiation and higher F_{ST}/d_{XY} values.

However, a number of population genetic processes can influence F_{ST} distributions , and discerning among these processes can be challenging when classifying outlier loci. For example, low F_{ST} values could result from gene flow and various forms of selection (i.e., balancing, convergent). Small population sizes and deep divergence may shift genomic distributions of F_{ST} and d_{XY} towards larger values, which can hinder or confound inferences concerning the importance of particular loci in divergent selection. Furthermore, most F_{ST} models of genetic differentiation assume equal population sizes, such that the rate of genetic drift is approximately equivalent within each population.

The R package GppFst (**G**enomic **P**osterior **P**redictive **F**st distributions) provides a robust framework to account for multiple evolutionary processes when conducting outlier tests for both F_{ST} and d_{XY} distributions. When population parameters are inferred from genomic data, the functions *GppFst* and *GppDxy* provided in this package will simulated theoretical distributions of F_{ST} and d_{XY} that can be used classify putative targets of selection. In short, our model accounts for the following sources of uncertainty that may influence F_{ST} and d_{XY} distributions: divergence time estimates, population size parameter estimates, different population size parameters between populations, unequal population sampling, and SNP sampling.

In our package F_{ST} is calculated as

$$F_{ST} = \frac{s^2}{\overline{p}(1-\overline{p})}$$

Where $\overline{p}$ is the allele frequency of allele A across populations and s^2 is the sample variance of allele A frequencies over populations calculated as:

$$s^2 = \sum_i \frac{n_i(\tilde{p}_i - \overline{p})^2}{(r-1)\overline{n}}$$

Where $\tilde{p}_i$ is allele frequency of allele A in population i calculated with a sample size of n. Finally r is the number of populations.

While d_{XY} is calculated as $D_{xy} = \sum_{ij} x_i\, y_j d_{ij}$ where d_{ij} measures the number of nucleotide differences between haplotype i from X and haplotype j from Y.

ThetaMater: Bayesian estimation of population size parameter θ from genomic data

Richard H. Adams[1], Drew R. Schield[1], Daren C. Card[1], Andrew Corbin[1], and Todd A. Castoe[1, §]

[1]Department of Biology, 501 S. Nedderman Dr., The University of Texas at Arlington, Arlington, TX 76010, USA

Introduction

The population size parameter $\theta = 4N_e\mu$ ($2N_e\mu$ for haploid organisms) reflects the mutation-drift balance occurring within a population with an effective size of N_e individuals and a mutation rate of μ per site per generation. As a measure of genetic diversity, θ represents the expected number of segregating sites observed between a pair of homologous sequences sampled from a given population (Wakeley 2008). Given an estimate of mutation rate, information about θ can be leveraged to obtain an estimate of the effective population size N_e. θ is a therefore a fundamental parameter of population genetics and is useful for understanding the degree to which neutral processes shape patterns of genetic variation in nature. Quantifying genetic diversity is also important to conservation biology, and thus estimates of θ provide critical insight into the genetic health of endangered species for informed conservation practices (Crandall et al. 1999).

Numerous methods and genetic models have been developed to estimate θ from genetic data (see Wang, 2005 for examples). As any estimate obtained from a single locus or a small set of loci entails substantial uncertainty, large genome-scale datasets offer opportunity to estimate θ with high accuracy and precision. However, few likelihood-based methods are currently scalable to such massive datasets (>10^6 loci, >10kb/locus), are often restricted to using a single or small set of diploid genomes, are restricted to a specific type of sequence data (i.e., whole genomes vs. reduced representation), or require users to make assumptions about generation time and mutation rates. For example, most implementations of the popular pairwise-sequential Markov coalescent model (PSMC) require whole-genome data and that users provide a mutation rate assumed to be identical across all loci (Li and Durbin 2011), while other methods are restricted

to using individual diploid genomes (Haubold, B. *et al.* 2010). There are many genealogy-based methods for estimating demographic parameters (Felsenstein 1992; Kuhner et al. 1995), but these are intractable for genomic datasets that include many individuals. Furthermore, no current methods provide a statistical framework for leveraging estimates of θ to filter potentially spurious loci from datasets (i.e., paralogs). Accordingly, there is major need for efficient and scalable likelihood-based methods for estimating θ from diverse genomic datasets.

Implementation

The R package *ThetaMater* was written in R and C++, and requires the R package MCMCpack (Martin et al. 2011) to simulate posterior probability distributions of θ. At the core of *ThetaMater* is the infinite-sites likelihood function (Watterson, 1975), which describes the probability distribution of observing *k* segregating sites in a sample size of *n* sequences obtained from a locus of size *l*. The likelihood of a genomic dataset under a given value of θ is then computed as a product of the individual-locus specific likelihoods (or summation of log-likelihoods), each with an associated number of segregating sites *k*, sample size *n* and length *l* (see manual for model description). We have further expanded this approach to incorporate a discretized-gamma model of among-locus rate variation to accommodate rate variation and to characterize the genomic landscape of among-locus rate variation by estimating the gamma shape parameter (Yang 1997). Importantly, our method provides a user-friendly framework for efficient estimation of θ and substitution rate variation that is scalable to diverse genome-scale datasets (>10^6 loci) with larger samples sizes (>10 genomes), while accounting for uncertainty within a likelihood-based framework. Our method collapses datasets into sets of unique patterns, such that under many conditions, there is almost no limit to the number of loci that can be used

to estimate θ within minutes on a desktop computer. Unlike other methods restricted to a particular format, *ThetaMater* includes functions for converting a variety of widely-used alignment formats into usable input, including whole-genome sequences, reduced-representation data (i.e., RADseq, sequence capture), and single or multilocus Sanger sequenced datasets. Finally, *ThetaMater* includes a posterior predictive simulator (PPS) that allows users to leverage estimates of θ to identify loci with evidence of model violations, such as selection (Adams *et al.* 2016) or paralogy.

ThetaMater includes three Bayesian Markov Chain Monte Carlo (MCMC) simulation models for estimating posterior distributions of θ: M1 (*ThetaMater.M1*) assumes no among-locus rate variation, M2 (*ThetaMater.M2*) estimates θ using a fixed α parameter, and M3 (*ThetaMater.M3*) estimates the joint posterior distribution of θ and α. We implement a gamma prior distribution for both θ and α with user-specified shape and scale parameters, and users can specify the number of rate classes used to approximate the distribution. The posterior predictive simulator function (*ThetaMater.PPS*) is directly integrated with the results from the three Bayesian models.

Biological Application

As a demonstration, we applied *ThetaMater* on a previously published RADseq dataset (2051 loci; Schield *et al.*, 2017). We conducted Bayesian estimation of using *ThetaMater.M1* for the empirical dataset before and after filtering loci with *ThetaMater.PPS* (Fig. 1A). We also simulated a large genomic dataset comprised of 10^6 loci (2kb each), sampling 20 genomes from a population with θ = 0.002 and among-locus rate variation = 0.5 (Fig. 1B). We specified the shape and scale parameters of the prior distribution at 10 and 0.0001 for the empirical example, and set prior parameters to 20 and 0.0001 for θ and 5 and 0.01 for α in the simulated analysis.

We ran the MCMC chain for a total of generations and discarded 10% as burn in. PPS were run using the unfiltered posterior distribution, simulating a single locus for all 10^4 generations present in the post-burn in MCMC samples using *ThetaMaterPPS*.

ThetaMater analysis of the unfiltered RADseq dataset suggested a mean θ estimate of 0.0019, corresponding to $N_e = 47{,}500$ assuming a mutation rate of 10^{-8} (Fig. 1A, red). PPS based on this posterior distribution identified 3 loci with a significant excess of mutations, and these loci were filtered prior to reanalysis with *ThetaMater*. The posterior distribution of N_e inferred was centered around 45,000 individuals after removing these potentially spurious loci (Fig. 1A, blue). *ThetaMater* analysis of the simulated data returned the simulated parameter values with high probability (Fig. 1B).

ThetaMater is optimized for diverse datasets, including single diploid genome analyses, multi-genome data, reduced-representation data, and single or multilocus alignments. ThetaMater assumes free recombination between loci, no recombination within loci, error-free SNP calls, and neutral evolution. We encourage all users to carefully consider these assumptions prior to analysis with ThetaMater (see manual). Given the user-friendly framework and tractability of *ThetaMater*, we expect *ThetaMater* to be useful for a variety of applications, including population biology, comparative genomics, and conservation biology.

Acknowledgments and Funding

This work has been supported by a University of Texas at Arlington Phi Sigma Society Grant to R.H.A, startup funds to T.A.C., and NSF DDIG grants to D.R.S & T.A.C (NSF DEB-1501886) and D.C.C. & T.A.C (NSF DEB-1501747).

Figures

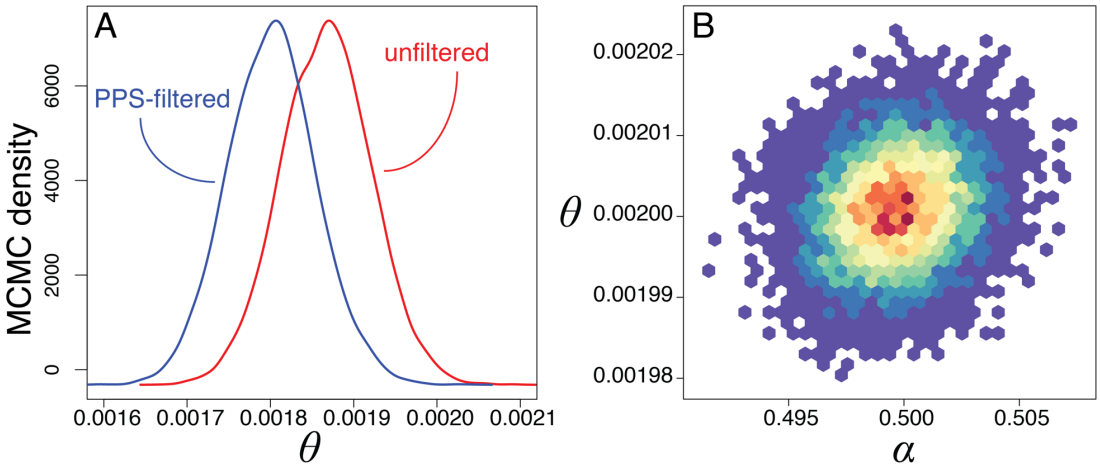


Figure 1. (A) Empirical posterior estimates of θ before (red) and after (blue) filtering with Thetamater.PPS, and (B) the joint posterior distribution of θ and α for the simulated dataset showing highest densities (warm colors) at the true simulated values ($\theta = 0.002$, $\alpha = 0.5$).

Appendix

The effective population size parameter θ

The population size parameter θ reflects the effects of genetic drift and mutation on patterns of genetic variation within a diploid population (for a haploid population) with an effective size of individuals and a mutation rate of per site per generation. If two homologous sequences are sampled at random from a population, describes the expected number of segregating sites observed between these two sequences. θ is a fundamental measure of genetic diversity in populations and is thus an informative parameter used in many population genetic models. The R package *ThetaMater* provides a Bayesian framework to estimate both θ and (shape of among-locus rate variation) parameters from a variety of genetic datasets, including haploid or diploid genomic data from single or multiple individuals, reduced-representation genomic data (e.g., RADseq, sequence capture), and single or multilocus Sanger sequence data (and variations of these datasets). ThetaMater implements three different functions that can be used to estimate these parameters within a Bayesian framework:

- *ThetaMater.M1*: estimate without among-locus variation
- *ThetaMater.M2*: estimate with a fixed parameter of rate variation and a user-defined number of locus rate classes
- *ThetaMater.M3*: estimate both and the shape parameter given a user-defined number of rate classes

The likelihood function implemented by ThetaMater

The three functions (*ThetaMater.M1*, *ThetaMater.M2*, *ThetaMater.M3*) simulate posterior probability distributions of effective population size parameters for a given dataset. These functions employ the likelihood function $P(S = k|l, n; \theta)$ to compute the probability of observing *k* segregating sites in a sample size of *n* from a locus with length *l* for a given value of θ. These methods compute the likelihood of a given dataset as a summation of the log-transformed likelihoods across all loci. See the following publications for more information about this model, its derivation, applications, and similar models:

- Tavaré, Simon. "Line-of-descent and genealogical processes, and their applications in population genetics models." Theoretical population biology 26.2 (1984): 119-164.
- Watterson, G.A. On the number of segregating sites in genetical models without recombination. Theoretical population biology 1975.
- Wakeley, John. "Coalescent theory." Roberts & Company (2009).
- Hein, Jotun, Mikkel Schierup, and Carsten Wiuf. Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford University Press, USA, 2004.
- Takahata, Naoyuki, and Yoko Satta. "Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences." Proceedings of the National Academy of Sciences 94.9 (1997): 4811-4815.

- Takahata, Naoyuki, Yoko Satta, and Jan Klein. “Divergence time and population size in the lineage leading to modern humans.” Theoretical population biology 48.2 (1995): 198-221.

- Yang, Ziheng. “On the estimation of ancestral population sizes of modern humans.” Genetical research 69.02 (1997): 111-116.

Below is the formula for the likelihood function described in these papers that is central to the three ThetaMater functions:

$$P(S = k|l, n; \theta) = \int_0^\infty P\,(S = k|t) f_T(t) dt$$

$$P(S = k|l, n; \theta) = (\frac{l\theta}{2})^k \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} \frac{i-1}{2} \int_0^\infty \frac{(t^k\, exp\{\frac{(-\theta + i - 1)t}{2}\})}{k!}\, dt$$

$$P(S = k|l, n; \theta) = (\frac{l\theta}{2})^k \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} \frac{i-1}{2} (\frac{2}{\theta + i - 1})^{(k+1)}$$

$$P(S = k|l, n; \theta) = \sum_{i=2}^{n} (-1)^i \binom{n-1}{i-1} \frac{i-1}{\theta + i - 1} (\frac{\theta}{\theta + i - 1})^k$$

For a dataset consisting of x loci, each an observed number of segregating sites k_i, number of bases l_i, and number of sequences sampled n_i, we can sum the likelihoods of the individual loci to get the likelihood of the entire dataset under a given value of θ:

$$L(D|\theta) = \sum_{i=1}^{x} log(P(S = k_i|l_i, n_i; \theta)$$

Applications, Assumptions, and Limitations of ThetaMater

Understanding the assumptions of *ThetaMater* and the underlying coalescent model are critical to the appropriate use of *ThetaMater*. Importantly, *ThetaMater* assumes that there is no recombination within individual loci and free recombination between loci (i.e., no linkage). Furthermore, all loci are assumed to have evolved under strictly neutral evolution. These are fundamental assumptions of the coalescent model and the likelihood function implemented in *ThetaMater*. This can been seen in the form of the likelihood equation provided above: the likelihood of an entire dataset is a summation of the log-likelihoods across loci that are assumed to be genetically unlinked. In other words, the genealogy and number of segregating sites observed at each locus is assumed to be independently and identically distributed (i.i.d).

To explore the potential effects of one such model violation (unrecognized recombination) in datasets, we simulated loci using the software *msprime* under 6 different recombination rates: (2e-9, 2e-8, 2e-7, 2e-6, 2e-5, 2e-4), using a sample size of 5 gene copies per 10kb locus, and with each dataset consisting of 10k loci. See Step 8: "Recombination & *ThetaMater*" for a plot of these analyses for each recombination rate and a dataset without recombination. In general, ThetaMater appears largely unaffected by recombination, as the posterior distribution of each analysis is largely centered around the true simulation value ($\theta = 0.008$). *ThetaMater* assumes that all loci are genetically unlinked, and at the request of a reviewer, we conducted a simulation of human chromosome 1 to evaluate the effects of linkage on *ThetaMater* estimates (See Step 9: "Linkage & *ThetaMater*"). Under extreme scenarios of linkage, *ThetaMater* appeared to be biased towards larger values, but for more realistic conditions, *ThetaMater* appears to be robust

to linkage. Nonetheless, these are complex subjects, and we recommend users to explore all potential violations of the model (including selection and recombination/linkage) prior to using *ThetaMater*.

As estimates from any one locus entail significant uncertainty, *ThetaMater* allows researchers to take full advantage of large, genomic datasets when estimating and provides a distribution of plausible values for parameter estimates while accounting for uncertainty. Users can also use an estimate of the shape of among-locus rate variation (*ThetaMater.M1*) or estimate the shape of among-locus rate variation (*ThetaMater.M2*) to account for among-locus rate variation when estimating , as well as characterize the genomic landscape of rate variation. The posterior predictive simulator included in *ThetaMater* allows users to identify potential outlier loci from the genomic distribution of genetic variation, whether due to issues of orthology (see Step 7), or other violations of model assumptions, such as selection (see *GppFst* R package, Adams 2017). *ThetaMater* also includes several functions for simulating datasets under the neutral coalescent model. Briefly, datasets are simulated under the infinite-sites model of mutation according to the protocol described in Wakeley 2008 (pg. 255).

Users can estimate locus-specific s for each locus within a dataset to characterize among-locus estimates of θ, or leverage all loci to estimate a single, population-wide estimate. For single locus-based estimates, reflects the time to the most common ancestor among a sample of sequences. This is because the average time for 2 copies to reach a common ancestor is equal to 2N generations (~4N generations for larger sample sizes). Thus, users can characterize differences in TRMCA (locus-specific) among loci for a number of different applications, such as understanding what evolutionary processes may be at work across the genome. For example, a

short TMRCA (i.e., small effective population size) may indicate the effects of positive selection, while an older TMRCA (i.e., large effective population size) may indicate balancing selection (or other processes).

Linkage & ThetaMater

ThetaMater assumes that all loci are genetically unlinked (i.e., free recombination between loci). At the request of a reviewer, we conducted a simulation analysis to evaluate the effects of linkage on posterior estimates derived via ThetaMater. We simulated a sample of human chromosome 1 (length = 248,956,000bp) with three different recombination rates (2e-8, 2e-9, 2e-10) and a sample size of 10 individuals. We also simulated a single dataset without recombination at all (i.e., $\rho = 0$, such that the entire chromosome was linked). We randomly sampled 1000bp loci every 100kb, with resembles "reduced-representation" sampling, such as RADseq and sequence capture data. We used the following command in msprime: `msprime.simulate(sample_size=10, Ne=10000, length=248956000, recombination_rate=`ρ`, mutation_rate=2e-8)` The results of these analyses are plotted below for each recombination rate. We included a python script (SimulateChr1.v2.py) to generate these results. As you can see, in all cases with recombination (ρ = 2e-8, 2e-9, 2e-10), the posterior distribution of was centered near the true simulated value ($\theta\theta$ = 0.0008), suggesting that *ThetaMater* is likely robust to linkage in these conditions. However, in the most extreme simulation in which $\rho = 0$ (no recombination at all), we did find that *ThetaMater* was biased towards a larger value than the true simulated value. Under extreme scenarios of recombination (all loci are genetically linked), *ThetaMater* may be biased, but under realistic conditions, *ThetaMater* appears robust to linkage. These simulations are not necessarily

conclusive for all scenarios, and we encourage users to explore all potential violations of the coalescent model prior to using *ThetaMater* (i.e., recombination, linkage, selection). If there is some concern for model violations, one can simulate datasets (as we have done with *msprime*) to explore other potential violations, including linkage and selection using similar approaches to those presented here.

Chapter 3

Investigating the impacts of positive selection on coalescent-based species tree estimation and delimitation

Richard H. Adams[1], Drew R. Schield[1], Daren C. Card[1], and Todd A. Castoe[1]

[1]Department of Biology, 501 S. Nedderman Dr., The University of Texas at Arlington, Arlington, TX 76010, USA

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 datasets 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 datasets.

Introduction

Multispecies coalescent models provide a valuable parameterization of the evolutionary processes that underlie neutral divergence between reproductively isolated lineages (Rannala and Yang 2003a; 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 2009a; 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 datasets (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. 2009b). 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 2003a; 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 to 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 2003a; Yang and Rannala 2010; Zhang et al. 2011a; Springer and Gatesy 2016), primarily because they violate the core assumption of the model. Regardless, it is likely that large, multilocus datasets may include some proportion of loci that have evolved under selection, and while it may be logical to filter away such loci from empirical datasets, 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? Here 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 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? (3) What practical concerns do positively-selected loci pose to analyses of empirical datasets?

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, dataset 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 (i) simulating genealogies (with and without selection) using the program MSMS (Ewing and Hermisson 2010)

(ii) simulating 1,000 base DNA sequence alignments under the JC69 model (Jukes and Cantor 1969) on simulated genealogies, (iii) conducting Bayesian species tree estimation and species delimitation using BPP (Yang and Rannala 2010) for each simulated dataset, and (iv) 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 multispecies coalescent 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. 2011a). 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 μ = 2.5 x 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 x 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. 2011a). For our simulations, we tested a total of 9 different 3-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 (τ_{ABC} = 0.0001), moderate-depth (τ_{ABC} = 0.001), and deep species tree models (τ_{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 ($\tau_{AB} = \tau_{ABC}$ x 0.1), medium ($\tau_{AB} = \tau_{ABC}$ x 0.5), and ancient divergence ($\tau_{AB} = \tau_{ABC}$ x 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. 2011a).

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 $2N_e s$ and , 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 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 towards 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} = -

20000), and very strong (s_{aa} = -100000) 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 : "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 x 10^{-8}. We set the starting time of selection to occur immediately after the divergence of Species-A and Species-B (), 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 dataset sizes (1-locus, 2-loci, and 10-loci) and varied the proportion of selected loci within these datasets: 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 datasets (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 (Fig. 2-3). 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 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 1,000bp alignments under the JC69 model (Jukes and Cantor 1969) for all simulated datasets. For the genealogies that experience positive selection, our approach effectively models a 1,000bp 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 datasets 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. 2011a). 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 datasets, 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 towards 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 to 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 5 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 to neutral loci, yielding patterns of differentiation that are incorrectly interpreted as greater lineage divergence when compared to 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. 3b). For example, although the divergence time between Species-A and -B (τ_{AB} = 0.00001) was two orders of magnitude lower than the divergence with Species-C (τ_{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. Fig. 3b).

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 towards 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 to the true simulated values and neutral estimates (Fig. 4 and Fig. S1-S2). We also observe a slight increase in population size parameter estimates of the sister taxon, Species-

B (θ_B) in some analyses (Fig. S1). 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 Fig. S1). Selection can bias estimates of τ_{ABC} and τ_{AB} towards larger values, meaning that datasets including loci under selection lead to incorrectly older estimates of speciation times when compared to neutral datasets. While strong selection at multiple loci can substantially bias parameters inferred from 2- and 10-locus datasets, 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, Fig. S1 d-f). Under the most extreme conditions explored in which 100% of loci in 10-locus datasets 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 2,100 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 to 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 towards smaller values (Fig. 4 and Fig. S1). 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 datasets 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 (Figs. 4 and S1; light gray shading).

Deep species trees – Our results suggest that selection has little influence over parameter estimates for deep species tree models ($\tau_{ABC} = 0.01$ Fig. S2), except when $\tau_{AB} = 0.001$ and only θ_A appears to be susceptible to strong selection (Fig. S2a-c, blue lines). 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 datasets (100% of loci under very strong selection), the parameter estimates are nearly identical to neutral inferences when $\tau_{AB} \geq 0.005$ (Fig. S2, black and red lines).

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}). Here, 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 towards 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 to have minimal influence over posterior probabilities (Fig. 5, light gray) and we often find little difference between estimates obtained from strictly neutral datasets and those inferred from datasets 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 ($\tau_{ABC} = 0.0001$), particularly when multiple loci have evolved under strong selection (s = 0.10, 0.50; Fig. 5 and Fig. S3). 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, 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. 5 a, 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 datasets (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 ($\tau_{AB} = 0.00001$ Fig. 5c, f, and Fig. S3a-c). We also find a small, but measurable increase in P_{BC} in several analyses (Fig. S3 d-f). 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 (Fig. S3, light gray vs. dark gray).

Moderate-depth species trees – Our results indicate that selection can also influence species delimitation on moderate-depth species trees under some conditions ($\tau_{ABC} = 0.001$; 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 to 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 to neutral locus datasets. 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 single and 2-locus analyses (τ_{AB} = 0.001, blue line; Fig. S4). In these scenarios, we find that even weak selection can increase P_3, P_A, and P_B when compared to 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 (Fig. S4).

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 to 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 towards 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 S5a). 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 $\tau_{AB} = 0.00001$ (Fig. S5a). When sampling is increased to 20, P_{BCA} is further increased to over 2.5x (0.608) that of neutral inferences (0.242) for datasets 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 ($\tau_{AB} = 0.00001$; Fig. S5a).

We observed similar effects of selection for species trees inferred from 2-locus datasets (Fig. S5a). 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 datasets may also be biased towards 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 towards the incorrect topology for moderate-depth species trees ($\tau_{ABC} = 0.001$; Fig. 6b and S5b). As with shallow species models, the effects of selection on P_{BCA} and P_{ABC} increase with the strength of selection, number 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 to shallow species trees (Fig. 6a vs. 6b). 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 5 haplotypes are sampled per species (Fig. S5b). We observed similar trends for 2-locus datasets, 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 datasets, our results suggest that species tree estimates can be strongly biased towards 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 datasets 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 (Fig. S6). 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 et al. 2009a, 2016a), 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 (here, 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 to 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 to 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 to inferences based solely on neutral loci. We also observed a slight increase in P_{BC} in some cases; this effect was relatively weak compared to 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 datasets and datasets 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 datasets were the same or nearly the same

as the prior probabilities (i.e., $P_3 = 1/3$), generally highlighting the need for larger datasets 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 example genealogy in Fig. 1). Therefore, selection can bias species tree inferences towards 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 S5). 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 datasets 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 (Fig. S6).

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 datasets. 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 datasets 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 towards 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 (Edwards et al. 2009, 2016b).

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 3-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. 2009b), 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. 2017a); 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 multilocus datasets: 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.

Acknowledgments

Support was provided from startup funds from the University of Texas at Arlington to TAC, NSF grants to TAC (DEB-1655571), TAC and DCC (DEB-1501747) and TAC and DRS (DEB-1501886), and Phi Sigma Support to RHA. Additionally, both the Lonestar and Stampede compute systems of the Texas Advanced Computing Center (TACC) were utilized for these analyses. We thank Matthew Fujita and Adam Leaché for valuable insights and helpful discussions.

Figures

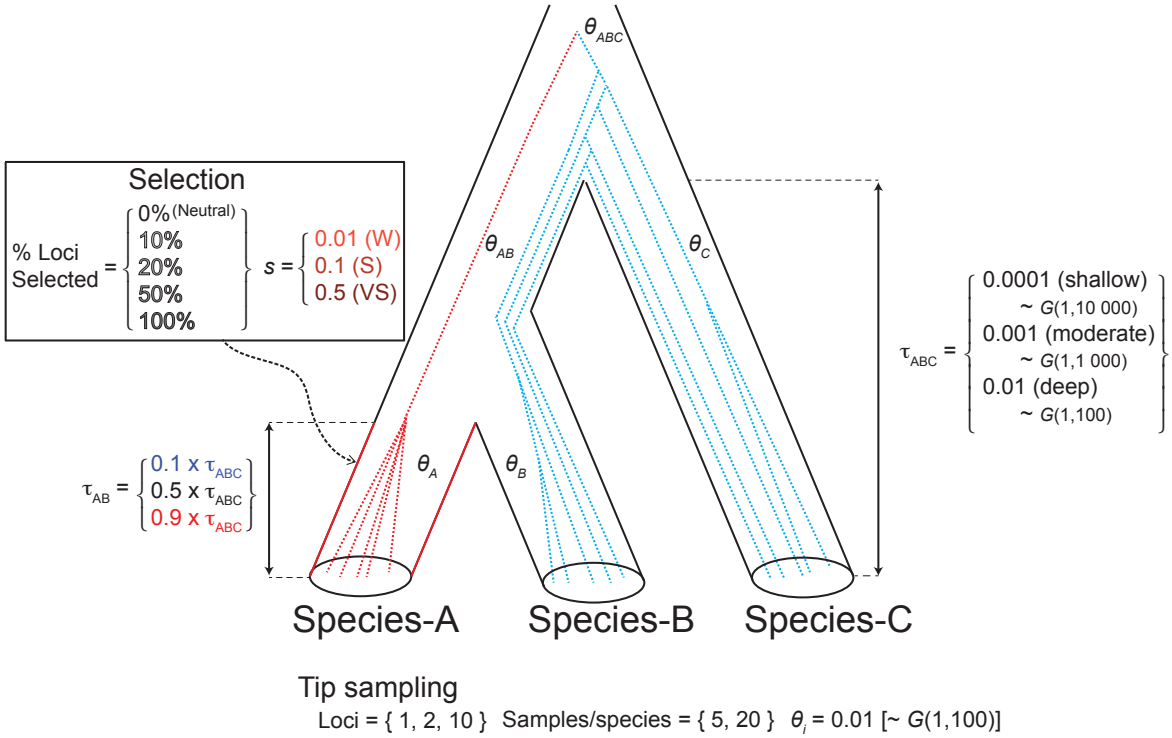


Figure 1. Species tree and experimental parameters used for simulating genealogies under the multispecies coalescent model both with and without selection. Dotted lines within the species tree represent an example genealogy in which a selective sweep has occurred in Species-A lineages immediately after 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.

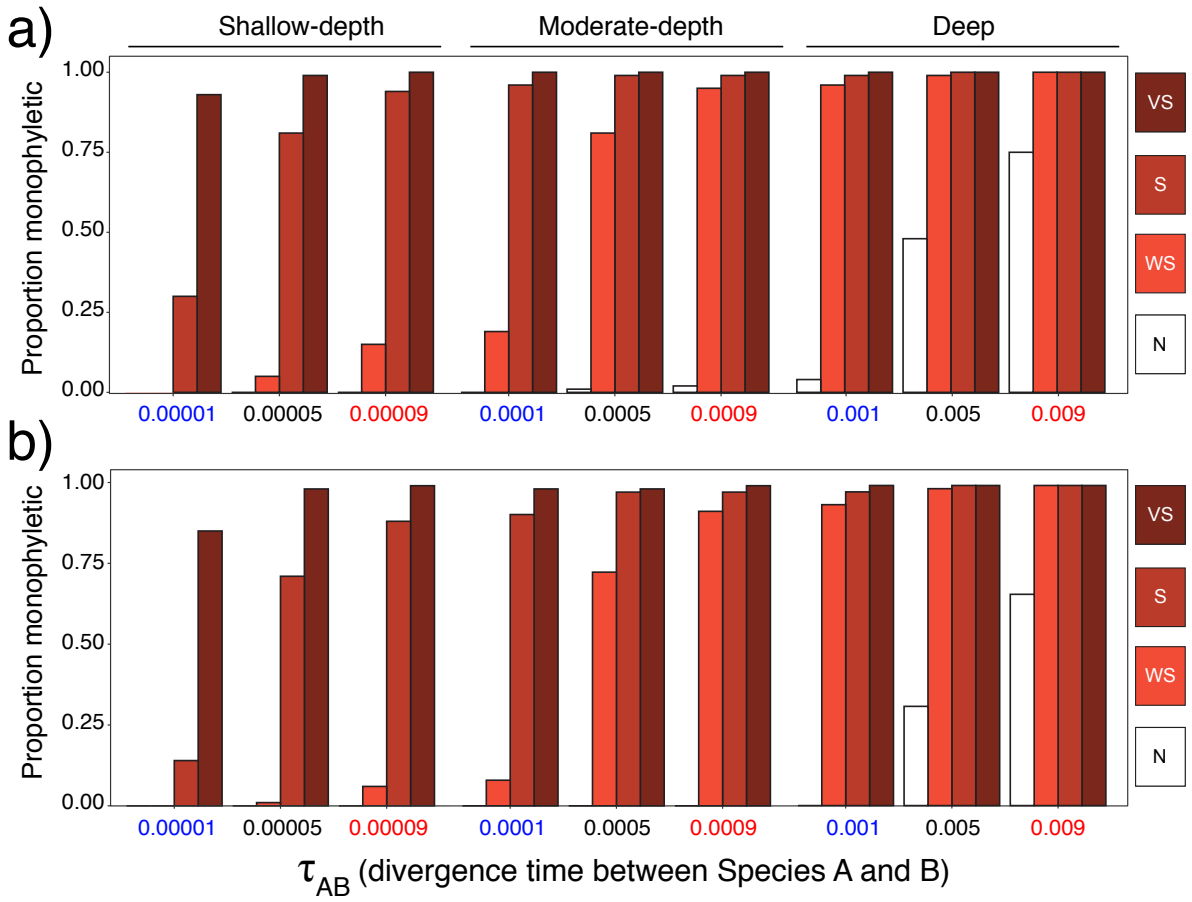


Figure 2. The impact of species-specific positive selection on gene tree distributions and the probability of monophyly. Barplots indicate the percentage of simulated genealogies with monophyletic relationships for all Species-A lineages (i.e., all Species-A lineages reach a common ancestor before coalescing with any outgroup lineages, see example genealogy shown in Fig. 1). Results are shown from left to right for the shallow (= 0.0001), moderate-depth (= 0.001), and deep (= 0.01) species tree models and for each respective Species-AB divergence: ×0.10, ×0.50, and ×0.90. For each simulation model and associated parameters, we simulated 104 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.

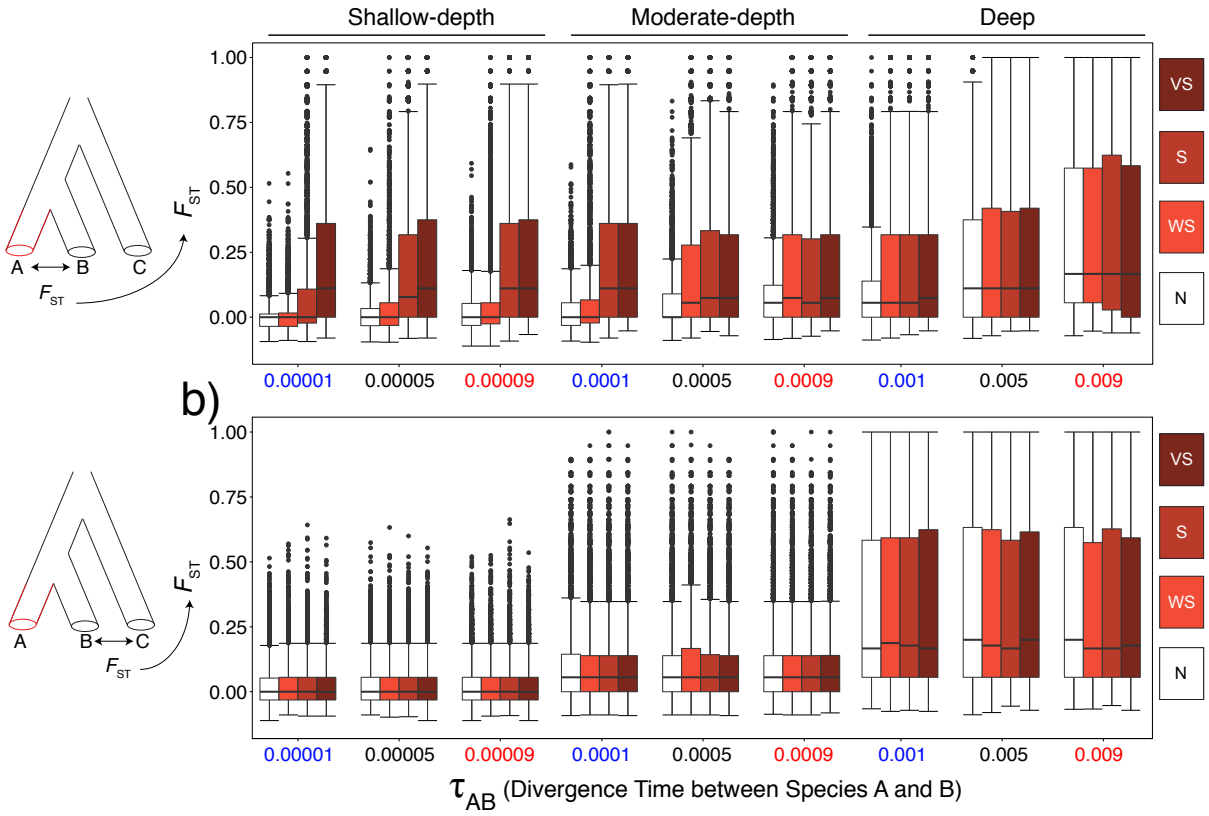


Figure 3. The effects of species-specific positive selection on measures of population differentiation. Boxplots represent the distribution of FST estimates between Species-A and Species-B (a) and Species-B and Species-C (b) across 104 simulated alignments with 20 haplotypes sampled per species. Results are shown from left to right for the shallow (= 0.0001), moderate-depth (= 0.001), and deep (= 0.01) species tree models and for each respective Species-AB divergence (): ×0.10, ×0.50, and ×0.90. For each simulation model and associated parameters, we simulated 104 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.

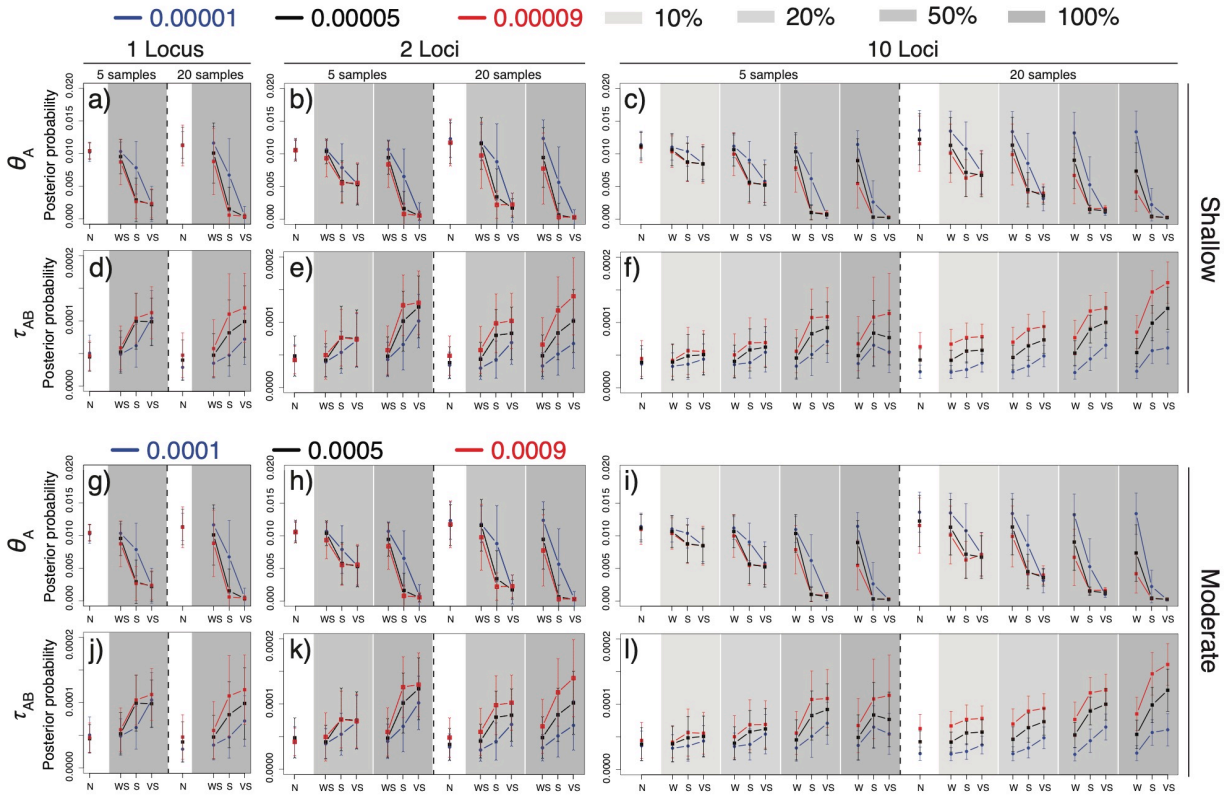


Figure 4. Selection can decrease estimates of and inflate divergence time estimates () for both the shallow (top) and moderate depth (bottom) species model. Results are shown for (a–c and g–i) and (d–f and j–l) for simulated data sets consisting of 1-locus (a, d, g, j), 2-loci (b, e, h, k), and 10-loci (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: = 0.00001, 0.00005, and 0.00009 for the shallow species model (top) and = 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.

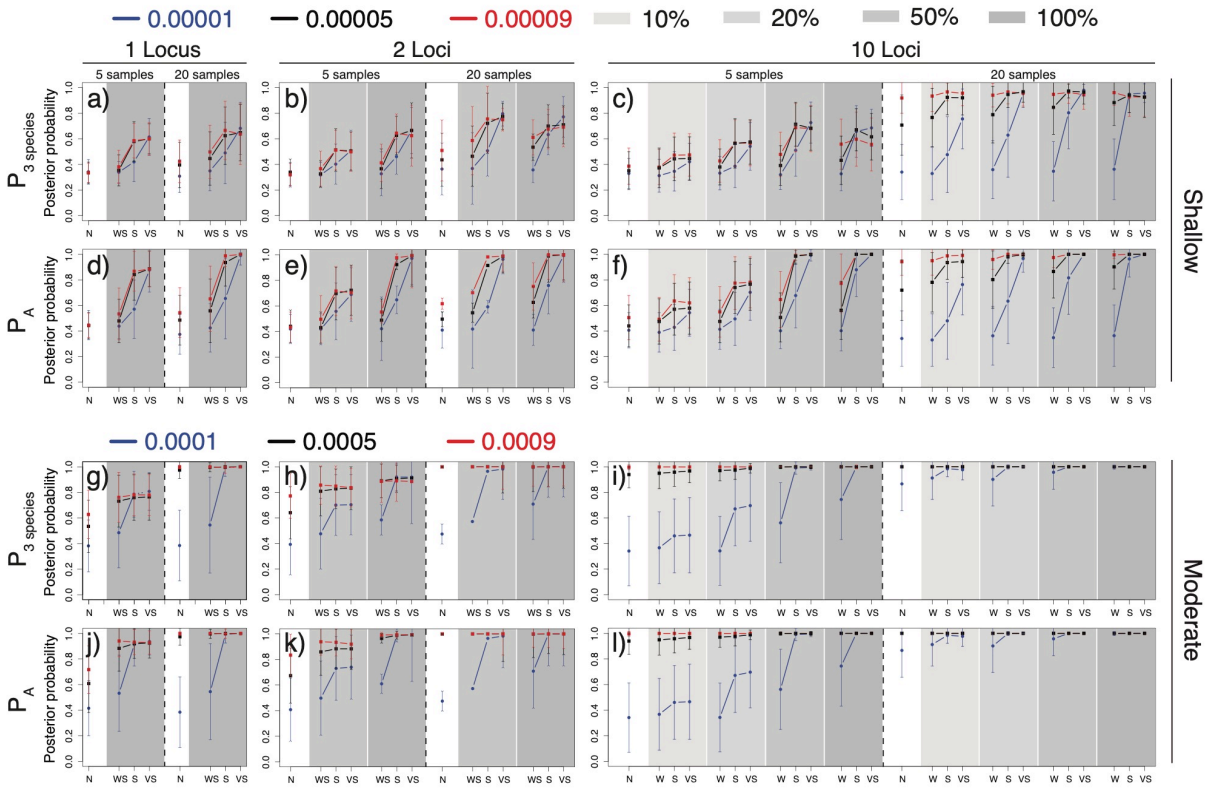


Figure 5. Selection can increase posterior probabilities of species hypotheses for the shallow (top) and moderate-depth (bottom) species tree model. Results are shown for the probability of three species (; a–c and g–i) and the probability of Species-A (; d–f and j–l).

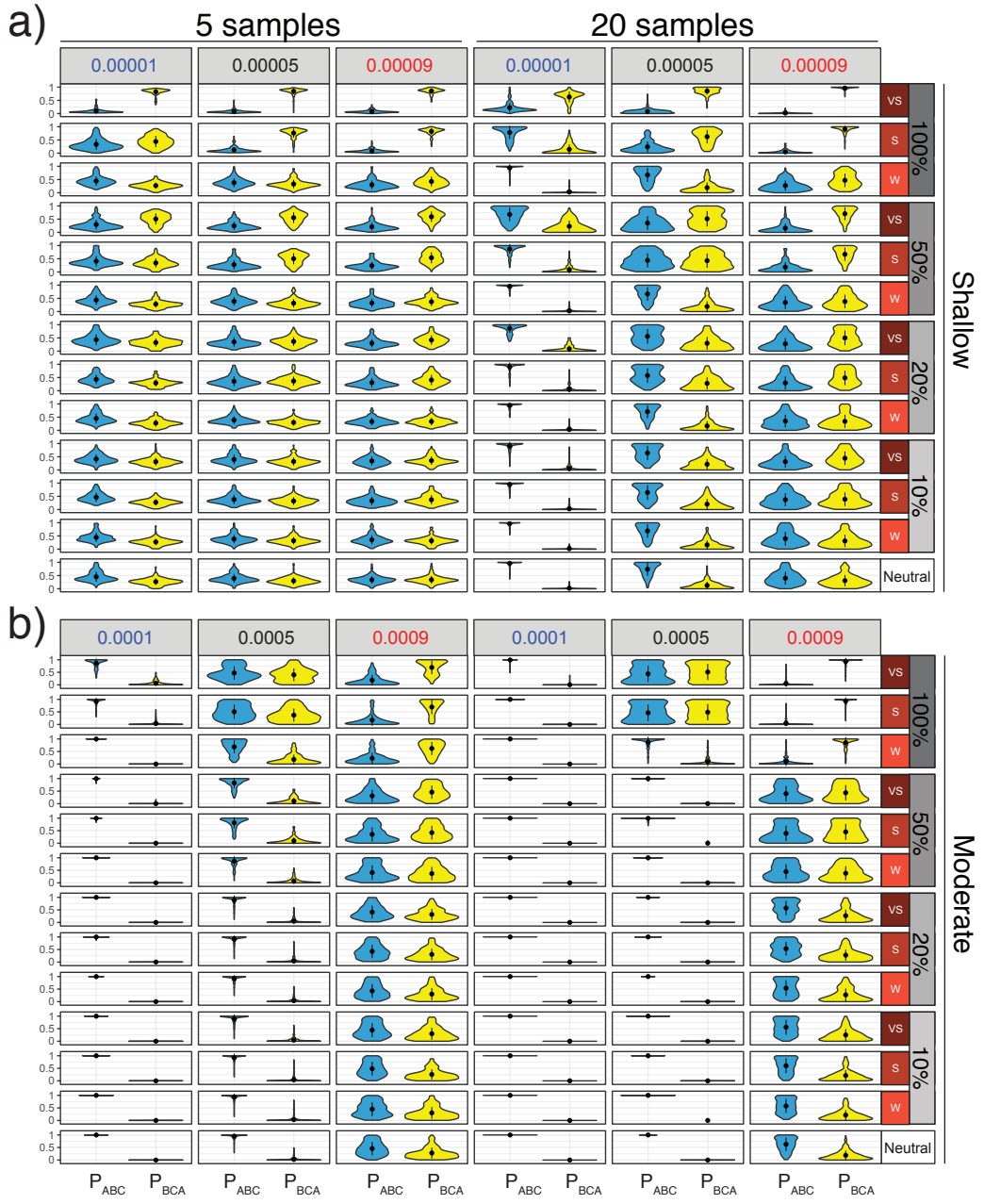
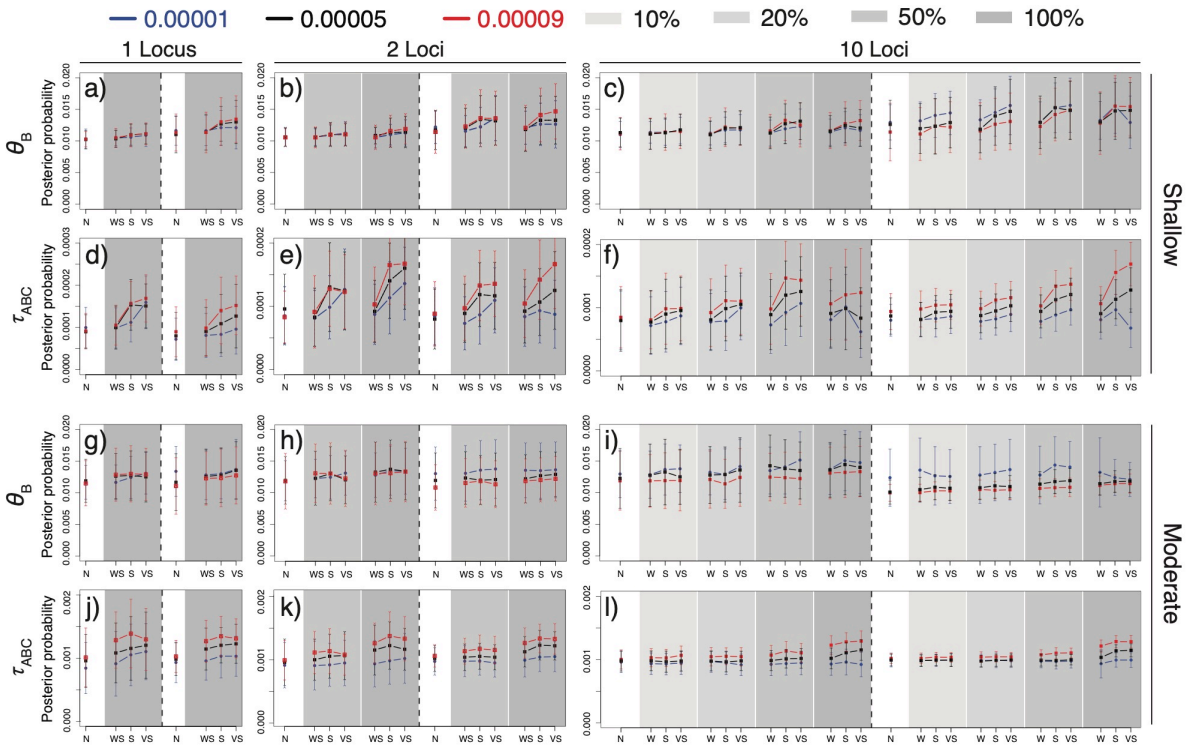
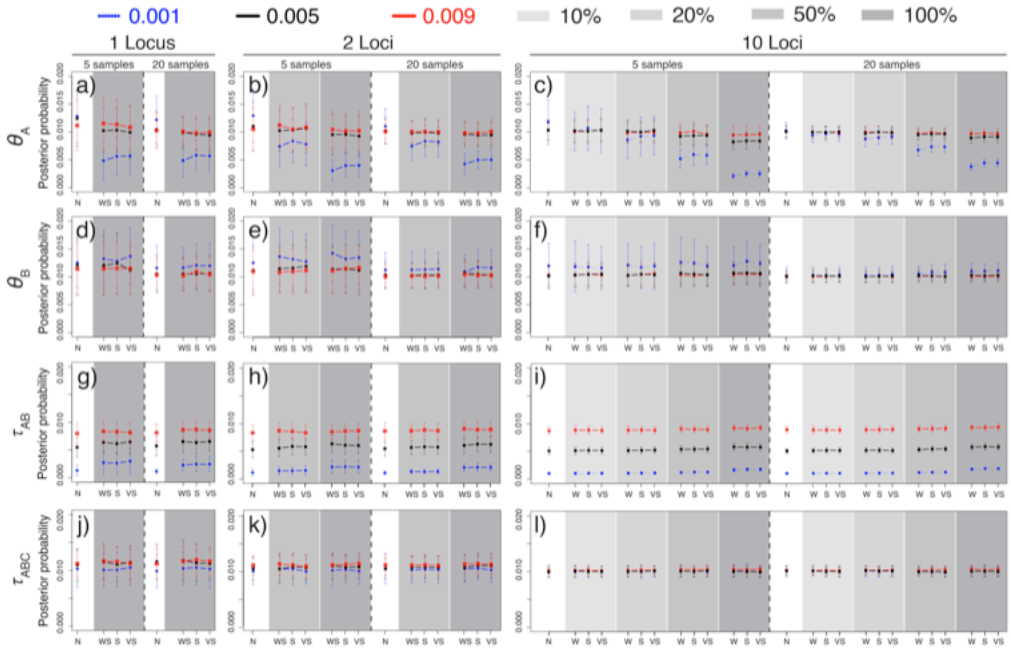


Figure 6. Species-specific positive selection can bias species tree estimates of shallow(a) and moderate-depth (b) species models. Violin plots show the distribution of posterior probabilities of the correct rooted species topology () and incorrect topology () across 200 replicates (mean shown in black) 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): =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.

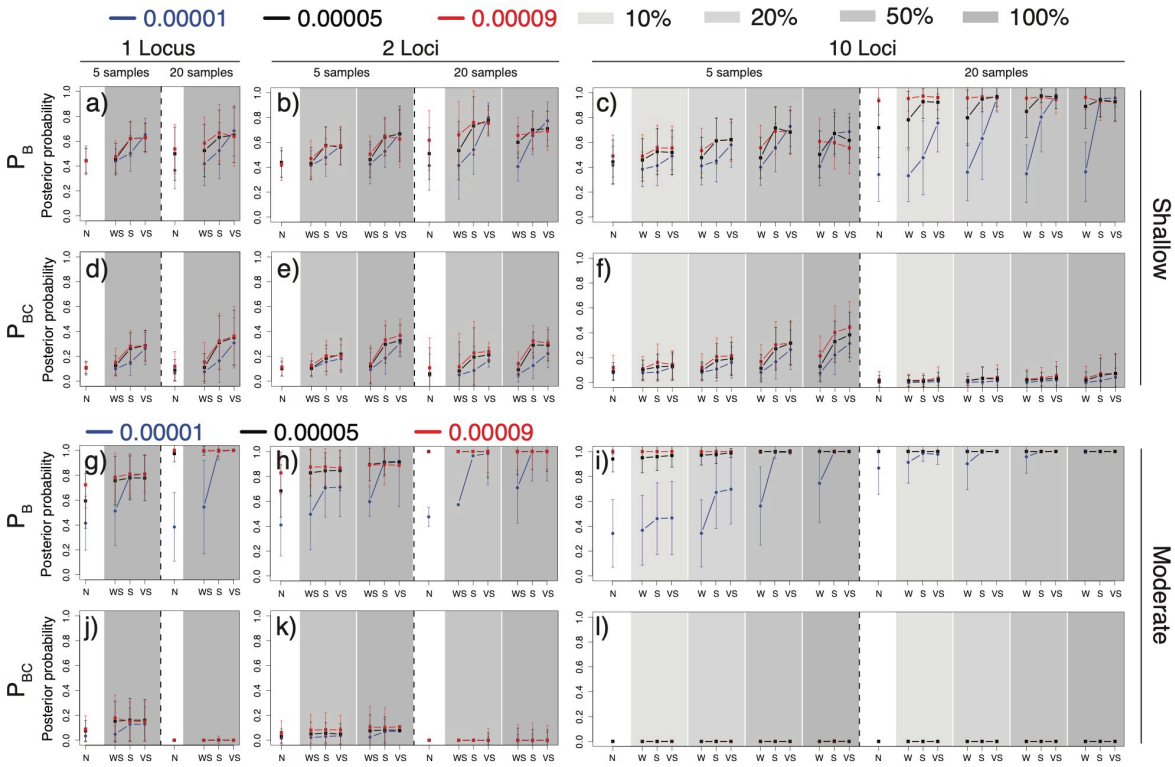
Supplementary Figures



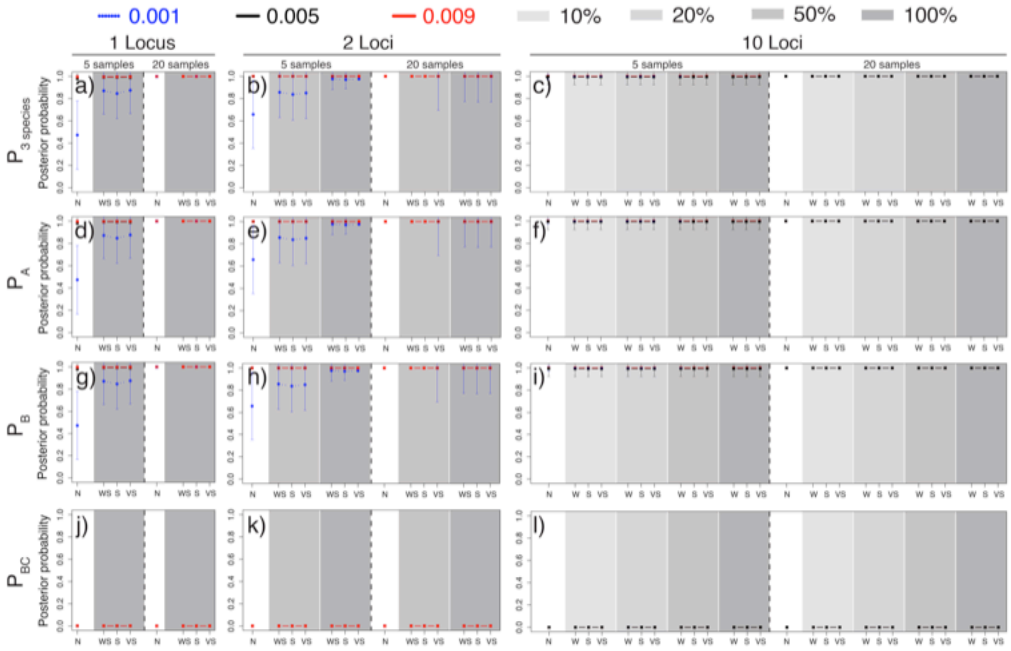
Supplementary Figure 1. The effects of species-specific positive selection on effective population size and divergence time estimates of the shallow (top) and moderate-depth (bottom) species tree models. Results are shown for (a-c), (d-f), (g-i), and (j-l) for simulated datasets consisting of 1-locus (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: recent (blue), medium (black), and ancient (red). 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.



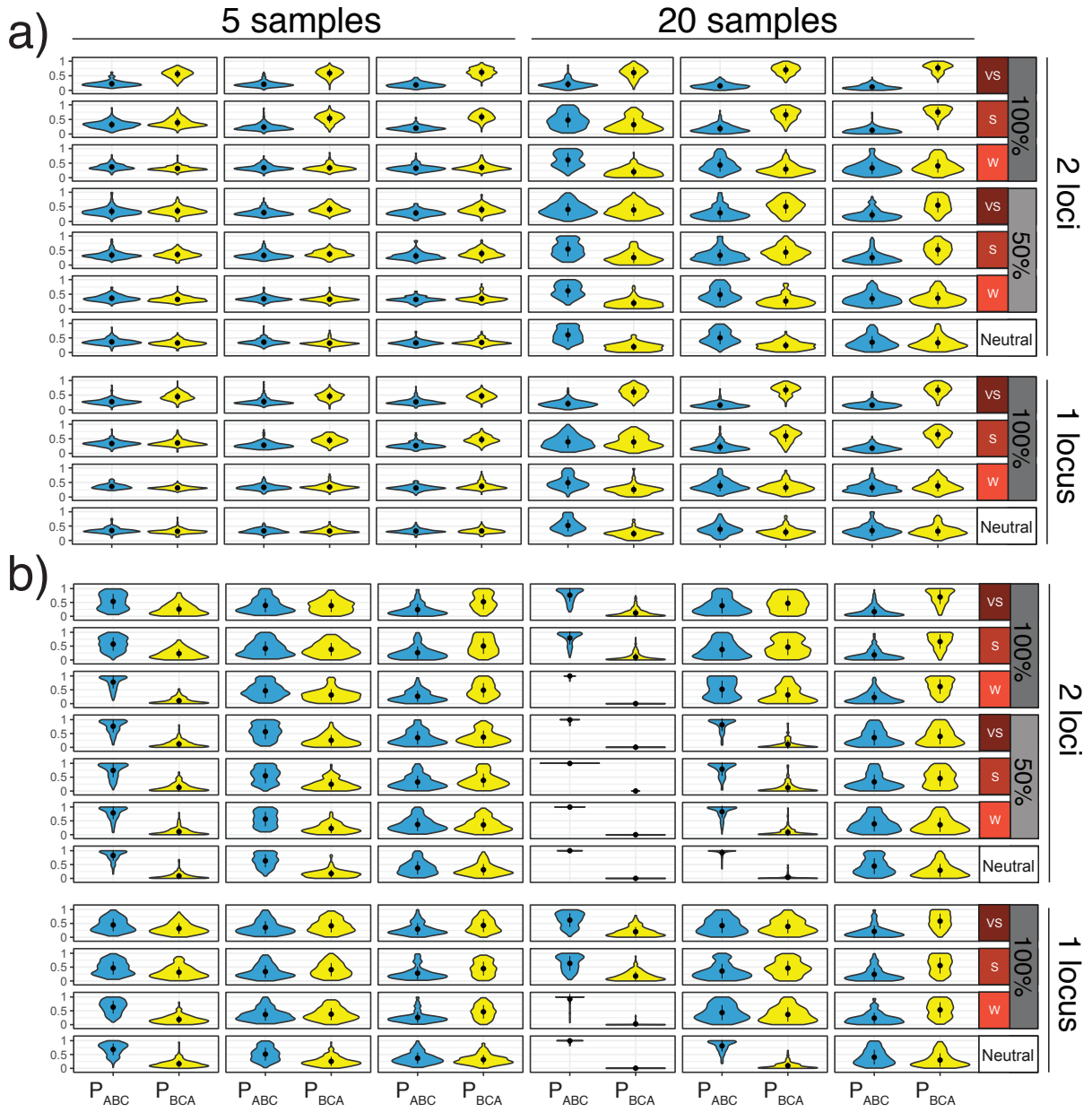
Supplementary Figure 2. The effects of species-specific positive selection on effective population size and divergence time estimates of the deep species tree model. Results are shown for (a-c), (d-f), (g-i), and (j-l) for simulated datasets consisting of 1-locus (a, d, g, j), 2-loci (b, e, h, k), and 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: 0.001 (blue), 0.005 (black), and 0.009 (red). 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.



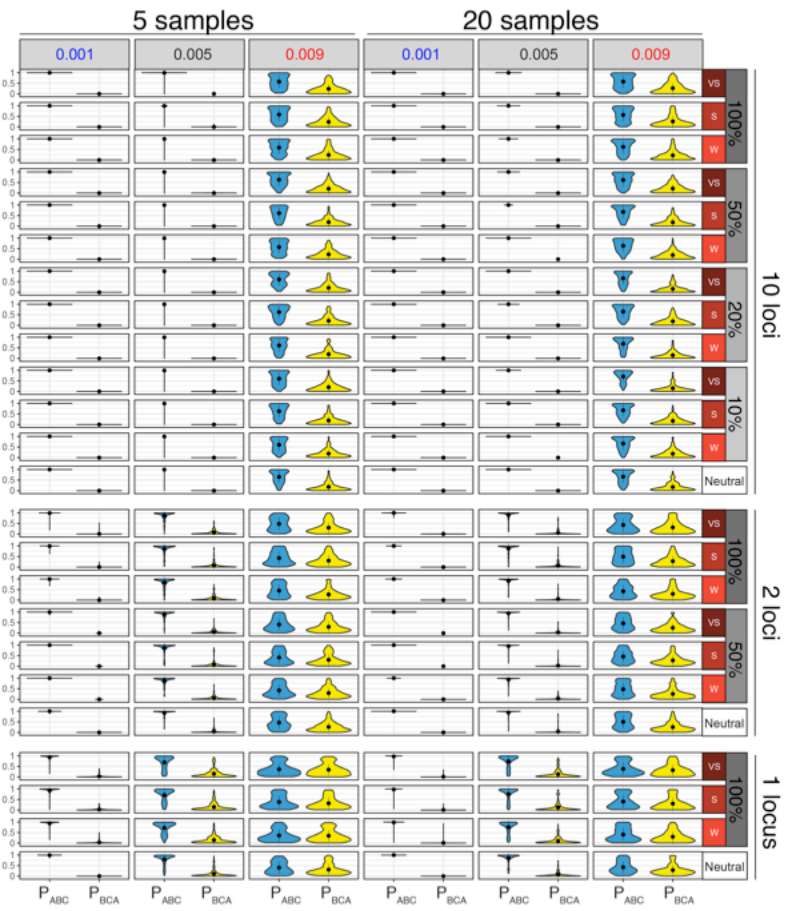
Supplementary Figure 3. The effects of species-specific positive selection on posterior probabilities of species hypotheses of the shallow and moderate-depth species tree models. Results are shown for simulated datasets consisting of 1-locus (a, d, g, j), 2-loci (b, e, h, k), and 10-loci (c, f, i, l). The mean (points) and standard deviation (error bars) of parameter estimates based on 200 replicates are shown for three different Species-AB divergence times: recent (blue), medium (black), and ancient (red). 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.



Supplementary Figure 4. The effects of species-specific positive selection on posterior probabilities of species hypotheses of the deep species tree model. Results are shown for simulated datasets 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: 0.001 (blue), 0.005 (black), and 0.009 (red). 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.



Supplementary Figure 5. The effects of selection on species tree estimates for the moderate species tree simulations. Violin plots show the distribution of posterior probabilities of the correct rooted species topology (P_{ABC}, blue) and incorrect topology (P_{BCA}, yellow) across 200 replicates (mean shown in black) for datasets 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): 0.0001 (blue), 0.0005 (black), and 0.0009 (red). 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, light red), strong ("S", s = 0.10, medium red), and very strong ("VS", s = 0.5, dark red) selection coefficients.



Supplementary Figure 6. The effects of selection on species tree estimates for the deep species tree simulations. Violin plots show the distribution of posterior probabilities of the correct rooted species topology (P_{ABC}, blue) and incorrect topology (P_{BCA}, yellow) across 200 replicates (mean shown in black) for datasets 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): 0.001 (blue), 0.005 (black), and 0.009 (red). 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, light red), strong ("S", s = 0.10, medium red), and very strong ("VS", s = 0.5, dark red) selection coefficients.

Chapter 4

Statistical binning leads to profound model violation due to gene tree error incurred by trying to avoid gene tree error

Richard H. Adams[a], and Todd A. Castoe[a,*]

[a]Department of Biology, 501 S. Nedderman Drive, University of Texas at Arlington, Arlington, TX 76019 USA

Abstract

Fundamental to all phylogenomic studies is the notion that increasing the amount of data – to entire genomes when possible – will increase the accuracy of phylogenetic inference. Simply adding more data does not, however, guarantee phylogenomic inferences will be more accurate. Even genome-scale reconstructions of species histories can suffer the effects of both incomplete lineage sorting (ILS) and gene tree estimation error (GTEE). Weighted statistical binning was originally proposed as a technique to assist the avian phylogenomics project in solving the bird tree of life, which has long eluded resolution as a result of both ILS and GTEE. These so-called "statistical binning procedures" seek to overcome GTEE by concatenating loci into longer multi-locus "supergenes" that are used to reconstruct a species tree under the assumption that the supergene tree set is an accurate estimate of the true underlying gene tree distribution. Here we evaluate the performance of the method using the original avian phylogenomics dataset. Our results suggest that statistical binning constructs false supergenes that concatenate loci with different coalescent histories more often than not: >92% of supergenes comprise discordant loci. Our results underscore a major logical inconsistency: GTEE – the sole justification for using statistical binning instead of standard concatenation – also makes these methods unreliable. These findings underscore the need for developing new robust frameworks for phylogenomic inference that more appropriately accommodate GTEE and ILS at a genome-wide scale.

1. Introduction

Much of our understanding and practice of evolutionary biology relies on knowledge of the species-level relationships of organisms (i.e., species trees). Two major sources of phylogenetic conflict can pose serious challenges for species tree reconstruction: incomplete lineage sorting (ILS) and gene tree estimation error (GTEE). Standard phylogenetic analysis of concatenated loci, for example, will be statistically inconsistent in the presence of ILS and yield highly-supported but incorrect species trees (Edwards et al. 2007; Kubatko and Degnan 2007). To address this, coalescent-based methods have been developed that are statistically consistent under ILS and will return the true species-level phylogeny with high confidence given sufficient information (Liu 2008; Degnan and Rosenberg 2009a; Knowles 2009; Heled and Drummond 2010; Liu et al. 2010b, 2015b). While ILS is an inherent property of the demographic processes of speciation and divergence, GTEE is a fundamentally different source of conflict that represents statistical sampling error and variation between the true tree and one estimated from a dataset of finite size and information content. Although modern phylogenomic datasets often consist of millions to billions of base pairs (bp), any one aligned locus is often limited to <3kbp of aligned orthologous sequence data, and thus individual gene trees may entail substantial error that can permeate to the level of species tree inference (Jarvis et al. 2014; Mirarab et al. 2014). Researchers thus face a gauntlet of challenges when analyzing phylogenomic data: concatenate loci and suffer the consequences of ILS, or do not concatenate loci and suffer the consequences of GTEE. Both sources of conflict can have major debilitating effects on the accuracy of species tree estimates, and it is not immediately clear whether one should prioritize either.

The avian Tree of Life is a prime example of an important vertebrate phylogeny that has long eluded resolution because of both ILS and GTEE (Mirarab et al. 2014; Jarvis et al. 2015; Prum et al. 2015a). In light of the challenges facing phylogenomic analyses, a new method ("weighted statistical binning"; referred to as "statistical binning" hereafter) was originally developed to enable the avian phylogenomics project in resolving the relationships of modern birds (Mirarab et al. 2014; Bayzid et al. 2015; Jarvis et al. 2015). The method has since been used to infer the evolutionary relationships of placental mammals (Tarver et al. 2016), teleost fishes (Malmstrøm et al. 2016), and many other major radiations (i.e.., Blaimer *et al.* 2016; Branstetter *et al.* 2017; JeŠovnik *et al.* 2017; Platt *et al.* 2018). The core justification behind this approach is to infer a set of "supergenes" that attempt to overcome GTEE by concatenating smaller sets of individual loci into longer supergene alignments comprising multiple loci that contain more information for inferring supergene trees. In practice, supergenes inferred via statistical binning are often used to obtain a set of supergene trees for downstream species tree estimation under the assumption that that they are 100% accurate. Importantly, gene tree estimates and associated bootstrap support values are used as input data for the statistical binning pipeline as the sole criteria for deciding whether the respective loci within a putative supergene evolved under the same tree (Bayzid et al. 2015). Using a compatibility graph based on these estimates, the pipeline effectively conducts a hypothesis test to decide whether several individual loci can be concatenated to form a supergene (i.e., they share a common topology) or not (i.e., do not share a topology; Mirarab *et al.* 2014; Bayzid *et al.* 2015). Accordingly, the fundamental purpose of statistical binning is to infer which phylogenetic conflicts among estimated gene trees are simply a result of GTEE (result: concatenate to form a supergene), and which conflicts represent true differences in coalescent history due to ILS (result: do not concatenate and estimate distinct trees).

Following publication of the avian phylogenomics project, substantial debate and contention has arisen over the use of statistical binning and similar methods (Bayzid et al. 2015; Jarvis et al. 2015; Liu and Edwards 2015; Mirarab et al. 2015; Roch and Warnow 2015; Warnow 2015). Authors have continued to argue both for and against these methods, and disagree over the statistical consistency (or lack of) of these approaches in the context of species tree estimation (Liu and Edwards 2015; Mirarab et al. 2015; Roch and Warnow 2015; Warnow 2015). A follow-up study revealed that statistical binning distorted supergene tree distributions and likely biased species tree estimates (Liu and Edwards 2015). Further studies corroborated this assertion: species trees reconstructed using supergenes obtained via statistical binning were likely to be highly inaccurate yet highly supported (Streicher et al. 2018). Subsequent response papers rejected the assertion that the method was statistically inconsistent, and instead argued for statistically consistency when the number of loci and the length of loci are both infinite (Bayzid et al. 2015; Mirarab et al. 2015). However, recent theoretical work has demonstrated the inconsistency of species tree methods that use supergenes inferred via statistical binning when the number of loci is unbounded but the length of each locus is bounded to a constant (Roch et al. 2018). These findings raise important questions about the nature of species tree inference under best-case scenarios (i.e., when the number and/or length of loci is infinite), and yet, we currently have relatively little understanding of the empirical performance of the statistical binning pipeline itself when both the number and length of loci are bounded.

When considering the properties of the method, it is imperative to acknowledge that the statistical binning pipeline itself only infers a set of supergene alignments, not a species tree. Statistical binning is therefore *not* a species tree estimation method *per se*, it is a supergene

estimation method that uses gene tree estimates to infer topology congruency among loci. Distinguishing between species tree estimation and supergene estimation is critical, because both are fundamentally different statistical problems: species tree estimation seeks a single species-level topology and set of parameters (i.e., divergence times, effective population sizes), while supergene inference involves deciding whether individual loci share the same gene tree or not. In this sense, statistical binning represents the first "cog in the wheel" of the phylogenomic analysis pipeline, which is followed by supergene tree estimation using standard phylogenetic techniques, such as maximum likelihood (ML) analysis, and species tree estimation using coalescent-based summary methods. Understanding whether the statistical binning pipeline provides reliable supergene alignments is therefore paramount to assessing the performance of the method. At the end of a statistical binning analysis, ML-analysis of each supergene is conducted under the assumptions of the standard phylogenetic model. While different supergenes can have different topologies, ML-analysis of the individual supergene alignments assumes that each gene placed within a supergene shares the same coalescent history. Under these conditions (i.e., a "true supergene" containing only congruent genes), standard ML-analysis – which assumes all sites share the same tree (Felsenstein 1981) – will converge with increasing probability to the single, true gene tree as the length of each congruent locus in the supergene increase (Fig. 1, left).

In contrast, if a supergene incorrectly concatenates genes from multiple distinct topologies, standard ML-analysis of this "false supergene" will not converge to the true gene tree set (i.e., one tree for each distinct gene) as the length of each discordant gene increases, because it is restricted to inferring a single best-fit tree. In the right example shown in Figure 1, a false supergene has been constructed by concatenating three genes with conflicting genealogies (red,

purple, green). Even if the length of each of the three genes is infinite, standard ML-analysis will infer only a single supergene tree – instead of the "true" gene tree set comprised of three distinct topologies. Violation of this fundamental assumption of the phylogenetic model (i.e., all sites share the same tree) is of major consequence because it is the underlying cause of the failure of ML-analysis in the presence of ILS (Mendes and Hahn 2017), and can also cause other modeling pathologies and biases, such as SPILS ("substitutions produced by ILS"; Mendes and Hahn 2016). False supergene trees inferred using standard ML-analysis are likely to reflect an amalgamation of phylogenetic signal, such that the gene tree with the most support (i.e., highest number of informative sites) may have disproportionate influence. The overall supergene tree distribution will also likely be distorted as distinct gene trees are effectively "hidden" within false supergenes and may be poorly represented or absent in the set of supergene trees. False supergenes therefore represent profound phylogenetic model misspecification, and the hope is that methods such as statistical binning are able to avoid such sources of systematic bias by inferring accurate supergenes (i.e., Fig. 1 left vs. right).

A critical question therefore remains: how well does statistical binning infer topological congruency (or lack of) from gene tree estimates when attempting to construct true supergenes? Here we evaluate the performance of the method at this core function, and while previous studies have primarily focused on the theoretical properties of the method for species tree inference when aspects of the data are infinite (i.e., number of genes and/or gene lengths are unbounded), we take a decidedly different, model-based approach to understand whether statistical binning provides accurate supergenes or not. We conducted a post-hoc likelihood-based model assessment of statistical binning accuracy using the 14,446 alignments (8,251 exons, 2,516

introns, and 3,679 UCEs) and the corresponding set of 2,021 supergenes inferred for the original avian phylogenomic analyses (Jarvis et al. 2014, 2015). We specifically applied two different likelihood-based tests to characterize the accuracy of supergenes inferred via statistical binning: likelihood ratio tests (LRTs implemented in ConcatePillar; Leigh *et al.* 2008) and SH tests (Shimodaira and Hasegawa 1999). The first approach conducts a series of likelihood-based model tests to evaluate whether the data (i.e., site patterns) of each respective supergene support a single topology or multiple, discordant topologies (Fig. 2, top box). The second method applies Shimodaira-Hasegawa tests (SH test; Shimodaira & Hasegawa 1999) to evaluate whether individual loci placed within a supergene reject the overall supergene tree in favor of a distinct, locus-specific topology (Fig. 2, bottom box). We used the results of the SH-tests to quantify the number of genes with evidence of significant topological congruency within each supergene alignment (i.e., genes that reject the supergene tree likely support a distinct topology). Unlike the statistical binning pipeline, which uses gene tree estimates to infer topological congruency, these two model-based approaches make direct use of the phylogenetic likelihood function by summing over site likelihoods for alternative tree models to validate supergene inferences by testing whether a single tree (i.e., "true positive", Fig. 1, left path) or multiple, distinct trees (i.e., "false positive") are a better explanation of the data (Fig. 1, right path).

2. Methods

2.1 Avian phylogenomic data

We downloaded the 14,446 alignments (8,251 exons, 2,516 introns, and 3,679 UCEs), the inferred supergene assignments for the 14,446 loci (i.e., assignment of each locus to a respective

supergene), and the 2,021 ML supergene trees inferred via statistical binning for the avian phylogenomic analyses (Jarvis et al. 2014, 2015). For our simulation-based assessments of statistical binning accuracy, we downloaded the simulated gene tree sets and their associated inferred supergene assignments that were used in the original avian phylogenomic studies and were based on the estimated avian species tree (Jarvis et al. 2014; Mirarab et al. 2014).

2.2 Likelihood-based tests of statistical binning accuracy

We evaluated the accuracy of each inferred supergene using likelihood ratio tests (LRTs) implemented in ConcatePillar (Leigh et al. 2008) and SH-tests (Shimodaira and Hasegawa 1999) implemented in RAxML v8.0.0 (Stamatakis 2014). First, we used Concatepillar to conduct LRTs to test whether a model consisting of a single topology or a model of multiple distinct topologies was better supported by the sequence data of each supergene based on the difference in log-likelihood scores between models (Fig. 2, top box). This approach effectively tests how many distinct topologies are supported by the data of each supergene and corrects for multiple comparisons throughout the process. If only a single topology best fits the data, this provides evidence that the supergene is likely to be accurate (i.e., Fig. 1, left). Conversely, if the data support multiple topologies, then the supergene likely violates the phylogenetic model because it exhibits evidence of incorrectly concatenated loci originating from distinct topologies (i.e., Fig. 1, right).

We used SH-tests in a similar fashion to test whether the difference in log-likelihood scores between the ML topology of each individual gene placed within a supergene and the overall ML supergene tree was statistically significant (Fig. 2, lower box). In other words, for each gene

placed within an inferred supergene, we used SH-tests to compare the likelihood of the individual gene-specific ML topology with the overall supergene ML topology (Fig. 2, colored vs. gray trees in lower box). If the individual gene-specific ML tree was a statistically significant better fit than the supergene topology (i.e., $P < 0.05$), then that supergene was likely falsely constructed by statistical binning (i.e., concatenated loci with different phylogenetic histories, i.e., Fig. 1 right). The number of genes that reject the overall supergene tree in favor of a locus-specific tree provide an indication of the number of discordant genealogies present within a supergene alignment. SH-tests were conducted in RAxML 8.0.0 (Stamatakis 2014) using the default GTR+I+**Γ** nucleotide substitution model independently for each locus.

In light of widespread evidence of supergene error (i.e., Fig 3), we were interested in characterizing the degree to which supergene trees reflected the topologies of their constituent genes. A critical concern of concatenating genes into a single supergene is that, if genes do not share the same tree, the gene with the most informative sites will dominate and overwhelm gene tree signals from shorter or less informative genes. In such cases, the supergene tree may only reflect the relationships supported by the dominant genealogy, while conflicting topologies of shorter loci will be effectively "hidden" and likely absent from the supergene tree distribution. To examine whether supergenes tend to be biased towards their longest constituent gene (and therefore capable of masking hidden gene trees from shorter gene constituents), we computed normalized Robinson-Foulds distance between each of the 14,446 gene trees and their associated supergene topology using the R package phangorn (Schliep 2011).

2.3 Simulation-based assessment of statistical binning accuracy

We also evaluated the accuracy of statistical binning on the simulated gene tree sets provided in the original study (Mirarab et al. 2014; Jarvis et al. 2015), by testing whether supergenes inferred via the method included only simulated genes that share a common gene tree. For each inferred supergene, we computed pairwise Robinson-Foulds distances (Robinson and Foulds 1979) between each simulated gene tree that statistical binning inferred to share a single supergene tree; all of the individual gene trees should be identical if statistical binning provided a correct supergene. An RF-distance of 0 between two trees means that the topologies are identical and an RF-distance >0 means the topologies are different. If all gene trees placed within a supergene have an RF-distance of 0, then the supergene was accurately inferred (i.e., Fig. 1, left). If there is at least one RF-distance that is greater than 0, the supergene was inaccurate because it incorrectly concatenated loci that evolved along distinct, conflicting gene trees (i.e., Fig. 1, right). We computed unrooted RF-distances using the "multiRF" function provided in the phytools (Revell 2012) package in R, and used these values to compute the mean RF-distance among gene trees across all inferred supergenes in each replicate simulation analysis (rightmost column of Supplementary Table 1). For reference, these supergenes were inferred in the original study using a bootstrap threshold of 75% (Jarvis et al. 2014).

2.4 Quantifying the impacts of statistical binning on gene tree distributions and species tree support

Considering evidence for spurious supergenes, we explored the impacts of statistical binning on both gene tree distributions and species tree support. To visualize differences in the underlying topological distributions due to statistical binning, we generated Densitree (Bouckaert 2010) plots and summary consensus trees using TreeAnnotator (Rambaut and Drummond 2016) of the

unbinned gene tree and binned supergene tree distributions. We also quantified shifts in species tree support by measuring the difference in multispecies coalescent likelihoods of the unbinned gene trees and binned supergene trees using (1) the "unbinned" species tree (UST) estimated using the unbinned gene trees and (2) the "binned" species tree that was estimated using the binned supergene trees. For each of the 14,667 unbinned gene trees for the avian dataset, we measured the difference between the multispecies coalescent likelihood given the "binned" species tree and separately, the likelihood of the gene tree given the "unbinned" species tree: GeneTreeLnL = LnL(GeneTree|Binned Species Tree) = LnL(GeneTree|Unbinned Species Tree). We also conducted this same analysis for the 2,021 supergenes inferred via statistical binning: SupergeneTreeLnL = LnL(SupergeneTree|Binned Species Tree) = LnL(SupergeneTree|Unbinned Species Tree). To visualize the impacts of statistical binning on species tree support, we compared the distributions of the 14,667 GeneTreeLnLs and the 2,021 SupergeneTreeLnLs.

3. Results and Discussion

3.1 Evidence of widespread model misspecification due to statistical binning

Model-based evaluation of the performance of statistical binning on the avian phylogenomic data indicate that it does not provide reliable supergenes because it is highly prone to constructing "false supergenes" from loci with different coalescent histories – leading to profound and widespread phylogenetic model violation (Fig. 4). Both likelihood-based methods we employed indicate widespread error: 96.0% (1,940/2,021) and 92.3% (1,866/2,021) of supergenes concatenated multiple, conflicting topologies using the LRTs and SH-tests, respectively (Fig. 3a

and 3b). Our results therefore indicate that the vast majority (>92%) of inferred supergenes represent false positives. We further evaluated the accuracy of statistical binning on the simulated datasets provided in the original avian study (Jarvis et al. 2015). Surprisingly, we found that 100% of multilocus supergenes (i.e., supergenes with at least 2 loci) across all simulation models and replicates were falsely constructed by statistical binning (Supplementary Table 1) and represent the right example shown in Figure 2. In other words, we found that the false positive rate of these methods for the avian dataset is ~92.3% at best.

Our analyses collectively suggest that statistical binning fails to overcome GTEE because it, like the methods it was designed to outperform, is based on unreliable gene tree and bootstrap support estimates that themselves suffer from high error, leading to false inferences of topological congruency. In other words, the core hypothesis test implemented in statistical binning, which uses bootstrap thresholds to determine gene tree congruence, does not appear to provide accurate supergene based upon our likelihood-based evaluations. Instead, our results indicate that genes incorrectly placed within these false supergenes exhibit surprisingly high gene tree incongruence, as indicated by mean Robinson-Foulds distances (RF-distance) within supergenes ranging from ~25-49 (Supplementary Table 1). ML-analysis of concatenated data predicts that supergene tree inference should be dominated by the gene with the most informative sites, which was observed in our analysis (Fig. 4). Considering our evidence of widespread supergene error (Fig. 3), evidence of the dominance of the longest gene driving supergene tree estimates suggests that alternative topologies of other, shorter genes within supergenes are likely under-represented or even absent from false supergene trees. At best, this scenario would result in the massive loss of genealogical information due to binning genes into supergenes (i.e., only

the topology from the longest gene is represented). A potential and worse scenario would be that this amalgamation of signal from genes with different genealogies may instead led to totally spurious supergene estimates that do not overlap with any of the true gene trees underlying the data (i.e., unnatural products of signal averaging). These findings also further clarify the underlying reason for the reported distortion of supergene tree distributions resulting from statistical binning (Liu and Edwards 2015), and corroborate recent theoretical work that has shown the inconsistency of statistical binning when the length of each locus is finite (Roch et al. 2018).

To characterize the impacts of statistical binning and potential biases it introduces in gene tree distributions, we compared the distribution of supergene trees with the distribution of locus-specific gene trees. Overlays of gene trees using Densitree illustrate that binning leads to major shifts in the gene tree distributions, including several major decreases in conflict (and increases in gene tree resolution), particularly for more ancient nodes (Figs. 5a-b), consistent with previous evidence that binning 'flattens' gene tree distributions (Liu et al. 2015a). Similarly, comparison of consensus trees between binned and unbinned gene tree sets highlight major differences in gene tree topology and broad increases in bipartition agreement based on binned supergene trees (Figs. 5c-d). Comparisons of likelihood support for alternative species trees indicates that statistical binning introduces major changes in the shape and magnitude of variation of species tree likelihoods (Figs. 5e-f). For example, the number of supergene trees that strongly support one species tree over another increases compared to the unbinned gene trees. Considering evidence that a large proportion of supergenes may be false (e.g., Fig. 3), our results collectively

suggest that statistical binning strongly biases gene tree distributions that do not reflect the true gene tree variation, and thereby provide high support for an incorrect species tree.

Although we have primarily presented the problem of "false supergenes" as a dichotomous phenomenon (i.e., ether all genes are congruent or not), their impacts on species tree estimation may be more complex depending on the particular evolutionary parameters (i.e., species tree shape, divergence times, population sizes), and/or experimental conditions (i.e., number and length of loci). For example, "false supergenes" comprised of only two distinct trees may be less problematic then if they contain loci from three distinct trees. It also seems possible that particular branches and subclades may be more or less accurately estimated than others. This could occur, for example, if most genes within a false supergene agree on the placement of a particular clade. Deeper nodes may be more accurately estimated than more recent species splits – perhaps because individual genes may exhibit little conflict in the placement of more ancient lineages (i.e., most ancient lineages are completely sorted). Nonetheless, ML-analysis of false supergenes will be a forced comprise of the conflicting signal exhibited across incongruent loci and thus, will likely suffer large-scale systematic error in topology, branch length estimates, and other parameters.

4. Conclusions

Perhaps surprisingly, genome-scale datasets do not yet equate to straight-forward and robust resolution of phylogeny. Instead, both biology and methodology continue to pose serious challenges for phylogenomic analyses. There is certainly logical merit in approaches that are designed – at least in theory – to tractably address these issues, such as statistical binning. Our

results, however, suggest that nearly every supergene tree inferred via this approach and used to reconstruct the avian species tree is likely to suffer extensive systematic error at the hands of pervasive phylogenetic model misspecification, such that statistical binning is more likely to suffer the effects of GTEE and ILS than overcome them. Instead, the effects of ILS will be rampant in the set of ML supergenes trees used to estimate a species tree when statistical binning is applied. Because these methods only infer topological congruency and do not estimate a species tree, we also argue that model-based supergene validation of statistical binning inferences (i.e., LRT tests) provides a far more direct assessment of the method at its core function and brings clarity to previous arguments, which primarily evaluated the performance of downstream species tree estimation methods that use supergenes as input data.

These findings raise the question of what alternative strategies would be useful for avoiding these issues? One solution is to simply collect more genetically linked data per locus (i.e., longer orthologus loci) to obtain higher quality gene trees without the need for concatenation. In practice, however, "simply collecting more data" is not always a simple or even viable option, particularly given that the original avian analyses sampled whole-genomes and still faced these issues, in part due to the difficulties in aligning long orthologous regions across deep evolutionary time. Increasing the length of individual loci also has the downside of increasing the probability of intra-locus recombination, which may pose additional complications and violations of the phylogenetic model analogous to those introduced by erroneous supergenes. Indeed, false supergenes exemplify the most "extreme" form of this violation whereby recombination occurs freely between genes with non-congruent histories incorrectly placed within a supergene. Unlike binning approaches that "agnostically" infer supergenes using only

gene tree estimates without taking into account genome structure, it may prove fruitful to make effective use of known genetic linkage among loci to propose the combinability of nearby putatively linked loci and test this inference using model-based approaches. Above all, our findings highlight the critical need for the continued development of more accurate phylogenomic methods that can tractably and reliably deliver more reliable gene trees, and ultimately, better species tree estimates

Acknowledgments

This project was supported by an NSF grant (DEB-1655571) to TAC, computation resources provided by the Texas Advanced Computing Center (TACC), and a University of Texas at Arlington Phi Sigma Society grant to RHA.

Figures

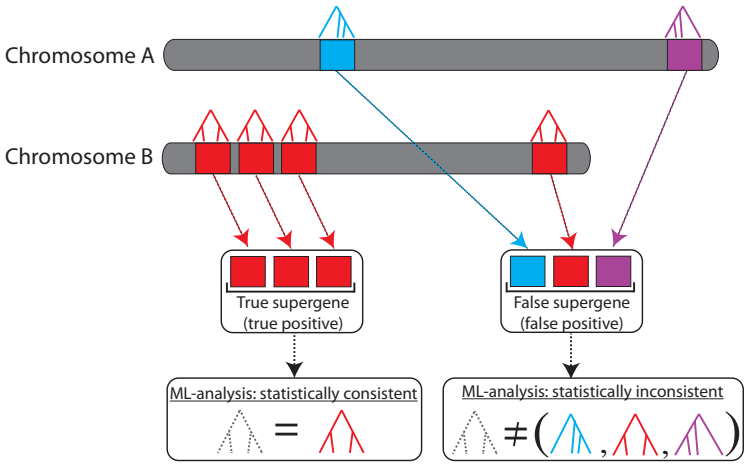


Figure 1. Statistical binning is a supergene estimation method, not a species tree estimation method. Based on similarities (or lack of) among gene tree estimates and bootstrap support values, the core function of the method is to infer whether individual genes share a common genealogy, and if so, concatenate congruent genes to construct longer supergenes. Example indicating loci sampled from two different chromosomes and three distinct gene trees (red, blue, and purple). If statistical binning is accurate, inferred supergenes will only concatenate loci that share the same topology (i.e., left example showing a "true supergene" comprised entirely of red loci). ML-analyses of "true supergenes" (MLE indicated as gray tree) will converge to the true topology as the length of each congruent locus increases, because all sites in the alignment evolved under the same red topology. However, if statistical binning is not accurate, incongruent loci that do not share a common topology may be incorrectly concatenated to form "false supergenes". In the right example, a false supergene has been constructed from three genes with three different topologies (blue, red, and purple). False supergenes represent profound phylogenetic model misspecification, because standard ML-analysis assumes that all sites within an alignment evolved under the same tree, and thus, only one tree will be estimated when there should be three (right example). Regardless of whether this ML topology is the blue, red, purple, or some other topology, the answer is the same: ML-analysis cannot be statistically consistent because it cannot estimate three unique trees. False supergene trees are likely to reflect an amalgamation of conflicting phylogenetic signal (here three distinct trees), such that the gene tree with the most support (i.e., highest number of informative sites) may have disproportionate influence (see Fig. 4). The relevant questions is thus whether statistical binning tends to infer true supergenes (left) or false supergenes (right), and although the method does not directly estimate a species tree, clearly supergene accuracy is likely to influence downstream species tree accuracy.

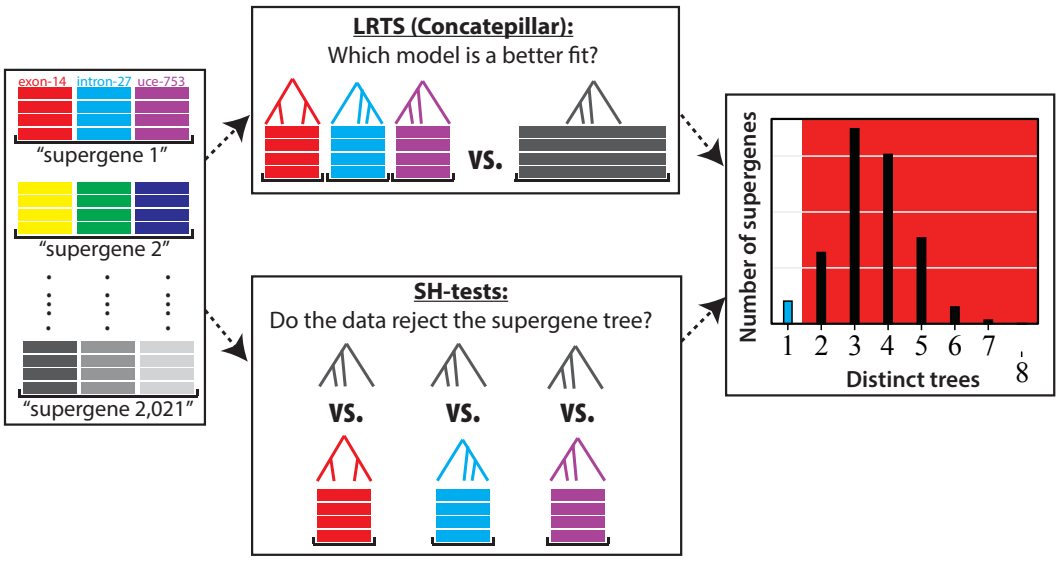


Figure 2. Model-based assessment of statistical binning accuracy. We tested the accuracy of each of the 2,021 supergenes inferred from the Avian phylogenomics project using Likelihood Ratio Tests (LRTs, implemented in Concatepillar, top box) and Shimodaira-Hasegawa (SH-tests, bottom box). The LRTs approach tests how many distinct topologies are present in a supergene inferred via statistical binning. For example, the likelihood of a model consisting of three distinct trees (red, blue and purple in top box) is compared to single-tree model (gray alignment and tree in top box). Similarly, the SH-tests approach evaluates whether individual loci placed within a supergene reject the overall supergene topology in favor of a locus-specific topology (i.e., red vs. gray supergene topology shown in lower box). A "true supergene" and its associated supergene tree are considered accurate if only a single topology is supported by the data (i.e., Fig. 1 left), while a "false supergene" occurs when multiple trees are supported by the data (i.e., Fig. 1 right). For both methods, we quantified the number and fraction of "true supergenes" (blue bar in right histogram) and "false supergenes" that incorrectly concatenate multiple trees (2-8 in this case, black bars and red area).

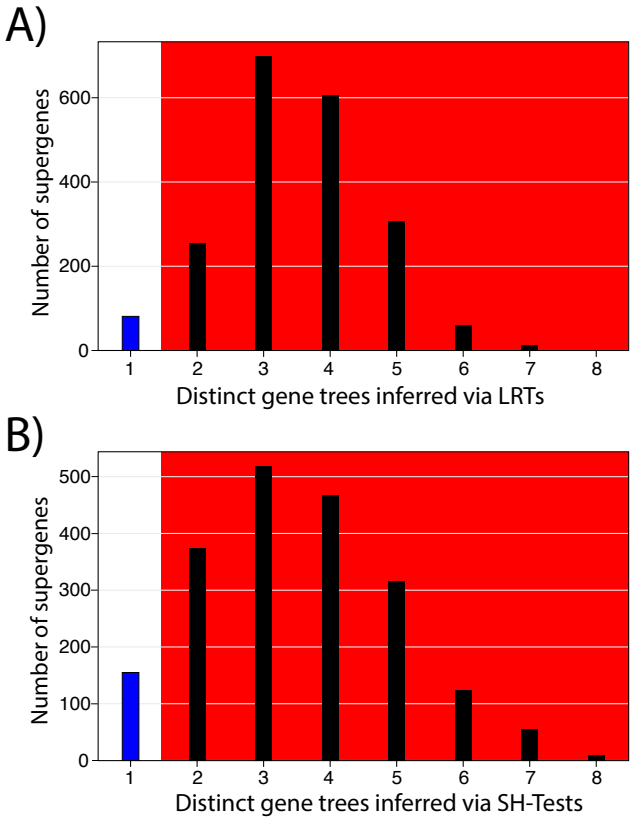


Figure 3. More than 92% of supergenes inferred via statistical binning appear to be false positives. Histograms showing the number of distinct topologies inferred with (a) likelihood ratio tests (LRTs with Concatepillar) and (b) Shimodaira-Hasegawa (SH-tests) across the 2,021 supergenes inferred for the Avian phylogenomic analyses. LRTs (a) and SH-tests (b) indicate that over 96% (1,934/2,021) and 92% (1,866/2,021) of supergenes are false positives, respectively (black bars and red area). In other words, only 4% (81/2,021) of supergenes appear to be "true supergenes" based on LRTs (blue bar), and only 7.7% (155/2, 021) based on SH-tests (blue bar).

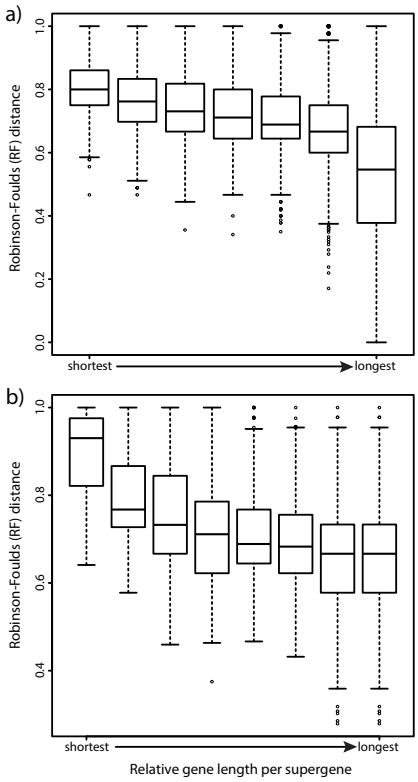


Figure 4. Robinson-Foulds (RF) distances between an individual gene topology and its associated supergene topology decrease with relative gene length, such that supergenes inferred via statistical binning tend to be biased towards the topology of its longest gene. Boxplots indicate the distribution of RF distances between each gene-specific ML topology and its respective supergene ML topology ranked from shortest to longest relative gene length. Results shown for supergenes comprised of 7 genes (a) and 8 genes (b), respectively.

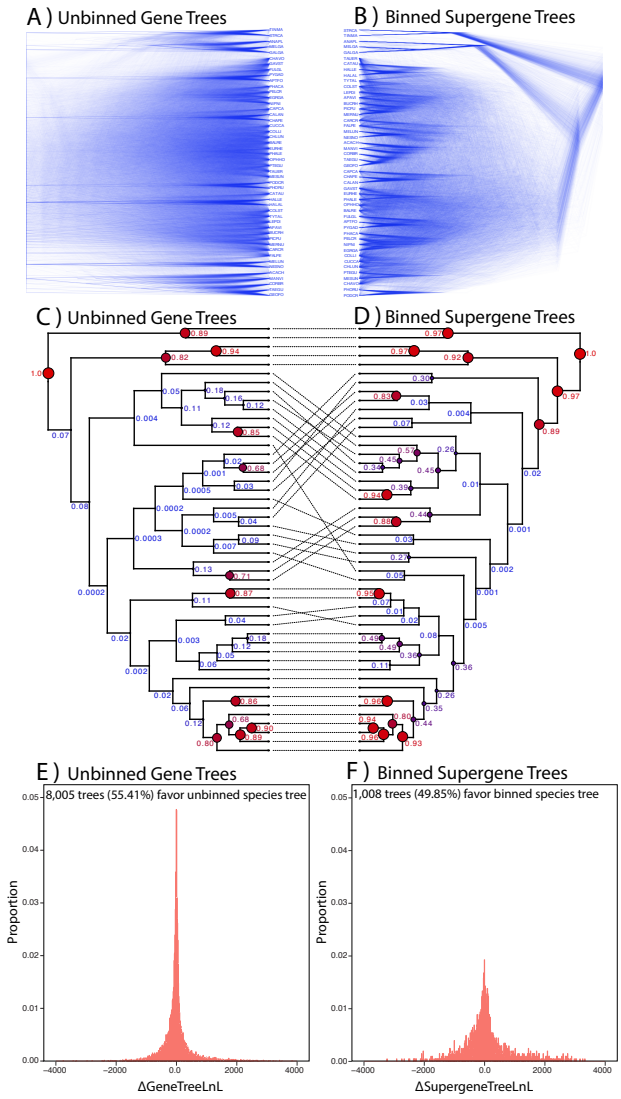


Figure 5. The impacts of statistical binning on gene tree distributions and species tree support. Densitree plots showing the gene tree topology distribution for (a) the individual gene trees ("unbinned") and (b) the supergene trees. Plot of consensus trees with bipartition frequencies estimated using the individual, unbinned gene trees (c) and (d) the supergene trees constructed with statistical binning (d). Node circles are labeled and colored by the bipartition frequencies observed in their respective gene tree distributions. Histograms showing the distributions of multispecies coalescent likelihoods for the unbinned gene trees (ΔGeneTreeLnLs; e) and binned supergene trees (ΔSupergeneTreeLnLs; f).

SUPPLEMENTARY TABLES

ILS-level	Replicates	#genes per replicate	Length	#inferred supergenes	#multilocus supergenes	% false multilocus supergenes	Total % false supergenes	Mean RF-distance
1X	20	1000	1000	10,822	8,948 (82.68%)	100%	82.68%	25.18
1X	20	1000	1500	17,028	2,970 (17.44%)	100%	17.44%	24.16
1X	20	1000	250	1,326	1,325 (99.92%)	100%	99.92%	47.51
1X	20	1000	500	3,451	3,405 (98.67%)	100%	98.67%	40.6
1X	1	14350	mixed	1,645	1,645 (100%)	100%	100%	41.96
1X	20	2000	1000	10,162	9,196 (90.47%)	100%	90.47%	25.55
1X	20	500	1000	5,810	4,103 (70.62%)	100%	70.62%	25.15
0.5X	20	1000	500	7,539	7,149 (94.83%)	100%	94.83%	37.56
2X	20	1000	500	2,203	2,180 (98.96%)	100%	98.96%	35.85

Supplementary Table 1. 100% of supergenes inferred via statistical binning are false positives as indicated by simulation analyses. Results summarized across all replicate datasets for each of the 9 simulation conditions that varied in the degree of incomplete sorting (0.5X, 1X, 2X), number of simulated genes, and sequence lengths. For each simulated dataset, we quantified the number and percentage of false multilocus supergenes (i.e., contain >2 loci) that incorrectly concatenated loci from distinct, conflicting gene topologies. In all cases, no two simulated genes placed within any supergene shared the same topology. In other words, we found that that false positive rate of statistical binning was 100% in all cases for multilocus supergene datasets. Genes incorrectly placed within supergene bins exhibited a high level of conflict as a result of ILS (leftmost column showing average Robinson-Foulds distance across supergenes).

Chapter 5

Probabilistic species tree distances: implementing the multispecies coalescent to compare species trees within the same model-based framework used to estimate them

Richard H. Adams[1] and Todd A. Castoe[1]

[1]Department of Biology, 501 S. Nedderman Dr., University of Texas at Arlington, Arlington, TX 76019 USA

Abstract

Despite the ubiquitous use of statistical models for phylogenomic and population genomic inferences, this model-based rigor is rarely applied to post-hoc comparison of trees. In a recent study, Garba and colleagues derived new methods for measuring the distance between two gene trees computed as the difference in their site pattern probability distributions. Unlike traditional metrics that compare trees solely in terms of geometry, these measures consider gene trees and associated parameters as probabilistic models that can be compared using standard information theoretic approaches. Consequently, probabilistic measures of phylogenetic tree distance can be far more informative than simply comparisons of topology and/or branch lengths alone. However, in their current form, these distance measures are not suitable for the comparison of species tree models in the presence of gene tree heterogeneity. Here we demonstrate an approach for how the theory of Garba *et al*. (2018), which is based on gene tree distances, can be extended naturally to the comparison of species tree models. Multispecies coalescent models (MSC) parameterize the discrete probability distribution of gene trees conditioned upon a species tree with a particular topology and set of divergence times (in coalescent units), and thus provide a framework for measuring distances between species tree models in terms of their corresponding gene tree probabilities. We describe the computation of probabilistic species tree distances in the context of standard MSC models, which assume complete genetic isolation post-speciation, as well as recent theoretical extensions to the MSC in the form of network-based MSC models that relax this assumption and permit hybridization among taxa. We demonstrate these metrics using simulations and empirical species tree estimates and discuss both the benefits and limitations of

these approaches. We make our species-tree distance approach available as an R package called pSTDistanceR, for open use by the community.

Introduction

Quantifying the degree of dissimilarity between phylogenetic tree structures has long been of interest to both mathematicians and evolutionary biologists alike. In particular, considerable attention has been directed towards characterizing the geometry of phylogenetic tree space and developing theoretical and empirical frameworks for measuring the distance between two trees (Estabrook et al. 1985; Kim 2000; Moulton and Steel 2004; Owen 2011; Shi et al. 2013; Kuhner and Yamato 2015). Molecular systematic studies now routinely employ distance measures to quantify variation within sets of trees and assess statistical confidence (or lack of) when summarizing and comparing analyses. For example, phylogeneticists often want to compare trees estimated using different datasets and/or analytical approaches, which can potentially provide insight into underlying sources of phylogenetic conflict (e.g., Castoe et al. 2009; Reddy et al. 2017). This is important because, despite the increase in accuracy predicted to coincide with the ever-increasing size of phylogenomic datasets, phylogenetic estimates often vary greatly from study-to-study, and many species-level relationships remain as contentious as ever (Reddy et al. 2017; Shen et al. 2017b; Walker et al. 2018). Robust methods for measuring phylogenetic distance can be used to dissect the causes and consequences such variation, and thus, their utility is increasingly evident in the face of widespread phylogenetic conflict that has persisted – and sometimes amplified – in the age of genome-scale datasets.

A number of tree distance measures have been proposed, including the Robinson-Foulds metric (Robinson and Foulds 1979, 1981), quartet distance (Estabrook et al. 1985), the geodesic or Billera-Holmes-Vogtmann (BHV) metric (Billera et al. 2001; Owen and Provan 2011), and many

others. Traditionally, these approaches view phylogenetic trees strictly in terms of their geometric properties– that is, only the branching structure (i.e., topology) and/or branch lengths are considered when comparing two trees. Although these measures are usually rapid to compute and benefit from relatively straightforward interpretations (e.g., the Robinson-Foulds metric measures the number of shared splits between a pair of trees), many are also paradoxically restricted by their own dependence on a strictly geometric perspective of trees. Ironically, in contrast to the relative simplicity of tree comparison approaches, tremendous effort has been directed towards understanding phylogenetic trees as probability generating models over the past decades – particularly in the analysis of genetic sequence data. From this model-based viewpoint, we consider the molecular evolutionary processes occurring along branches of a phylogeny that ultimately determine the probability of observing a particular pattern of nucleotides (or amino acids) at a single site. In other words, a phylogenetic tree model parameterizes the probability distribution of site patterns as a function of the topology, branch lengths, and other parameters relevant to the nucleotide substitution process (i.e., relative substitution rates, equilibrium base frequencies). Accordingly, rather than a depiction of tree space solely in terms of topology and/or branch lengths, a probabilistic phylogenetic model is most appropriately identified by a set of points in the space of site patterns, which has been referred to as "phylogenetic oranges" or "hyperdimensional oranges" (Kim 2000; Moulton and Steel 2004).

Viewing phylogenies as probabilistic models instead of solely geometric structures suggests that potentially far greater information can be incorporated for the comparison of trees. For these reasons, Garba *et al*. (2018) proposed the use of probabilistic model-based distances to compare

two trees by measuring the distance between their site pattern probability distributions. Unlike traditional measures based solely on topology and/or branch lengths, these measures effectively incorporate information encoded by parameters of the nucleotide substitution process. As predicted, probabilistic measures can be more informative than traditional topology or branch-length based distances (i.e., Fig. 2 of Garba *et al*. 2018). For example, two trees with exactly the same topology and branch lengths can yield very different site pattern probabilities if the nucleotide substitution parameters differ substantially, and conversely, trees with different topologies can exhibit similar site pattern distributions depending on these parameters. In either case, measuring the distance between two trees in terms of their site pattern probability distributions is likely to illuminate important differences that may be overlooked or obscured when only conducting simple comparisons of topologies. Importantly, this model-based perspective of trees also forms the foundation of likelihood-based methods, such as maximum likelihood estimation (MLE) and Bayesian inference (BI), that have become cornerstones of contemporary molecular phylogenetics. Thus, there is an intuitive link between probabilistic phylogenetic *inference* and the probabilistic phylogenetic *distance* measures of Garba *et al*. (2018), such that trees can be directly compared within the same model-based framework used to estimate them.

Although the distance measures of Garba *et al*. (2018) mark a significant advancement towards more informative distance metrics, they are inherently limited in one fundamental aspect: they only measure distance between *gene* trees, not *species* trees *per se*. Species trees, rather than gene trees, depict the evolutionary relationships among organisms, and thus, reconstructing species-level relationships is the primary goal of most phylogenetic studies (Maddison 1997;

Nichols 2001; Rannala and Yang 2003b). The distinction between gene trees and species trees is critical when computing phylogenetic distances because individual gene trees may bear little resemblance to one another and with the species tree (Nichols 2001; Degnan and Rosenberg 2009b). Incomplete lineage sorting (ILS) is perhaps the most pervasive and well-studied source of gene tree heterogeneity that is notorious for its ability to challenge species tree accuracy (Maddison 1997; Nichols 2001; Degnan and Salter 2005; Edwards 2009c; Edwards et al. 2016). The multispecies coalescent (MSC) model was developed to accommodate ILS by merging phylogenetics and coalescent theory into a unified framework that models the evolution of gene trees imbedded within a species tree (Maddison 1997; Nichols 2001; Rannala and Yang 2003b). A species tree model parameterizes the probability distribution of gene trees conditioned upon the species-level topology and set of divergence times in coalescent units (with one coalescent time unit to be $2N_e$ generations where N_e is the effective population size). Under the MSC, gene trees are therefore permitted to vary from locus-to-locus as a result of the coalescent process occurring within branches of a species tree, and accordingly, site pattern probability distributions may also vary. The probabilistic metrics proposed by Garba et al. (2018) effectively ignore such variation because trees are constrained to a single topology when computing and comparing site pattern probabilities and thus, they cannot be used in their current form to measure the distance between two species tree models. These measures can be used to quantify the distance between any two gene trees, however, this provides only indirect (if inefficient) information about species-level distances. Only when all gene trees share the same topology, branch lengths, and substitution parameters will these measures directly translate to species tree comparisons. Fundamentally, the probabilistic phylogenetic distances proposed by Garba *et al*. (2018)

therefore represent *gene* tree distances that are largely invalid for the comparison of species tree models in the presence of gene tree heterogeneity.

Another unique challenge arises when biological processes yield phylogenetic tree structures that are not strictly bifurcating. In particular, substantial effort has been directed towards developing models that incorporate hybridization events among species in the form of phylogenetic networks (Huson and Bryant 2006; Nakhleh 2010; Degnan and Ane 2017; Zhu and Degnan 2017). To model both ILS and hybridization, theoretical work has extended the MSC to derive network-based species models that depict hybridization events as interconnecting edges in the species tree (Degnan and Ane 2017; Zhu and Degnan 2017). In addition to a species topology and set of divergence times (in coalescent units), the presence of hybridization events in the species tree may also modulate gene tree probabilities. Much remains unknown about the space of phylogenetic networks, and it is not always clear how network distances should be computed because many existing metrics, including the probabilistic gene tree distances of Garba *et al*. (2018), as well as topology-based metrics (i.e., Robinson-Foulds distances), are typically designed to measure strictly bifurcating trees and therefore must be modified to be relevant for reticulating species trees (Cardona et al. 2009; Nakhleh 2010; Degnan and Ane 2017). One particularly relevant concern for network model selection and inference involves the issue of identifiability: two networks can be mathematically or even practically indistinguishable because they induce identical (or nearly so) probability distributions on gene tree topologies (Zhu and Degnan 2017). Although many have been generalized to networks, existing distance metrics often assume a distance of zero when comparing two networks that display the same topology when removing a subset of hybridization edges, even if their gene tree distributions differ

(Cardona et al. 2009; Degnan and Ane 2017). Collectively, these findings suggest that a model-based approach may prove particularly relevant and useful for measuring species network distances because such an approach should, in theory, be able to detect differences (or a lack of differences) in the underlying gene tree probabilities.

In this study, we discuss how the principles and theory of the probabilistic gene tree distance measures proposed by Garba *et al*. (2018) can be generalized for the computation of species tree distances. To derive analogous measures for computing species tree distances, we employ the MSC to parametrize the probability distribution of gene trees conditioned upon a specific topology and set of species divergence times (in coalescent units). Just as Garba *et al*. (2018) viewed gene trees as parametric models that can be compared in terms their site pattern probability distributions, here we measure the distance between two gene tree probability distributions induced by their respective species tree models under the MSC. We first briefly describe the gene tree distances of Garba *et al*. (2018) followed by a modification of these measures to species tree distances. We then demonstrate the utility of this approach using several examples of the MSC. Finally, we apply these measures to more complex network-based species models that present particularly challenging problems for phylogenetic model selection and inference.

Methods

Probabilistic Species Tree Distances

The probabilistic Gene Tree Distance (pGTD) measures proposed by Garba *et al*. (2018) compare two gene trees in terms of the difference in their site pattern probability distributions. Importantly, site patterns are considered independently and identically distributed (i.i.d.) in the computation of pGTD – meaning that gene tree topologies and/or branch lengths do not vary for a given tree. In the presence of gene tree heterogeneity, pGTD measures will not equate to species tree distances because they constrain gene trees to a single topology, branch lengths, and other parameters. We can, however, leverage the same principles of Garba *et al*. (2018) to derive probabilistic species tree distances by substituting species-level parameters into these same equations. See the Supplementary Materials and the original study (Garba *et al*. 2018) for a detailed treatment of probabilistic gene tree distances, which provides a basis for computing species tree distances in a similar manner.

Here we describe how these principles can be used to derive probabilistic Species Tree Distances (pSTD) whereby the goal is to compare species-level relationships, rather than individual gene trees. Just as Garba *et al*. (2018) viewed gene trees as probability generating models, here we leverage the multispecies coalescent (MSC) model to measure the distance between two species trees in terms of their probability distributions on genealogies.

Under the standard MSC (i.e., lineages remain genetically-isolated), a species tree model $\phi = \{T, \lambda\}$ with n extant species defines a discrete probability distribution of all possible gene trees

G_n as a function of the species topology (T) and set of divergence times (λ) in coalescent units. If only a single lineage is sampled per species, the total number of possible rooted gene trees is $|G_n| = \frac{(2n-3)!}{2^{n-2}\,(n-2)!}$, and the probability of a particular gene tree g in G_n is computed as a function of the species tree model: $P(g|\phi = \{T,\ \lambda\})$. To derive species tree distances, we replace terms in the equations of Garba *et al*. (2018) to reflect species models ϕ and their associated gene tree probability distributions (Equations provided in the Supplementary Materials). The distance between two species tree models $\phi_1 = \{T_1,\ \lambda_1\}$ and $\phi_2 = \{T_2,\ \lambda_2\}$ is computed as:

$$d(\phi_1,\ \phi_2) = d(P(G_n|T_1,\ \lambda_1)\,, P(G_n|T_2,\ \lambda_2)) \quad (1)$$

where $P(G_n|T_1,\ \lambda_1)$ is the probability distribution of gene trees given the model parameters ϕ_1(likewise for ϕ_2) and $d(\phi_1,\ \phi_2)$ can represent the Hellinger distance (d_H), the Kullback-Leibler distance (d_{KL}), or the Jensen-Shannon distance (d_{JS}^2), shown below in Equations 2-4:

$$d_H(\phi_1,\ \phi_2)^2 = \frac{1}{2}\sum_{g\epsilon G_n}\left(\sqrt{P(g|\phi_1)} - \sqrt{P(g|\phi_2)}\right)^2 \quad (2)$$

$$d_{KL}(\phi_1,\ \phi_2) = \sum_{g\epsilon G_n} P(g|\phi_1) \times \log\left(\frac{P(g|\phi_1)}{P(g|\phi_2)}\right) \quad (3)$$

$$d_{JS}^2(\phi_1,\ \phi_2) = \frac{1}{2} d_{KL}\left(P(g|\phi_1); \frac{P(g|\phi_1)+P(g|\phi_2)}{2}\right) +$$

$$\frac{1}{2} d_{KL}\left(P(g|\phi_2); \frac{P(g|\phi_1)+P(g|\phi_2)}{2}\right) \quad (4)$$

We have implemented these equations in an R software package (pSTDistancesR) that uses HYBRID-COAL (Zhu and Degnan 2017) to generate gene tree probability distributions (see Software Availability section below). These equations are effectively the same equations

proposed by Garba *et al*. (2018) except that gene-tree and substitution parameters have been replaced by species-level parameters. For each possible genealogy in G_n, we record the difference in the probability of that genealogy between two species tree models and sum these differences across gene tree space using Equations 1-4. For example, consider two 4-species tree models $\phi_1$and ϕ_2. In this case, there is a total of $|G_4| = \frac{(8-3)!}{2^{4-2}\,(4-2)!} = 15$ possible genealogies, and for each genealogy in this set G_4, we measure the difference between its probabilities under ϕ_1 and ϕ_2 using Equations 2-4. By implementing the MSC in such a manner, we are effectively incorporating information about the coalescent process running along branches of the species tree model when computing distances. For example, two species trees can have the exact same topology (i.e., $T_1 = T_2$= Robinson-Foulds distance of zero) but very different gene tree distributions depending on the branch lengths, which determine the probability that a pair of lineages coalesce within a particular species branch. We also note that the Kullback-Leibler distance is not a true metric because it is not symmetric (i.e., $d_{KL}(\phi_1,\ \phi_2) \neq d_{KL}(\phi_2,\ \phi_1)$) and does not satisfy the triangle equality (see Supplementary Materials for more information) – this is a fundamental property of the Kullback-Leibler distance that is relevant to any of its applications, including the original gene tree distances of Garba *et al*. (2018). Despite this limitation, we include the Kullback-Leibler distance here because of its wide use for model comparison, particularly in the field of systematics.

For the purposes of this study, we primarily discuss the computation of pSTD on species trees with relatively fewer tips (<10), for which probabilistic distances can be computed analytically using Equations 2-4. However, the total number of possible gene trees $|G_n|$ can be tremendous for larger species trees, and these distances can be estimated using simulations in a manner

similar to Garba *et al*. (2018). For example, we can obtain a sample of m gene tree topologies from each species tree and approximate the Hellinger and Kullback-Leibler distance between ϕ_1 and ϕ_2 as:

$$d_H^*(\phi_1,\ \phi_2)^2 \simeq 1 - \left(\frac{1}{2m}\right)\sum_{i=1}^{m}\left(\sqrt{\frac{P(\boldsymbol{g}_{i,\phi_1}|\phi_2)}{P(\boldsymbol{g}_{i,\phi_1}|\phi_1)}} + \sqrt{\frac{P(\boldsymbol{g}_{i,\phi_2}|\phi_1)}{P(\boldsymbol{g}_{\mathrm{i},\phi_2}|\phi_2)}}\right) \quad (5)$$

$$d_{KL}^*(\phi_1,\ \phi_2) \simeq \frac{1}{m}\sum_{i=1}^{m}\log\left(\frac{P(\boldsymbol{g}_{i,\phi_1}|\phi_1)}{P(\boldsymbol{g}_{i,\phi_1}|\phi_2)}\right) \quad (6)$$

To explore potential advantages and disadvantages of pSTD in relation to other metrics, we computed pSTD in two scenarios: a pair of bifurcating species trees with the same topology and branch lengths that only differ by a scaling factor γ (Fig. 1a vs. Fig. 1b), and a pair of species trees with the same topology and branch lengths that are identical except for one internal branch that is scaled by γ (Fig. 1a vs. Fig. 1c). These scenarios represent similar examples to those shown in Figure 2 and Figure 3a of Garba *et al*. (2018) in which either a single branch or all branches of gene trees were scaled by a factor when comparing pGTD and BHV metrics. For the first scenario, we consider two bifurcating species tree models $\phi_1 = \{T_1,\ \lambda_1\}$ and $\phi_2 = \{T_2,\ \lambda_2\}$ that share the same topology (i.e., $T_1 = T_2$), but the branch lengths of the second model ϕ_2 are obtained by scaling the branch lengths of ϕ_1 by a factor γ, such that $\lambda_2 = \gamma\lambda_1$ (Figure 1a vs. Figure 1b). Similarly, in the second scenario, only the length of the internal branch for the second species tree is scaled by γ (Fig. 1c). To explore the properties of pSTD under varying degrees of ILS, we specify ϕ_1 to the following (in newick format): "(((A:1,B:1):1,C:2):1,D:3)" and we allow γ to vary from $0 - 10$.

Probabilistic Distances as a Framework for Comparing Increasingly Complex Species Tree Models

While comparing gene tree distributions under multispecies coalescent models is the primary focus of this study, we argue that this approach could be extended to incorporate and compare species tree models that include other evolutionary processes, such as migration, hybridization, recombination, and selection, among others. Here we demonstrate two of potential extensions of our pSTD approach: (1) reticulating species tree models and (2) nucleotide site pattern probabilities. In the previous section, we have applied a simplistic and commonly used interpretation of the MSC whereby species are assumed to diverge in genetic isolation from one another in the absence of gene flow, natural selection, migration, hybridization, or any other evolutionary process. That is, the probability of a gene tree (used to compute the distances of Equations 2-6) is only a function of the species tree topology and branch lengths in coalescent units, such that all gene tree heterogeneity is assumed to arise from ILS. Recent work has expanded the MSC to accommodate hybridization with the development of Network Multispecies Coalescent (NMSC) models (Degnan and Ane 2017; Zhu and Degnan 2017). The NMSC can be incorporated into our pSTD equations to compute the distances between network species models that include hybridization edges. For example, the species model ϕ can include a network topology (instead of a strict bifurcating tree) and other parameters associated with the timing and duration of hybridization. Species models with different network topologies can therefore be compared with one another, and with models that do not include hybridization. To explore the utility of pSTD for comparing complex phylogenetic structures, we computed probabilistic distances between two species tree networks (Fig. 2a vs. 2b), and separately

between a network and a bifurcating tree (Fig. 2a vs. 2c). These networks (Fig. 2a-b) were chosen because they present particularly challenging problems for network-inference and distance computation, and were used in recent studies of network models (Degnan and Ane 2017; Zhu and Degnan 2017). In the first scenario, the two different species networks display the same tree after removal of hybridization edges that differentiate the two networks. As before, we let the edge lengths of ϕ_2 scale by a factor γ that ranges from $0 - 10$. In a second example, we use pSTD to compute the distance between a network (Fig. 2a) and a bifurcating species tree model (Fig. 2c).

Another example extension of these distances is the incorporation of mutational processes that give rise to molecular sequence data. For example, probabilistic distances may also incorporate site pattern probabilities that are contingent upon the gene tree distributions, thereby providing a natural comparison to the gene tree distances of Garba *et al.* (2018). We demonstrate the utility of incorporating mutation into the probabilistic species tree distances by computing pSTD between two species tree models (Fig. 3a vs. 3b) across a range of branch scaling values to obtain the second tree (Fig. 3b). For these examples we use a mutation rate of $\mu = 10^{-5}$ under the 4-state JC69 model (Jukes and Cantor 1969) using the site pattern probability equations and example species trees provided in Chifman and Kubatko (2015), and a scaled population size parameter $\theta = 2N_e\mu = 0.10$ for all branches in the model (Fig. 3).

Four Empirical Demonstrations of Probabilistic Species Tree Distances

We applied our probabilistic species tree distance measures to four different empirical examples that included: (1) quantifying variation within a set of species tree estimates obtained using

resampling procedures (i.e., bootstrapping) across different genomic regions, (2) comparing species trees estimated using different methods and/or datasets, (3) dissecting contentious estimates of phylogenetic relationships, and (4) characterizing a Bayesian posterior probability distribution of species tree model estimates obtain via Markov Chain Monte Carlo (MCMC) sampling. For the first and second demonstrations, we used the avian phylogenomic analyses (Jarvis et al. 2014) as an example dataset because this dataset has been used as a case-study for understanding the performance of species tree estimation methods on genome-scale datasets (Mirarab et al. 2014; Liu and Edwards 2015) and for dissecting causes of phylogenetic conflict (Reddy et al., 2017). We downloaded a set of 14,446 estimated gene trees and a set of 32 species tree topologies that were estimated in the original study (i.e., Jarvis et al., 2014) or estimated in previous studies (i.e., Prum et al. 2015), which allowed us to compare species tree estimates across different datasets and approaches. We pruned these trees down to 8 focal taxa that represent challenging and contentious problems for resolution of the avian phylogeny: bald eagle (*Haliaeetus_leucocephalus*), barn owl (*Tyto alba*), speckled mousebird (*Colius striatus*), cuckoo roller (*Leptosomus discolor*), downy woodpecker (*Picoides pubescens*), carmine bee-eater (*Merops nubicus*), rhinoceros hornbill (*Buceros rhinoceros*), and bar-tailed trogon (*Apaloderma vittatum*). For all analyses, probabilistic distances between species trees were computed analytically using Equations 1-4.

For the first demonstration, we quantified variation among sets of bootstrapped species trees that were estimated from different chromosomes. For each of the five first chromosomes of the chicken genome (Gallus_gallus-5.0; GCA_000002315.3; Warren et al., 2017), we obtained a set containing all available gene trees that were estimated in Jarvis *et al.* (2014) for that

chromosome, and we used these gene tree sets to conduct nonparametric bootstrap resampling (with 10 replicates) independently for each chromosome using MP-EST (Liu et al. 2010a). In other words, we obtained 10 bootstrapped species tree estimates for chromosome one, and so on, for each of the five largest autosomes using their respective gene tree sets. We used multidimensional scaling of the Hellinger distance (computed analytically using Eq. 4), and the R package TREESPACE (Jombart et al. 2017) to characterize variability among chromosome-scale species tree estimates in the phylogenetic placement of avian lineages. In the second demonstration, we computed pairwise species tree distances between 32 different estimates of the avian phylogeny. These 32 different estimates were obtained using different datasets, models, methods, and studies, and were analyzed in the context of the original genome-scale inferences of Jarvis *et al.,* (2014) or subsequent critical reanalysis of these data (Prum et al. 2015b; Reddy et al. 2017). We used the program MP-EST (Liu et al. 2010a) to estimate the branch lengths of these species trees in coalescent units following the general protocol of Jarvis *et al.,* (2014). We computed pairwise distances between all 32 species trees, and used these to construct a cluster-based NJ tree using the R package PHANGORN (Schliep 2011) to quantify similarities among estimates.

For the third demonstration, we used three case-studies of contentious relationships (Amphibians, Neoaves, and Reptiles) that were highlighted in a recent study focused on the causes and consequences of phylogenetic conflict (Table 1 in Shen et al. 2017). We downloaded six species trees (shown in Fig. 6) and the set of 9,363 gene trees from the original study (Shen et al. 2017b), which we used to estimate the branch lengths of species trees in coalescent units using MPEST. We computed probabilistic distances between each of the three species tree pairs,

as well as both the rooted and unrooted Robinson-Fould distances, and the BHV metric. For the fourth application, we used an example dataset for estimating species-level relationships of Canids using Bayesian species tree estimation with the program StarBEAST2 (Ogilvie et al. 2017). We downloaded the CanisPhylogeny-example.xml file from the 'example files' that are provided with StarBEAST2, and ran the MCMC chain for a total of 6,000 generations using this example file. We sequentially sampled 10 species tree estimates every 1,000 generations (total of 60), and computed the pairwise Hellinger distances between all 60 species tree estimates using Equation 2.

Results

Scaling Species Divergence Times

Comparing species tree distances across an array of branch scaling factors highlights the benefits of incorporating gene tree probability distributions for comparing and contrasting species tree distance measures (Fig. 1). In the comparison of two bifurcating species trees with the same topology and branch lengths that only differ by a scaling factor γ (Fig. 1a vs. Fig. 1b), probabilistic distance measures show little resemblance to the BHV metric across an array of values for γ (Fig. 1d). Scaling branch lengths by γ results in complex differences in the underlying gene tree probability distributions that are reflected by differences in the probabilistic measures shown in Figure 1, while the Robinson-Foulds distance is zero in all cases for trees shown in Figures 1 and 2. In contrast, the BHV metric simply scales linearly with γ, while the Hellinger, Kullback-Leibler, and Jensen-Shannon distances exhibit more complex relationships. In the second scenario for which only a single branch of ϕ_2 is scaled by γ (i.e., all other branches

remain unchanged; Fig. 1a vs. Fig. 1c), we observe similar trends with pSTD that provide more informative comparisons between two trees (Fig. 1e). The Hellinger and Kullback-Leibler distance metrics exhibit asymptotic trends toward their respective limits (Fig. 1d-e), suggesting diminishing impacts of branch length scaling on gene tree probability distributions with larger values of γ.

Comparing More Complex Species Tree Models Using pSTD

Probabilistic network distances are able to compare complex species tree structures, and we demonstrate that here across two examples: between two species tree networks that display the same tree (after removal of hybridization edges that differentiate the two networks; Fig. 2a vs. 2b), and between a network and a bifurcating tree (Fig. 2a vs. 2c). pSTD computed in both scenarios reveal the effects of branch scaling on network distances (Fig. 2d), and the potential utility of pSTD for comparing a network with a bifurcating tree. (Fig. 2e). As with the examples shown in Figure 1, we see that the Hellinger and Jensen-Shannon distances appear to exhibit asymptotic behavior as the edge length differences increase between species models. However, the Kullback-Leibler distance, which is not a metric (i.e., it is asymmetric and does not satisfy the triangle inequality), increases far more rapidly, particularly when comparing a network and a bifurcating topology (Fig. 2e).

Although we have primarily focused on comparing gene tree distributions, we also show how nucleotide site pattern probabilities can be incorporated into the distance computations to demonstrate an additional extension of the species tree distance approach. Comparing two species tree models (Fig. 3a vs. 3b) in terms of their site pattern probability distributions under

the multispecies coalescent model + 4-state JC69 model highlight the ability for pSTD approaches to effectively incorporate mutational processes when comparing phylogenetic models (Fig. 3c). As before, we see that the BHV metric simply scales linearly as species trees differentiate. For example, the probabilistic distances shown in Figure 3 exhibit complex shifts in slope as the internal branch lengths of the species tree become more distant. As before, the Robinson-Foulds distance is zero in all cases.

Four Empirical Applications Of pSTD

In our first example application of pSTD, variation in key nodes of the avian phylogeny was quantified by comparing distances between bootstrap replicates estimated from different chromosomes (Fig. 4). This was visualized using multidimensional scaling (MDS; Hillis et al. 2005) of the Hellinger distance (Eq. 2), providing a detailed depiction of the bootstrap sampling space of species trees across chromosomes, highlighting both differences and similarities among chromosomes in species tree estimates (Fig. 4). For example, species tree estimates derived from chicken chromosome 3 show greater variation that those derived from chromosome 2, while estimates from chromosome 4 and 5 show substantial overlap with one another.

Our second empirical application demonstrated pSTD by applying these distances to quantify variation in avian species tree estimates inferred from different data subsets, models, and inferential approaches (Jarvis et al. 2014; Reddy et al., 2017; Prum et al. 2015). Clustering of species tree estimates based on pSTD (i.e., Hellinger distance, Eq. 2) are markedly different than those based on Robinson-Foulds distances alone (Fig. 5a vs. Fig. 5b), and more informative (i.e., the collapsed nodes in Fig. 5b provide no additional information). Our clustering of species trees

based on pSTD differs notably from the results shown in Reddy *et al*. (2017) previously used to characterize and understand conflict among species trees estimated using different datasets (i.e., Fig. 6 of Reddy et al. 2017). Perhaps the most apparent contradiction between our clustering results based on pSTD and other metrics is the disparate clustering of species trees obtained using the so-called heuristic "statistical binning" approaches, which attempt to build longer supergenes prior to gene tree estimation (Mirarab et al. 2014), and all other metrics (Fig. 5a). For example, the "unbinned" intron and total evidence ("TENT") species trees formed a cluster distant from "binned" analyses of these same datasets based on pSTD (Fig. 5a), and conversely, the "binned" and "unbinned" analyses of these two datasets cluster together when compared using the Robinson-Foulds metric (Fig. 5b). pSTD-based clustering also highlights major discrepancies in the placement of the "PRUM 2015" tree, suggesting very different gene tree probability distributions between this tree and the "binning" trees estimated in Jarvis *et al.* (2014). For example, the Hellinger distance (Eq. 2) suggests that the "PRUM 2015" tree and the unbinned analyses are more similar to one another (Fig. 5a), yet the Robinson-Fould metric indicates that the topology of this tree is identical to the tree obtained in Jarvis *et al.* (2014) using the "binned" analysis of introns (Fig. 5b).

We used pSTD to explore species tree distances for several vertebrate clades that included contentious relationships based on previous studies as a third empirical application of pSTD. These analyses demonstrate that probabilistic measures of species tree distance can be particularly useful for enabling more complete dissection of differences in topology and branch lengths that differentiate contentious species tree inferences (Fig. 6). In all three test-case examples taken from Shen *et al.* (2017), the unrooted Robinson-Foulds distance is zero, while

the various probabilistic measures effectively compare these contentious estimates in terms of their gene tree probability distributions. Finally, in our four demonstration, we used pSTD to characterize a posterior distribution of species tree estimates sampled at different times along a single MCMC chain from a StarBEAST2 run. This example demonstrates well that pSTD can be particularly useful for dissecting variation among estimates, and even for testing for convergence of MCMC chains (Fig. 7). MDS of the pairwise Hellinger distance indicates that samples taken earlier in MCMC show greater variation (e.g., MCMC Set 1, Fig. 7) compared to samples taken later in the MCMC consistent with convergence of the MCMC towards the posterior.

Discussion

Over the past few decades, tremendous effort has been directed towards understanding phylogenetic trees as probability generating models on character data. Indeed, phylogenetic inference is now predominantly a model-based endeavor whereby evidence in support of alternative hypotheses can be assessed and quantitatively leveraged to estimate parameters and significance. While the application of model-based frameworks to statistical inference has become a cornerstone of contemporary molecular phylogenetics, model-based approaches for comparing phylogenetic trees are still in their relative infancy. Given the ubiquitous use of statistical models for the purpose of evolutionary inference, it seems ironic that studies rarely (if ever) conduct a model-based comparison of trees that were estimated within a model-based framework. The probabilistic measures proposed by Garba *et al*. (2018) improve substantially upon the shortcomings of previous approaches, but their application is largely restricted to gene tree comparisons and are not directly applicable to models of species trees and networks. Here we have generalized these approaches to derive probabilistic *species* tree distance measures.

Understanding the species-level relationships among organisms is the primary focus of the majority of phylogenetic studies, such that gene trees are typically viewed as "nuisance parameters" because they often conflict strongly with one another and may individually provide little insight into the true, species-level relationships. Gene tree heterogeneity is widespread in nature and often poses significant challenges for phylogenetic inference as a result of different evolutionary processes, including incomplete lineage sorting (Heled and Drummond 2010; Camargo et al. 2012), migration (Zhang et al. 2011b; Qu et al. 2012; Leaché et al. 2014), hybridization (Meng and Kubatko 2009; Zhu and Degnan 2017), recombination (Lanier and Knowles 2012), and selection (Castoe et al. 2009c, 2010; Adams et al. 2018). The impacts of gene tree variation on species tree estimation have been a central topic of interest for the past few decades, resulting in the development of multispecies coalescent models for accommodating ILS and its associated gene tree conflicts (Nichols 2001; Rannala and Yang 2003b; Heled and Drummond 2010; Edwards et al. 2016). By implementing the multispecies coalescent model, pSTD provide a means for comparing species trees in terms of their induced gene tree probabilities, which can provide more information than simple measures of topology and/or branch lengths of species trees. Species trees are now commonly estimated within the MSC framework, and thus, pSTD measures allow species trees to be compared within the same framework used to estimate them. Furthermore, we have shown that these probabilistic measures represent a general framework that is easily extended for comparing increasingly complex species tree models that consider other evolutionary processes in addition to ILS (i.e., Fig. 2-3).

Here we have demonstrated several applications for pSTD, although many more diverse applications likely exist, particularly considering that the method itself can be readily modified to

incorporate more complex versions of the standard MSC model. Importantly, we demonstrate the utility of pSTD for illuminating differences in species tree estimates likely driven by biological, methodological and statistical factors. For example, in the limited number of applications included in this study we were able to demonstrate how using pSTD can illuminate distinct biologically-relevant phylogenetic signal from different chromosomes (Fig. 3), and also be used to diagnose statistical properties and variation among species tree estimates sampled by bootstrapping or from Bayesian MCMC chains (Figs. 4 and 7). We also demonstrated how pSTD may be extended to incorporate additional processes, such as hybridization and mutation which further increase the flexibility and thus the utility of pSTD. In one of these demonstrations we use an extended form of pSTD to test among speciation network hypotheses, and between network-based and bifurcating species trees (Fig. 2) – both of which represent key challenges to other methods and priorities for modern speciation research (Degnan and Ane 2017; Zhu and Degnan 2017).

Our example applications of pSTD also highlight the utility of these distances for dissecting the basis of variation in species tree inferences derived from different analytical approaches, datasets, or phylogenetic models (Fig. 5). In these comparisons that utilize species tree inferences based on avian phylogenomic data (Jarvis et al. 2014; Reddy et al., 2017; Prum et al. 2015), pSTD measures suggest that a model-based comparison of species trees can be far more informative than simple topology and/or branch length comparisons. Intriguingly, pSTD-based clustering indicated that avian phylogenomic species tree estimates tend to cluster together based on the specific method used (i.e., the “unbinned” MP-EST analyses clustered separately from the “binned” analyses in Fig. 5a), rather than the particular dataset used. This result contradicts

clustering based simply on topology alone, which indicates the species tree estimates obtained using the same data-type are more similar (Fig. 5b). For example, the TENT (total nucleotide evidence trees) inferred in Jarvis *et al.* (2014) exhibited the same topology regardless of whether the "binned" or "unbinned" approach was used (Fig. 5b), and yet, these two species trees induce very different gene tree probability distributions, which is reflected when computing pSTD (Fig. 5a). These findings also agree with recent studies that suggest heuristic species tree approaches may have particularly strong and misleading influence on species tree estimation (Liu and Edwards 2015; Roch et al. 2018). Therefore, pSTD comparisons of species tree distributions may provide insight into the potential effects that species tree methods may impose on species tree inference that is not otherwise identified by other measures.

Our example applications of pSTD also highlight the broad utility of the approach for investigating model identifiability (or lack of) in several contexts – a topic that represents a major concern for species tree estimation (Chifman and Kubatko 2015; Degnan and Ane 2017; Zhu and Degnan 2017). In the context of the MSC, this means that the number of gene trees required to distinguish between competing species tree models may exceed the limits of reasonably-sized empirical datasets for two models that are practically indistinguishable. The practical ramifications of model identifiability are critical considerations for empirical studies because gene trees themselves are always estimated (rather than known), which introduces another source of potential error into the problem. The problem of identifiability has been particularly relevant in the context of reticulating phylogenetic networks (Degnan and Ane 2017; Zhu and Degnan 2017), and our analyses highlight the utility of pSTD as a tool for understanding model identifiability of complex species tree models. Indeed, modeling species

hybridization entails numerous challenges for phylogenetic model selection and inference. If the number of hybridization events is unbounded, for example, the space of phylogenetic networks is infinitely large, suggesting that the size of network space can be much larger than that of bifurcating trees (Degnan and Ane 2017; Zhu and Degnan 2017). The inherent difficulties of computing network distances has been noted by previous authors (Degnan and Ane 2017), and several traditional geometric-based measures, such as the Robinson-Foulds distance, have been augmented for the comparison of network topologies (Cardona et al. 2009; Nakhleh 2010), but make several limiting assumptions. Here we have shown that pSTD can be readily extended for comparing reticulating species trees because it can determine whether networks are distinguishable (i.e., pSTD = 0) or indistinguishable (i.e., pSTD > 0) in terms of their gene tree probabilities. For example, our distance metrics are able to quantify and confirm previous studies demonstrating the indistinguishability of networks that display the same topologies when only a single allele is sampled per species because their probabilistic distance is zero (Fig. 2d). Additionally, we have shown that pSTD can be used to measure the distance between a species network and a strictly bifurcating model (Fig. 2e). Collectively, these results suggest that pSTD may provide a particularly valuable framework for enabling meaningful comparisons of complex phylogenetic tree structures and a means for understanding the identifiability of these complex models – areas of great importance for the continued development and implementation of more realistic phylogenetic models.

Although the species tree distance measures discussed in this study entail several advantages and useful applications, they also are limited in several key ways. One key limitation is the higher computational cost of measuring model-based distances for species trees, compared to simple

topology or related measures, which would scale with the number of taxa in the tree. For this study, we have demonstrated these measures using trees with fewer taxa (i.e., <10) to improve computational tractability, and for the purpose of understanding the relationships of specific contentious subclades (i.e., Fig. 6). The time taken to compute the 6000 pSTD shown in Figure 1 was ~1.5 minutes, while the 6000 computations shown in Figure 2 were completed in ~4 minutes, both using an Intel(R) Core i5 3.8GHz processor. To measure the distance between different estimates of the avian phylogeny (Fig. 4-5) and for the examples of contentious phylogenetic estimates (Fig. 6), we increased computational feasibility by subsampling the phylogeny and computing distances between subtrees extracted from a larger tree. This approach is similar to the pruning strategy employed by Reddey *et al.* (2017) that compared the phylogenetic placement of specific "indicator clades". Another limitation is the number of lineages sampled per species. Currently, the software we used to compute gene tree probabilities under the MSC and NMSC (i.e., HYBRID-COAL; Zhu and Degnan 2017) provides gene tree probability distributions conditioned upon a single individual (i.e., single haploid sequence) sampled per species, although more complex sampling schemes should be relatively straightforward to incorporate. One popular application of the MSC is for conducting species delimitation to evaluate alternative models of speciation (i.e., different schemes for lumping or splitting of individuals into species; Fujita et al., 2012; Yang and Rannala, 2010), and pSTD permit the comparison of species delimitation models in precise terms of their gene tree probabilities. Theoretically, internal branch lengths in the species tree could be set to zero to compare models that split or lump individuals into a single species or population. Currently, the pSTD measures discussed in this study only consider ILS and hybridization, yet many other evolutionary processes may generate gene tree heterogeneity. Despite its limitations, the broad

applicability and extendibility of the pSTD approach argues for its broad value and utility for addressing biological, methodological, and statistical questions in the context of the MSC – many of which were not readily addressed with previous measures.

Conclusions

Phylogenetic distance measures have become an integral part of phylogenetic analyses with broad applications across the field of evolutionary biology. Probabilistic measures of tree distances provide an intuitive framework for comparing model-based estimates of phylogeny and incorporate inherent advantages over traditional measures that compare only topology and branch lengths. Here we have generalized the same theory and statistical framework used for computing gene tree distances to the context of probabilistic species tree model comparison. This logical extension of gene tree distances to species tree models enables a broad spectrum of enhanced model comparisons that fill an important gap for comparing species tree models, including non-bifurcating network models. Indeed, computing network distances has historically proved difficult, and our demonstrations here show how probabilistic-based distances can be leveraged to compare species networks in the precise terms of their gene tree probabilities. As further extensions and advancements improve the complexity of species tree models, we envision that these distance measures can provide an increasingly valuable foundation for comparing models that incorporate a wide-range of evolutionary processes, such as migration, recombination, and natural selection.

Software Availability

We developed an open source software package pSTDistanceR written in R 3.4.1 (R Core Team 2017) and C++ that computes the Hellinger, Kullback-Leibler, and Jensen-Shannon pSTD using Equations 1-6 and the program HYBRID-COAL (Zhu and Degnan 2017), which is used to extract gene tree probabilities under both the standard MSC (without hybridization) and the NMSC. pSTDistanceR is freely available on github: https://github.com/radamsRHA/pSTDistanceR/. All scripts used to generate the figures in the study are provided in the Supplementary Materials.

Funding

Support was provided from startup funds from the University of Texas at Arlington to TAC, NSF grant to TAC (DEB-1655571), and Phi Sigma Support to RHA. Additionally, both the Lonestar and Stampede compute systems of the Texas Advanced Computing Center (TACC) were utilized for these analyses.

Figures

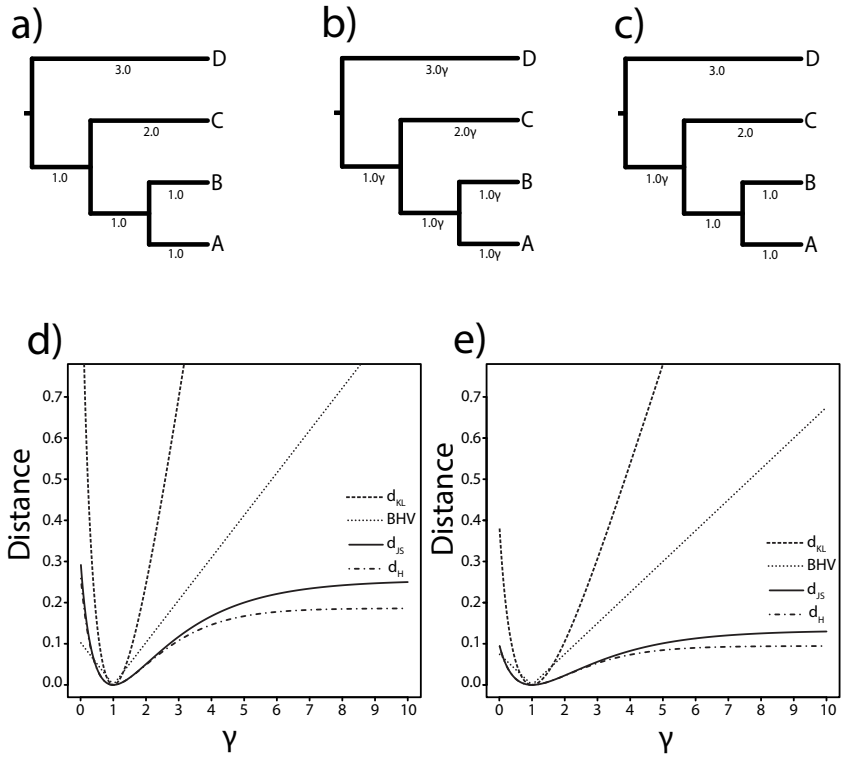


Figure 1. Species tree models and phylogenetic distances for two scenarios of branch scaling. The first species tree model is shown in (a), which was used to obtain the second species model (b) by scaling all branch lengths by a factor γ. The Hellinger (d_H), BHV (d_{BHV}), Jensen-Shannon (d_{JS}), and Kullback-Leibler (d_{KL}) distances between (a) and (b) are shown in plot (d). Similarity, the length of a single internal branch in species tree (a) was scaled by γ to obtain the species tree shown in (c). Plot (d) shows the distances across a range of γ when comparing (a) and (c). Note that in all cases, the Robinson-Foulds distance is zero (i.e., topologies are identical).

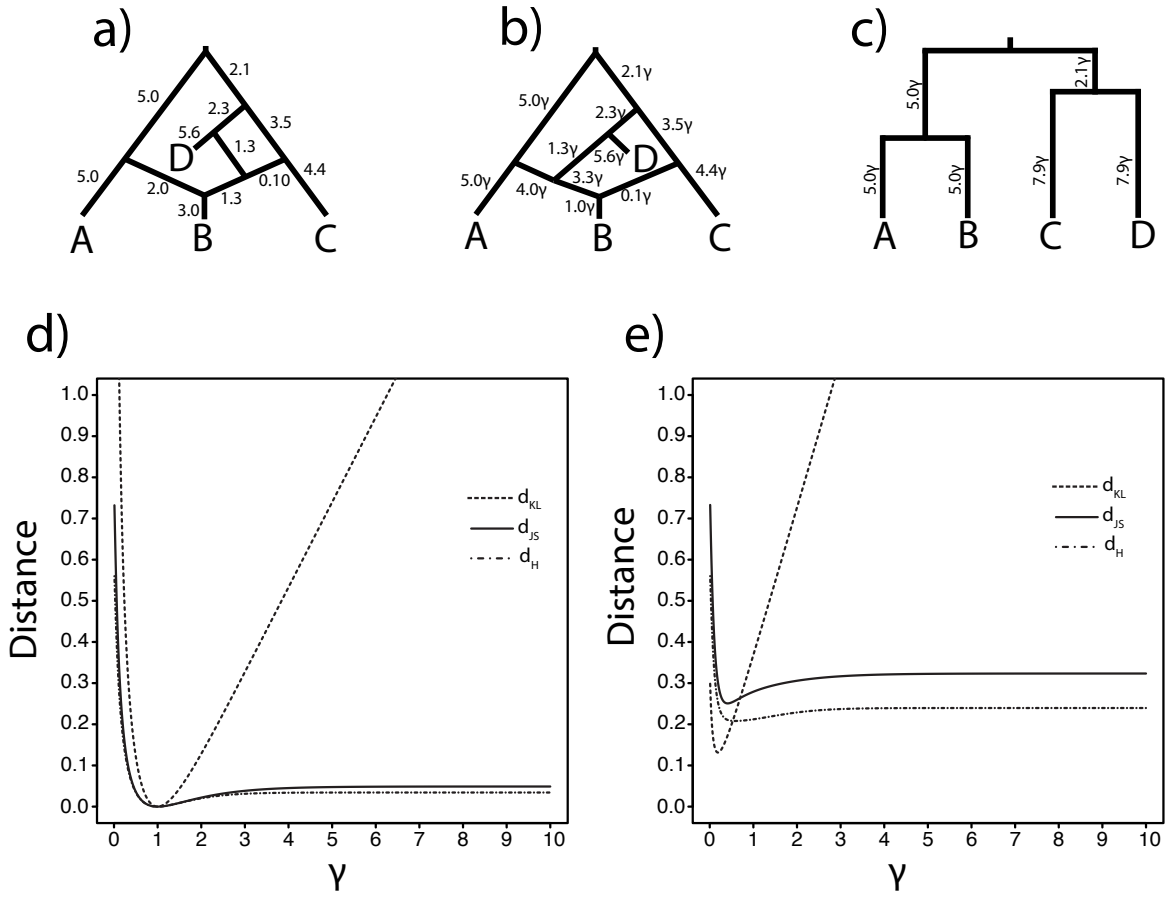


Figure 2. Species models and probabilistic distances for two scenarios of branch scaling. The first network model is shown in (a), which was used to obtain the second species model (b) by scaling all branch lengths by a factor γ. Probabilistic species tree distances computed between (a) and (b) are shown in plot (d). Plot (e) shows the same probabilistic distances computed across a range of γ when comparing (a) and the bifurcating species tree model shown in (c).

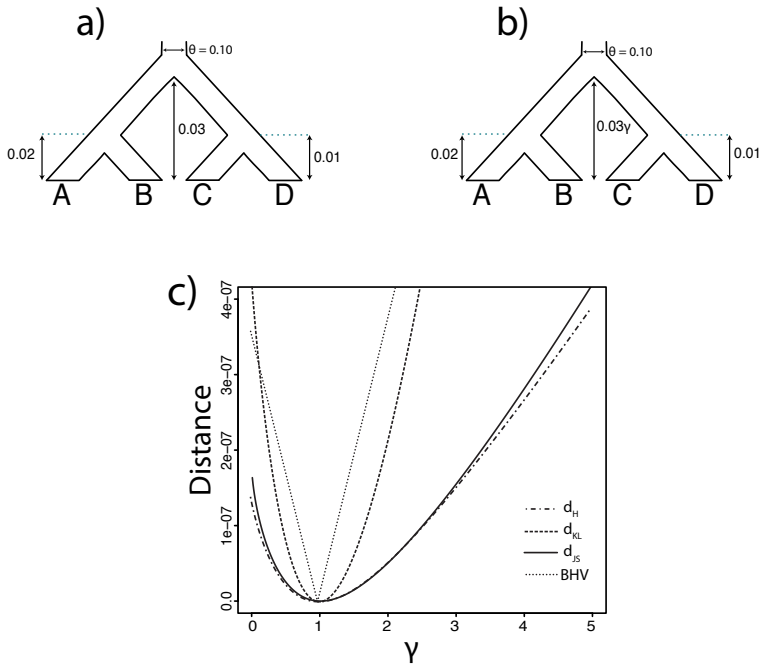


Figure 3. Probabilistic distances that incorporate site pattern probabilities using the 4-state JC69 model under the multispecies coalescent. Species tree distances measured between (a) and (b) are shown in plot (c) across a range of branch length scaling for species tree (b).

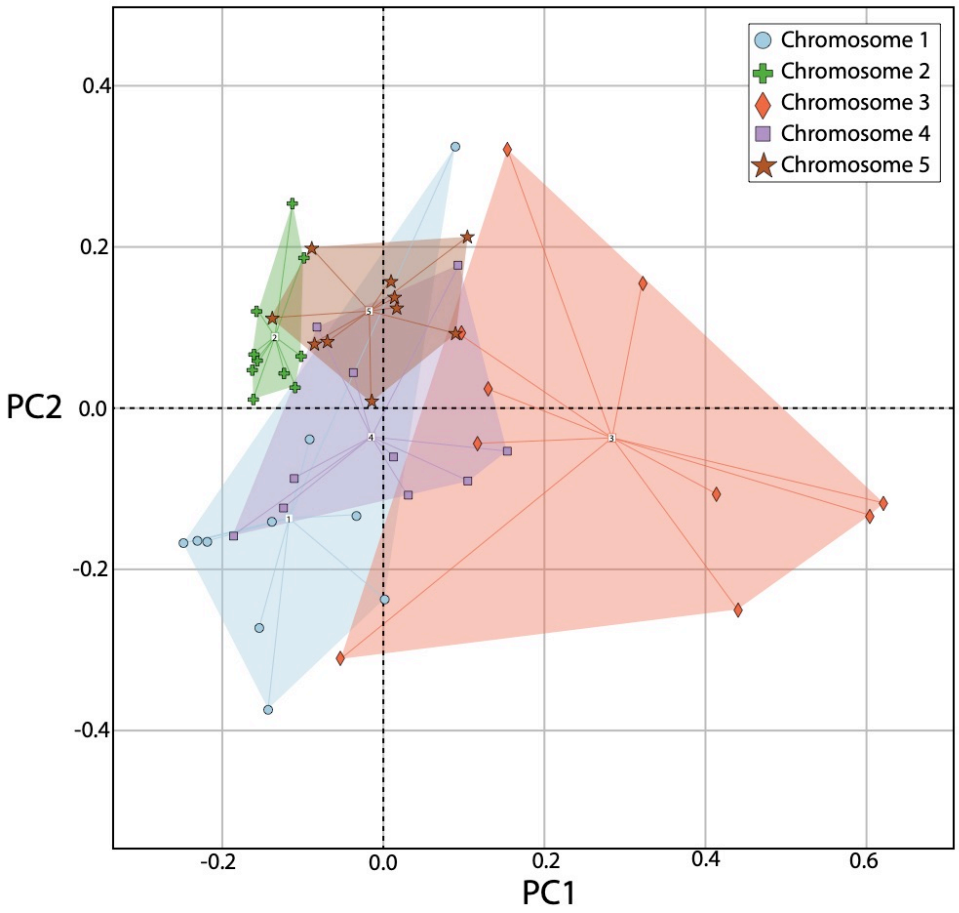


Figure 4. Multidimensional scaling of the pairwise Hellinger distances (Eq. 2) between bootstrap estimates of species trees obtained for the first five chromosomes (10 bootstrap replicates per chromosome) of the chicken genome. Bootstrapping was conducted using all available gene trees for each respective chromosome. Tree symbols and groups coloring based on chromosome.

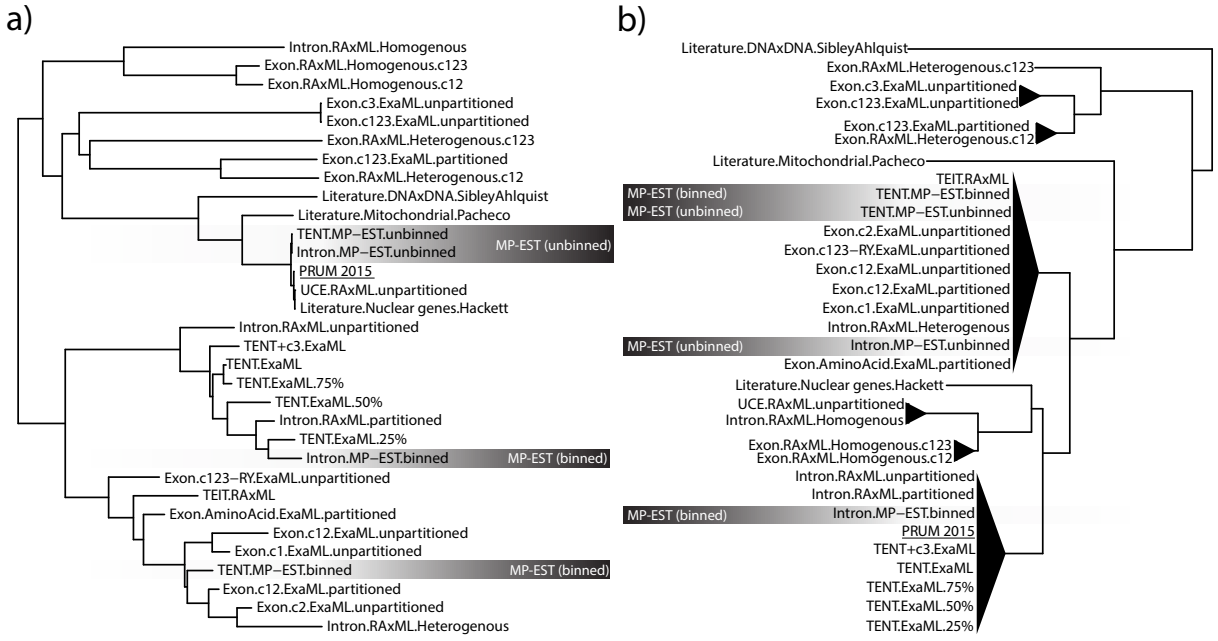


Figure 5. Clustering of species tree distances computed between 32 estimates of the avian phylogeny using the Hellinger pSTD (a) and Robinson-Foulds metric (b). Dendrograms were generated using the NJ algorithm with midpoint rooting, and tree names were obtained from the original study and reflect the particular dataset used (i.e., exons, introns, total nucleotide evidence "TENT") and approach (i.e., "unbinned" vs. "binned" MP-EST analyses). The tree inferred in Prum *et al.* (2015) is highlighted as "PRUM 2015". Clades were collapsed if the distance was zero.

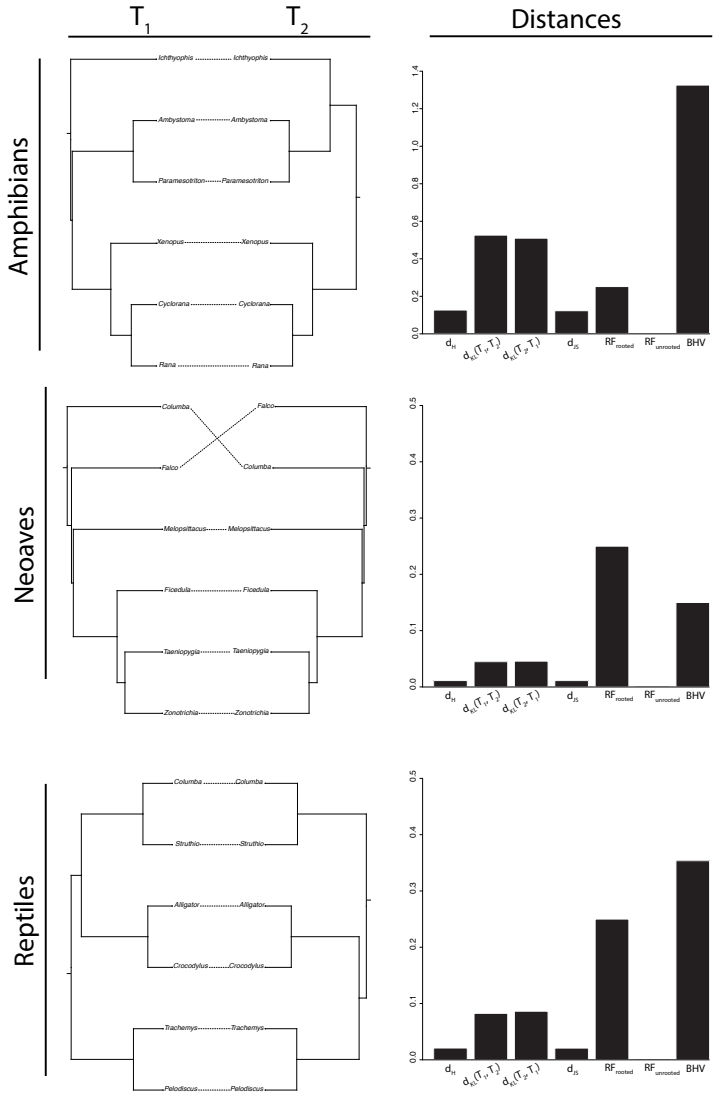


Figure 6. Measuring probabilistic distances between estimates of contentious species tree relationships for three case-studies of animals from Shen *et al*. (2017). Cophylo plots show two alternative species tree hypotheses (T_1 and T_2) for Amphibians (top), Neoaves (middle), and Reptiles (bottom). Barplots show the Hellinger distance (d_H),), Kullback-Leibler (d_{KL}) distance measured from T_1 to T_2 ($d_{KL}(T_1, T_2)$), the Kullback-Leibler (d_{KL}) measured from T_2 to T_1 ($d_{KL}(T_2, T_1)$), the Jensen-Shannon (d_{JS}), the rooted Robinson-Foulds distance (RF_{rooted}), the unrooteed Robinson-Foulds distance ($RF_{unrooted}$), and the BHV distance (d_{BHV}).

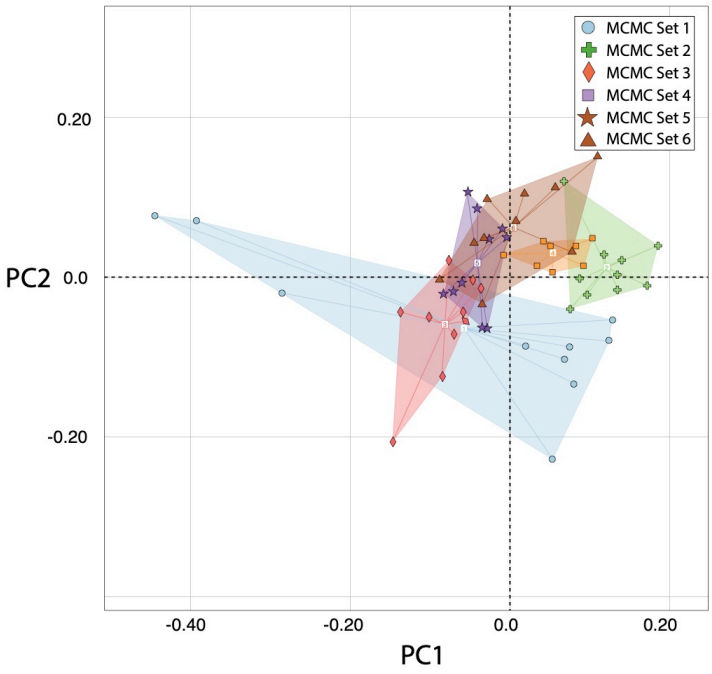


Figure 7. MDS of the pairwise Hellinger distances (Eq. 2) between 6 sets of species tree sampled from a Bayesian posterior distribution of species trees obtained via MCMC. Each of the 6 sets consists of 10 species tree sampled sequentially from the posterior MCMC samples (see text for further details).

Supplementary Methods

Probabilistic Gene Tree Distances

To provide a theoretical background to species tree distance measures, here we discuss the computation of probabilistic gene tree distances established in Garba *et al*. (2018). A gene tree ψ with n leaves represents a parametric model $\psi = \{g,\ v,\ \pi\}$ that specifies a discrete probability distribution $P(S_n|\psi = \{g,\ v,\ \pi\})$ of all possible site patterns S_n conditioned upon the gene topology (g), set of $2n - 3$ branch lengths (v), and a vector (π) containing all relevant parameters of the substitution process (i.e., relative substitution rates, equilibrium base frequencies). There are $4^n = |\mathrm{S_n}|$ possible nucleotide site patterns for a gene tree with n leaves, and the probability of each pattern s in the set $\mathrm{S_n}$ is determined by the parameters in ψ. The probabilistic gene tree distance (PGTD) between two gene trees $\psi_1 = \{g_1,\ v_1,\ \pi_1\}$ and $\psi_2 = \{g_2,\ v_2,\ \pi_2\}$ can be computed as the difference between their respective site pattern probability distributions:

$$d(\psi_1,\ \psi_2) = d(P(S_n|g_1,\ v_1,\ \pi_1)\,, P(S_n|g_2,\ v_2,\ \pi_2)) \quad (1)$$

As discussed in Garba *et al*. (2018), the distance metric function $d(\psi_1,\ \psi_2)$ can represent the Hellinger distance (d_H), the Kullback-Leibler divergence (d_{KL}), or the Jensen-Shannon distance (d_{JS}^2), which are shown below in Equations (2), (3), and (4), respectively:

$$d_H(\psi_1,\ \psi_2)^2 = \frac{1}{2}\sum_{s\epsilon S_n}\left(\sqrt{P(s|\psi_1)} - \sqrt{P(s|\psi_2)}\right)^2 \quad (2)$$

$$d_{KL}(\psi_1,\ \psi_2)\ = \sum_{s\epsilon S_n} P(s|\psi_1) \times \log\left(\frac{P(s|\psi_1)}{P(s|\psi_2)}\right) \quad (3)$$

$$d_{JS}^2(\psi_1,\ \psi_2) = \frac{1}{2} d_{KL}\left(P(s|\psi_1); \frac{P(s|\psi_1)+P(s|\psi_2)}{2}\right) +$$

$$\frac{1}{2} d_{KL}\left(P(s|\psi_2); \frac{P(s|\psi_1)+P(s|\psi_2)}{2}\right) (4)$$

For comparisons of gene trees with relatively few tips, these distances can be computed analytically. However, the total number of possible site patterns (4^n) can exert high computational cost for larger gene trees, and thus Garba *et al*. (2018) propose simulations to approximate these distances, which can be expressed in terms of their expectations. If $\boldsymbol{s}_{\mathrm{i},\psi_1}, i = 1,2,\dots,m$ are a set of m site patterns simulated from ψ_1 and similarly, $\boldsymbol{s}_{\mathrm{i},\psi_2}, i = 1,2,\dots,m$ are a set of m site patterns drawn from ψ_2, the Hellinger and Kullback-Leibler distances can be approximated using Equations (5) and (6) below, and further, the Jensen-Shannon distance can be estimated using Equation (6) and the same formula of Equation (4):

$$d_H^*(\psi_1,\ \psi_2)^2 \simeq 1 - \left(\frac{1}{2m}\right) \sum_{i=1}^{m} \left(\sqrt{\frac{P\left(\boldsymbol{s}_{\mathrm{i},\psi_1}|\psi_2\right)}{P\left(\boldsymbol{s}_{\mathrm{i},\psi_1}|\psi_1\right)}} + \sqrt{\frac{P\left(\boldsymbol{s}_{\mathrm{i},\psi_2}|\psi_1\right)}{P\left(\boldsymbol{s}_{\mathrm{i},\psi_2}|\psi_2\right)}} \right) (5)$$

$$d_{KL}^*(\psi_1,\ \psi_2) \simeq \frac{1}{m} \sum_{i=1}^{m} \log\left(\frac{P\left(\boldsymbol{s}_{\mathrm{i},\psi_1}|\psi_1\right)}{P\left(\boldsymbol{s}_{\mathrm{i},\psi_1}|\psi_2\right)}\right) \quad (6)$$

These equations measure the distance between two gene trees $d(\psi_1,\ \psi_2)$ in terms of their differences in the probability of site patterns in S_n, such that larger distances indicate greater differences in site pattern probabilities. By incorporating parameters involved in the substitution

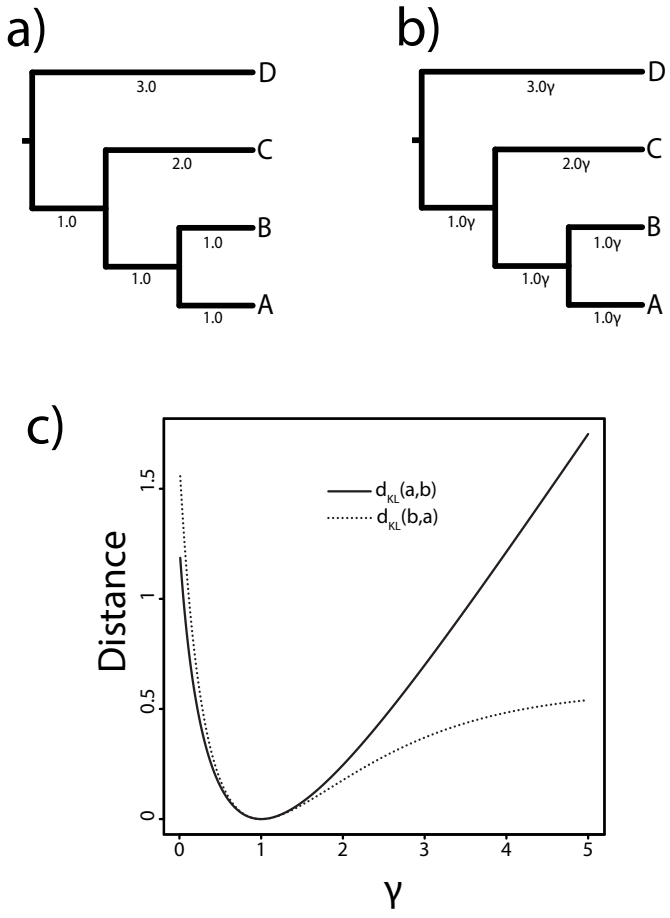
processes, these measures inherently provide more information that distances based solely on gene topologies and/or branch lengths. For example, two gene trees ψ_1 and ψ_2 could have the exact same topology ($g_1 = g_2$) and branch lengths ($v_1 = v_2$), and yet exhibit drastically different probability distributions of genetic sequence data, depending on the parameters of the substitution process (i.e., $\pi_1 \neq \pi_2$). These equations form the basis for computing probabilistic species tree distances discussed in the text and shown in Equations 1–6.

The Kullback-Leibler Distance is not a true metric

The Kullback-Leibler distance was one the first measure proposed for measuring the distance between two models in terms of their probability distributions (Kullback and Leibler 1951). Importantly, while the other two distance measures (Hellinger and Jensen-Shannon distances) are true metrics, the Kullback-Leibler distance is not a true metric because it does not satisfy the triangle inequality such that, in the context of species tree model comparisons, the distance from one species tree model ϕ_1 to another ϕ_2 is not necessarily the same distance as that from ϕ_2 to ϕ_1. In other words, Kullback-Leibler distance is not symmetric: $d_{KL}(\phi_1, \phi_2) \neq d_{KL}(\phi_2, \phi_1)$. It is important to acknowledge this property of the Kullback-Leibler distance, and we have conducted an analysis to demonstrate this nonsymmetric nature (SFig. 1). Using the same species tree models (a) and (b) shown in Figure 1, we have computed the Kullback-Leibler distances $d_{KL}(\phi_1, \phi_2)$ and $d_{KL}(\phi_2, \phi_1)$, which exhibits that these distances are not necessarily equal to one another (i.e., $d_{KL}(\phi_1, \phi_2) \neq d_{KL}(\phi_2, \phi_1)$). When the two models are more similar, such that the distance approaches zero, $d_{KL}(\phi_1, \phi_2)$ and $d_{KL}(\phi_2, \phi_1)$ also become more similar. Conversely, as the models become more dissimilar due to the larger branch length scaling on the

tree model shown in Supplementary Figure 01b, the disparately between $d_{KL}(\phi_1,\ \phi_2)$ and $d_{KL}(\phi_2,\ \phi_1)$ becomes more apparent (i.e., right side of SFig.1). This is nonsymmetrical nature is fundamental property of the Kullback-Leibler distance that applies not only to our species tree model distances, but also any other implementation of the Kullback-Leibler distance, including the original gene tree distances proposed by Garba *et al*. (2018). Nonetheless, despite this nonsymmetrical property, the Kullback-Leibler distance remains an important and useful tool that has been applied in the field of systematics, as well as biological research in general, and it has been used extensively across a broad range of fields, including mathematics, statistics, computation science, and engineering, to name a few.

SUPPLEMENTARY FIGURES



Supplementary Figure 1. Species tree models and phylogenetic distances for two scenarios (a and b) of branch scaling that demonstrate that the KL distance is not symmetric.

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.:btw795.

Adams R.H., Schield D.R., Card D.C., Castoe T.A. 2018. Assessing the Impacts of Positive Selection on Coalescent-Based Species Tree Estimation and Species Delimitation. Syst. Biol.

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

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

Bayzid M.S., Mirarab S., Boussau B., Warnow T. 2015. Weighted Statistical Binning: Enabling Statistically Consistent Genome-Scale Phylogenetic Analyses. PLoS One. 10:e0129183.

Billera L.J., Holmes S.P., Vogtmann K. 2001. Geometry of the space of phylogenetic trees. Adv. Appl. Math. 27:733–767.

Blaimer B.B., LaPolla J.S., Branstetter M.G., Lloyd M.W., Brady S.G. 2016. Phylogenomics, biogeography and diversification of obligate mealybug-tending ants in the genus Acropyga. Mol. Phylogenet. Evol. 102:20–29.

Bollback J.P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19:1171–80.

Bouckaert R.R. 2010. DensiTree: Making sense of sets of phylogenetic trees. Bioinformatics.

Branstetter M.G., Danforth B.N., Pitts J.P., Faircloth B.C., Ward P.S., Buffington M.L., Gates M.W., Kula R.R., Brady S.G. 2017. Phylogenomic Insights into the Evolution of Stinging Wasps and the Origins of Ants and Bees. Curr. Biol. 27:1019–1025.

Brown J.M. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Syst. Biol. 63:334–348.

Brown J.M., Lemmon A.R. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Syst. Biol. 56:643–55.

Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., Roychoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1932.

Buckley T.R. 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. Syst. Biol. 51:509–523.

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.

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.

Cardona G., Llabrés M., Rosselló F., Valiente G. 2009. Metrics for phylogenetic networks i: Generalizations of the robinson-foulds metric. IEEE/ACM Trans. Comput. Biol. Bioinforma. 6:46–61.

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. 2009a. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc. Natl. Acad. Sci. 106:8986–8991.

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. 2009b. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc. Natl. Acad. Sci. U. S. A. 106:8986–91.

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. 2009c. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc. Natl. Acad. Sci. 106:8986–8991.

Castoe T.A., de Koning A.P.J., Pollock D.D. 2010. Adaptive molecular convergences: Molecular evolution versus molecular phylogenetics. Commun. Integr. Biol. 3:67–69.

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.

Crandall K. a, Posada D., Vasco D. 1999. Effective population sizes : missing measures and missing concepts. Anim. Conserv. 2:317–319.

Degnan J.H., Ane C. 2017. Modeling hybridization under the network multispecies coalescent. Syst. Biol.

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

Degnan J.H., Rosenberg N.A. 2009b. 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 (N. Y). 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. 2009a. Is a new and general theory of molecular systematics emerging? Evolution (N. Y). 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. 2009c. Is a new and general theory of molecular systematics emerging? Evolution. 63:1–19.

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.

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

Estabrook G.F., McMorris F.R., Meacham C.A. 1985. Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units. Syst. Biol. 34:193–200.

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.

Felsenstein J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. Syst. Biol. 27:401–410.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein J. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. 59:139–47.

Foll M., Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. Genetics. 180:977–993.

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.

Garba M.K., Nye T.M.W., Boys R.J. 2018. Probabilistic Distances between Trees. Syst. Biol. 67:320–327.

Gelman A., Carlin J.B., Stern H.S., Rubin D.B. 2004. Bayesian Data Analysis. .

Goldman N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

Guillot G., Mortier F., Estoup A. 2005. GENELAND: A computer package for landscape genetics. Mol. Ecol. Notes. 5:712–715.

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

Haldane J.B.S. 1934. A Mathematical Theory of Natural and Artificial Selection Part X. Some Theorems on Artificial Selection. Genetics. 19:412–429.

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. Mol. Ecol. Evol. Approaches Appl.:435–449.

Hillis D.M., Heath T.A., St. John K. 2005. Analysis and visualization of tree space. Syst. Biol.

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.

Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol.

Biol. Evol.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Al. E. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science (80-.). 346:1320–1331.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Alfaro-Núñez A., Narula N., Liu L., Burt D., Ellegren H., Edwards S. V, Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2015. Phylogenomic analyses data of the avian phylogenomics project. Gigascience. 4:4.

Jensen J.D., Foll M., Bernatchez L. 2016. The past, present and future of genomic scans for selection. Mol. Ecol. 25:1–4.

JeŠovnik A., Sosa-Calvo J., Lloyd M.W., Branstetter M.G., FernÁndez F., Schultz T.R. 2017. Phylogenomic species delimitation and host-symbiont coevolution in the fungus-farming ant genus Sericomyrmex Mayr (Hymenoptera: Formicidae): ultraconserved elements (UCEs) resolve a recent radiation. Syst. Entomol. 42:523–542.

Jombart T., Kendall M., Almagro-Garcia J., Colijn C. 2017. treespace: Statistical exploration of landscapes of phylogenetic trees. Mol. Ecol. Resour.

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

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

Kelchner S.A., Thomas M.A. 2007. Model use in phylogenetics: nine key questions. Trends Ecol. Evol. 22:87–94.

Kim J. 2000. Slicing hyperdimensional oranges: The geometry of phylogenetic estimation. Mol. Phylogenet. Evol. 17:58–75.

Kingman J. 1982. The coalescent. Stoch. Proc. Appl. 13:235–48.

Knowles L.L. 2009. Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. Syst. Biol. 58:463–467.

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

Kuhner M.K., Yamato J. 2015. Practical performance of tree comparison metrics. Syst. Biol. 64:205–214.

Kuhner M.K., Yamato J., Felsenstein J. 1995. Estimating effective population size and mutation

rate from sequence data Using Metropolis-Hastings Sampling. Genetics. 140:1421–1430.

Kullback S., Leibler R.A. 1951. On information and sufficiency. Ann. Math. Stat. 22:79–86.

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.

Leache 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.

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.

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

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.

Leigh J.W., Susko E., Baumgartner M., Roger A.J. 2008a. Testing congruence in phylogenomic analysis. Syst. Biol. 57:104–115.

Leigh J.W., Susko E., Baumgartner M., Roger A.J. 2008b. Testing congruence in phylogenomic analysis. Syst. Biol. 57:104–115.

Lemmon A.R., Moriarty E.C. 2004. The importance of proper model assumption in bayesian phylogenetics. Syst. Biol. 53:265–277.

Li H., Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature. 475:493–496.

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

Liu L., Edwards S. V. 2015. Comment on “Statistical binning enables an accurate coalescent-based estimation of the avian tree.” Science (80-.). 350:171.

Liu L., Edwards S. V, Mirarab S., Bayzid M.S., Boussau B., Warnow T. 2015a. Comment on “Statistical binning enables an accurate coalescent-based estimation of the avian tree.”

Science (80-.). 350:171 LP-171.

Liu L., Wu S., Yu L. 2015b. Coalescent methods for estimating species trees from phylogenomic data. J. Syst. Evol. 53:380–390.

Liu L., Yu L. 2010. Phybase: an R package for species tree analysis. Bioinformatics. 26:962–3.

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

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

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. 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.

Malmstrøm M., Matschiner M., Tørresen O.K., Star B., Snipen L.G., Hansen T.F., Baalsrud H.T., Nederbragt A.J., Hanel R., Salzburger W., Stenseth N.C., Jakobsen K.S., Jentoft S. 2016. Evolution of the immune system influences speciation rates in teleost fishes. Nat. Genet. 48:1204–1210.

Martin A.D., Quinn K.M., Park J.H. 2011. MCMCpack : Markov Chain Monte Carlo in R. J. Stat. Softw. 42:1–21.

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.

Mendes F.K., Hahn M.W. 2016. Gene tree discordance causes apparent substitution rate variation. Syst. Biol. 65:711–721.

Mendes F.K., Hahn M.W. 2017. Why Concatenation Fails Near the Anomaly Zone. Syst. Biol.

Meng C., Kubatko L.S. 2009. Detecting hybrid speciation in the presence of incomplete lineage

sorting using gene tree incongruence: A model. Theor. Popul. Biol. 75:35–45.

Mirarab S., Bayzid M.S., Boussau B., Warnow T. 2015. Response to Comment on "Statistical binning enables an accurate coalescent-based estimation of the avian tree." Science (80-.). 350:171–171.

Mirarab S., Bayzid S.M., Boussau B., Warnow T. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science (80-.). 346:1250463–1250463.

Moulton V., Steel M. 2004. Peeling phylogenetic "oranges." Adv. Appl. Math. 33:710–727.

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

Nakhleh L. 2010. A metric on the space of reduced phylogenetic networks. IEEE/ACM Trans. Comput. Biol. Bioinforma. 7:218–222.

Narum S.R., Hess J.E. 2011. Comparison of FST outlier tests for SNP loci under selection. Mol. Ecol. Resour. 11:184–194.

Nei M. 1976. Mathematical models of speciation and genetic distance. Popul. Genet. Ecol. Acad. Press. New York.:723–766.

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

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

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.

O'Reilly P.F., Birney E., Balding D.J. 2008. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. Genome Res. 18:1304–1313.

Ogilvie H.A., Bouckaert R.R., Drummond A.J. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol. Biol. Evol.

Oleksyk T.K., Smith M.W., O'Brien S.J. 2010. Genome-wide scans for footprints of natural selection. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 365:185–205.

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.

Owen M. 2011. Computing Geodesic Distances in Tree Space. SIAM J. Discret. Math. 25:1506–1529.

Owen M., Provan J.S. 2011. A fast algorithm for computing geodesic distances in tree space.

IEEE/ACM Trans. Comput. Biol. Bioinforma. 8:2–13.

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

Pavlidis P., Jensen J.D., Stephan W., Stamatakis A. 2012. A critical assessment of storytelling: Gene ontology categories and the importance of validating genomic scans. Mol. Biol. Evol. 29:3237–3248.

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.

Platt R.N., Faircloth B.C., Sullivan K.A.M., Kieran T.J., Glenn T.C., Vandewege M.W., Lee T.E., Baker R.J., Stevens R.D., Ray D.A. 2018. Conflicting Evolutionary Histories of the Mitochondrial and Nuclear Genomes in New World Myotis Bats. Syst. Biol. 67:236–249.

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015a. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature. 526:569–573.

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015b. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature. 526:569–573.

Qu Y., Zhang R., Quan Q., Song G., Li S.H., Lei F. 2012. Incomplete lineage sorting or secondary admixture. Disen. Hist. divergence from Recent gene flow Vinous-throated parrotbill (Paradoxornis webbianus). 21:6117–6133.

R Core Team. 2015. R: A Language and Environment for Statistical Computing. R Found. Stat. Comput. Vienna Austria. 0:{ISBN} 3-900051-07-0.

R Core Team. 2017. R Development Core Team. R A Lang. Environ. Stat. Comput.

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.

Rambaut A., Drummond A.J. 2016. TreeAnnotator v1.8.4. .

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

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

Reddy S., Kimball R.T., Pandey A., Hosner P.A., Braun M.J., Hackett S.J., Han K.L., Harshman

J., Huddleston C.J., Kingston S., Marks B.D., Miglia K.J., Moore W.S., Sheldon F.H., Witt C.C., Yuri T., Braun E.L. 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. Syst. Biol. 66:857–879.

Reid N.M., Hird S.M., Brown J.M., Pelletier T.A., McVay J.D., Satler J.D., Carstens B.C. 2014. Poor fit to the multispecies coalescent is widely detectable in empirical data. Syst. Biol. 63:322–333.

Revell L.J. 2012. phytools: An R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3:217–223.

Robinson D., Foulds L. 1979. Comparison of weighted labelled trees. Lect. Notes Math. 748:119–126.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Roch S., Nute M., Warnow T. 2018. Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods. arXiv Prepr. arXiv1803.02800.

Roch S., Warnow T. 2015. On the Robustness to Gene Tree Estimation Error (or lack thereof) of Coalescent-Based Species Tree Methods. Syst. Biol. 64:663–676.

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.

Schield D.R., Adams R.H., Card D.C., Perry B.W., Pasquesi G.M., Jezkova T., Portik D.M., Andrew A.L., Spencer C.L., Sanchez E.E. 2017. Insight into the roles of selection in speciation from genomic patterns of divergence and introgression in secondary contact in

venomous rattlesnakes. Ecol. Evol.

Schield D.R., Card D.C., Adams R.H., Jezkova T., Reyes-Velasco J., Proctor F.N., Spencer C.L., Herrmann H.W., Mackessy S.P., Castoe T.A. 2015. Incipient speciation with biased gene flow between two lineages of the Western Diamondback Rattlesnake (Crotalus atrox). Mol. Phylogenet. Evol. 83:213–223.

Schliep K.P. 2011. phangorn: Phylogenetic analysis in R. Bioinformatics.

Schluter D. 2009. Evidence for Ecological Speciation and Its Alternative. Science (80-.). 323:737–741.

Schrider D., Shanku A.G., Kern A.D. 2016. Effects of linked selective sweeps on demographic inference and model selection. bioRxiv.:47019.

Shaffer H.B., Thomson R. 2007. Delimiting Species in Recent Radiations. Syst. Biol. 56:896–906.

Shapiro M.D., Kronenberg Z., Li C., Domyan E.T., Pan H., Campbell M., Tan H., Huff C.D., Hu H., Vickrey A.I. 2013. Genomic diversity and evolution of the head crest in the rock pigeon. Science (80-.). 339:1063–1067.

Shen X.-X., Hittinger C.T., Rokas A. 2017a. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1:0126.

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

Shi F., Feng Q., Chen J., Wang L., Wang J. 2013. Distances between phylogenetic trees: A survey. Tsinghua Sci. Technol. 18:490–499.

Shimodaira H., Hasegawa M. 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Mol. Biol. Evol. 16:1114–1116.

Solis-Lemus C., Knowles L.L., An C. 2015. Bayesian species delimitation combining multiple genes and traits in a unified framework. Evolution (N. Y). 69:492–507.

Solís-Lemus C., Yang M., Ané C. 2016. Inconsistency of species-tree methods under gene flow. Syst. Biol.:syw030.

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

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.

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

Streicher J.W., Miller E.C., Guerrero P.C., Correa C., Ortiz J.C., Crawford A.J., Pie M.R., Wiens J.J. 2018. Evaluating methods for phylogenomic analyses, and a new phylogeny for a major frog clade (Hyloidea) based on 2214 loci. Mol. Phylogenet. Evol. 119:128–143.

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

Sullivan J., Swofford D.L. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mamm. Evol. 4:77–86.

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

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

Tarver J.E., Dos Reis M., Mirarab S., Moran R.J., Parker S., O'Reilly J.E., King B.L., O'Connell M.J., Asher R.J., Warnow T., Peterson K.J., Donoghue P.C.J., Pisani D. 2016. The interrelationships of placental mammals and the limits of phylogenetic inference. Genome Biol. Evol. 8:330–344.

Tavaré S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26:119–164.

Teshima K.M., Coop G., Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? Genome Res. 16:702–712.

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. 97:5313–5316.

Vitti J.J., Grossman S.R., Sabeti P.C. 2013. Detecting Natural Selection in Genomic Data. Annu. Rev. Genet. 47:97–120.

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.

Waddell P.J., Ota R., Penny D. 2009. Measuring fit of sequence data to phylogenetic model: Gain of power using marginal tests. J. Mol. Evol. 69:289–299.

Wakeley J. 2008. Coalescent Theory: An Introduction. .

Walker J.F., Brown J.W., Smith S.A. 2018. Analyzing contentious relationships and outlier genes in phylogenomics. bioRxiv. 0:115774.

Wang J. 2005. Estimation of effective population sizes from data on genetic markers. Philos.

Trans. R. Soc. Lond. B. Biol. Sci. 360:1395–1409.

Warnow T. 2015. Concatenation analyses in the presence of incomplete lineage sorting. PLoS Curr. 7.

Warren W.C., Hillier L.W., Tomlinson C., Minx P., Kremitzki M., Graves T., Markovic C., Bouk N., Pruitt K.D., Thibaud-Nissen F., Schneider V., Mansour T.A., Brown C.T., Zimin A., Hawken R., Abrahamsen M., Pyrkosz A.B., Morisson M., Fillon V., Vignal A., Chow W., Howe K., Fulton J.E., Miller M.M., Lovell P., Mello C. V., Wirthlin M., Mason A.S., Kuo R., Burt D.W., Dodgson J.B., Cheng H.H. 2017. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3: Genes|Genomes|Genetics.

Watterson G.A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7:256–276.

Wright S. 1949. The Genetical Structure of Populations. Ann. Eugen. 15:323–354.

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

Yang Z.H. 1997. On the estimation of ancestral population sizes of modern humans. Genet. Res. 69:111–116.

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

Zhang C., Zhang D.X., Zhu T., Yang Z. 2011b. 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 S., Degnan J.H. 2017. Displayed trees do not determine distinguishability under the network multispecies coalescent. Syst. Biol. 66:283–298.

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.