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

[§]**Correspondence**: Todd A. Castoe, Department of Biology, University of Texas at Arlington, Arlington, TX 76010 USA.

Email: todd.castoe@uta.edu *phone:* 817-272-9084 *fax:* 817-272-9615

Running Head: Probabilistic distances between species tree models

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 topology 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.

Keywords: information theory, model comparison, species tree estimation, hypothesis testing.

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. 2017; 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 2003). 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 2009). 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 2009; 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 2003). 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 tree topologies conditioned upon a specific species tree and set of divergence times (in coalescent units). Just as Garba *et al*. (2018) viewed gene trees as parametric models that can be compared in terms of 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 gene topologies.

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 tree topologies is $|G_n| = \frac{(2n-3)!}{2^{n-2}(n-2)!}$, and the probability of a particular topology 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 topology 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 tree topologies 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:

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

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

$$d_{JS}^2(\phi_1,\ \phi_2) = \frac{1}{2}d_{KL}\left(P(g\mid\phi_1);\frac{P(g\mid\phi_1)+P(g\mid\phi_2)}{2}\right) + \frac{1}{2}d_{KL}\left(P(g\mid\phi_2);\frac{P(g\mid\phi_1)+P(g\mid\phi_2)}{2}\right) \quad (4)$$

We have implemented these equations in an R software package (pSTDistancesR) that uses HYBRID-COAL (Zhu and Degnan 2017) to calculate gene tree topology probabilities (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 gene tree topologies, and for each topology 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 tree topologies $|G_n|$ 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 topologies from each species tree and approximate the Hellinger and Kullback-Leibler distance between ϕ_1 and ϕ_2 as:

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

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

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 topology 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 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 topology (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 topology 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. 2010). 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. 2010) 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. 2017), 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 iterations 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 Jensen-Shannon 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. 2011; 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. 2009, 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 2003; 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 compare 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). Although here we have focused on the derivation of species tree

distances using gene tree topology probabilities alone, effectively incorporating full gene tree probabilities (i.e., including topology and coalescent time variation) may be useful future extensions of these distances.

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. Previous studies have demonstrated that species trees are identifiably from the distributions of unrooted gene trees (Allman et al. 2011) and pSTD reflect this property. 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 indistinguishable (i.e., pSTD = 0) or distinguishable (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.

REFERENCES

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.

Allman E.S., Degnan J.H., Rhodes J.A. 2011. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J. Math. Biol.

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

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.

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

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

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

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

Edwards S. V. 2009. 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.

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.

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.

Heled J., Drummond A.J. 2010. Bayesian Inference of Species Trees from Multilocus Data. Mol. Biol. Evol. 27:570–580.

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

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.

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.

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

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

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

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.

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

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

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 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.

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

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

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.

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.

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

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

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.

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.

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

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

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

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

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

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.

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

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

Zhu S., Degnan J.H. 2017. Displayed trees do not determine distinguishability under the network multispecies coalescent. Syst. Biol. 66:283–298.

FIGURE LEGENDS

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 the height of 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).