2
3 Expected pairwise congruence among gene trees under
4 the coalescent model
5

6 Yuan Tian¹, Laura S. Kubatko^{1,2*}

7
8 ¹*Department of Evolution, Ecology, and Organismal Biology, The Ohio State University*

9 ²*Department of Statistics, The Ohio State University*
10

11 **To whom correspondence should be addressed;*

12 *E-mail: kubatko.2@osu.edu*

13 Author to receive proofs:

14 Laura Kubatko

15 The Ohio State University

16 Department of Statistics, 404 Cockins Hall

17 1958 Neil Avenue

18 Columbus, OH 43210

19 FAX: 614-292-2096

20 E-mail: kubatko.2@osu.edu
21

Abstract— Although it is widely appreciated that gene trees may differ from the overall species tree and from one another due to various evolutionary processes (e.g., incomplete lineage sorting (ILS), horizontal gene transfer, etc.), the extent of this incongruence is rarely quantified and discussed. Here we consider the expected amount of incongruence arising from ILS, as modeled by the coalescent process. In particular, we compute the probability that two gene trees randomly sampled from the same species tree agree with one another as well as the distribution of the Robinson-Foulds distance between them, for species trees with three to eight taxa. We demonstrate that, as expected under the coalescent model, the amount of discordance is affected by species tree-specific factors such as speciation times and effective population sizes for the species under consideration. Our results highlight the fact that substantial discordance may occur, even when the number of species is very small, which has implications both for larger taxon samples and for any method that uses estimated gene trees as the basis for further statistical inference. The amount of incongruence is substantial enough that such methods may need to be modified to account for variability in the underlying gene trees.

Keywords: coalescent, incomplete lineage sorting, gene tree, species tree, incongruence, RF distance, mapping traits, comparative methods.

Incomplete lineage sorting (ILS) has long been recognized to be a predominant cause of substantial variation in the evolutionary trees for individual genes, leading to incongruence among gene trees (Pamilo and Nei, 1988; Takahata, 1989; Hein, 1993; Maddison, 1997; Sang and Zhong, 2001; Kubatko, 2009; Liu *et al.*, 2010; Bayzid and Warnow, 2012). Coalescent theory (Kingman, 1982a,b; Tajima, 1983; Tavaré, 1984; Takahata and Nei, 1985; Pamilo and Nei, 1988; Rosenberg, 2002; Rannala and Yang, 2003; Degnan and Salter, 2005) is commonly used to model ILS, and the predictions concerning agreement among gene trees made by the coalescent model are becoming increasingly well-understood. For example, it is widely appreciated that the extent of gene tree incongruence depends on characteristics of the species tree, such as branch lengths and effective population sizes, suggesting the importance of considering the coalescent process in phylogenetic studies (Degnan and Salter, 2005; Degnan and Rosenberg, 2006). Indeed, most current methods of multi-locus phylogenetic inference incorporate the coalescent process to model ILS (Edwards *et al.*, 2016). In this study, we consider the extent of congruence between a pair of gene trees that is expected under the coalescent process for a small number of taxa. Our study is motivated by four current challenges in empirical multilocus phylogenetics.

First, we consider the problem of assessing the extent to which empirical data fit the predictions made **by** the coalescent model. In particular, a recent study (Simmons *et al.*, 2016) examined eight empirical data sets and presented the pairwise congruence among gene trees for each of these studies. Simmons *et al.* (2016) reported that the average pairwise congruence among gene trees varied greatly both between studies and sometimes within a study. However, the study did not consider the extent of pairwise congruence that would be *expected* among gene trees under the coalescent model, making it difficult to assess fit to the coalescent model. Here we consider this problem from an **analytic** perspective by computing the probability of **sampling** two identical gene tree topologies, as well as the probability

distribution of the pairwise Robinson-Foulds distance (hereafter “RF distance”, Robinson and Foulds, 1979) among possible gene trees, **for model species trees with three to eight taxa.**

Second, note that a well-known example of incongruent gene trees is the commonly observed conflict between mitochondrial and nuclear phylogenies (Ferris *et al.*, 1983; Moore, 1995; Sota and Vogler, 2001; Shaw, 2002; McCracken and Sorenson, 2005; Leaché and McGuire, 2006; Robertson *et al.*, 2006; Peters *et al.*, 2007; Good *et al.*, 2008). Conflicting mitochondrial and nuclear gene trees are often attributed to the low mutation rate of nuclear DNA sequences, hybridization or other horizontal gene transfer, and natural selection (Avice *et al.*, 1987; Rand, 2001; Sanderson and Shaffer, 2002; Funk and Omland, 2003; Ballard and Whitlock, 2004; Ballard and Rand, 2005; Spinks and Shaffer, 2009; Roos *et al.*, 2011), while the effect of ILS is often overlooked. Incongruence among gene trees is also extensively observed in other phylogenetic settings, such as conflicting plastid and nuclear gene trees, or different gene trees among nuclear genes (Cranston *et al.*, 2009; Moyer *et al.*, 2009; Bell and Hyvönen, 2010). Our computation of the expected extent of pairwise gene tree incongruence **for specific examples** serves to highlight the important role that ILS is likely playing in these empirical observations **more generally.**

Third, we consider gene tree incongruence as a potential source of bias when mapping character traits onto phylogenetic trees. In this setting, it is common to use a single tree (normally an estimated species tree) as the phylogenetic framework onto which character traits (e.g., nucleotides, morphological traits, behavioral traits, etc.) are mapped. For example, Hahn and Nakhleh (2015) recently discussed the risk of ignoring variation in gene tree topologies and mapping characters onto a single representation of the species tree. Quantifying the extent to which the two underlying phylogenies may vary using the RF distance **for specific examples**, as we have done here, gives insights into the extent of the potential bias in mapping character traits onto a fixed phylogeny.

Fourth, we note that a similar issue may exist in phylogenetic comparative studies. A

number of comparative studies look for correlation in two or more traits after adjusting for a single species tree. Often this single tree is a well-resolved species tree that is believed to be a reliable estimate of the species-level evolutionary relationships (Wiegmann *et al.*, 2009; Misof *et al.*, 2014). It is clear, however, that each of the traits under consideration may have its own gene tree, and these gene trees may vary substantially from the species tree. Although the uncertainty in phylogenetic relationships has been taken into account in several approaches (Richman and Price, 1992; Huelsenbeck *et al.*, 2000; Lutzoni *et al.*, 2001; Pagel and Lutzoni, 2002; Huelsenbeck and Rannala, 2003; Pagel *et al.*, 2004), most of them focus on a lack of resolution in the overall species tree estimate (Hahn and Nakhleh, 2015) or gene tree estimation error (Simmons *et al.*, 2016), rather than on genuine variation in the underlying histories. Again, quantification of the extent of incongruence can give insights into how robust comparative procedures might be to gene tree variation.

We next briefly describe our computational methods for assessing the extent of gene tree incongruence under the coalescent model, and then provide our results. While we consider only small trees here (in the range of three to eight taxa), these examples serve to highlight the fact that gene tree incongruence is widespread, even when the number of taxa under consideration is not large. We discuss the implications of our findings in the Discussion.

METHODS

In our study, the probability that two topologically identical gene trees are observed from the same species tree under the coalescent model is computed assuming no recombination within the two loci, and free recombination between the two loci. We further assume that the only evolutionary process generating discord between the two gene trees is the coalescent process (i.e., there is no hybridization or other horizontal gene transfer, and no gene duplication/loss). We consider the case in which one gene lineage is sampled for each species – when additional samples within a species are sampled, the situation will be even more

complex in the sense that there is the possibility of even greater discord between gene trees. Varying population sizes and branch lengths of the species tree are examined to study their effect on congruence among gene trees. Furthermore, the probability distribution of pairwise RF distances between the two gene trees under the coalescent model is computed. Note that the expected gene tree congruence and the distribution of RF distances are computed based on the study of Degnan and Salter (2005). In their paper, they derived a method for computing the distribution of gene tree topologies given a bifurcating species tree with an arbitrary number of taxa, when one gene is sampled for each species. This method was implemented in the computer program COAL. When a given species tree (with branch lengths in coalescent units) is input into COAL, the probability of each possible gene tree topology will be calculated. All calculations presented here are carried out using COAL (Degnan and Salter, 2005).

Gene tree incongruence for species trees with varying numbers of taxa

Completely asymmetric species trees and selected symmetric species trees with up to eight taxa are used to compute the probability of observing two topologically identical gene trees. We select three different levels for the internal branch lengths (time between one speciation event to the following speciation event) for each species tree. Species trees with all internal branch lengths equal to 2.0 (all branch lengths in coalescent units) are labeled “Long species trees”, species trees with all internal branch lengths equal to 1.0 are labeled “Medium species trees”, and species trees with all internal branch lengths equal to 0.5 are labeled “Short species trees”. Under the coalescent model, it is expected that gene trees sampled from species trees with “long” internal branch lengths will have little ILS, while species trees with “short” internal branch lengths are expected to generate gene trees with extensive ILS. It is impossible to explore the space of all species trees exhaustively, even for small numbers of taxa, because there are infinitely many choices for internal branch lengths. Our choice of

“long”, “medium” and “short” settings here is meant to span the range of possibilities. The sum of squares of all possible gene tree probabilities gives the total probability of sampling two topologically identical gene trees for a given species tree.

Gene tree incongruence probability for species trees with varying internal branch lengths

To examine the effect of different branch lengths on the extent of gene tree incongruence, we use species trees of four taxa and five taxa. For the four-taxon case, the species tree is denoted as $((((A : T_0, B : T_0) : T_1, C : (T_0 + T_1)) : T_2, D : (T_0 + T_1 + T_2)))$, where T_1 and T_2 are the two internal branch lengths. Note that the value of T_0 does not affect the probability of gene trees that are embedded in the species tree, because only one individual is sampled in each species. In our calculations, we use 100 equally spaced values from 0.0 to 5.0 for T_1 and T_2 . For the five taxa case, the species tree is denoted as $(((((A : T_0, B : T_0) : T_1, C : (T_0 + T_1)) : T_2, D : (T_0 + T_1 + T_2)) : T_3, E : (T_0 + T_1 + T_2 + T_3)))$. Three levels (0.1, 1.0, and 5.0) are selected for T_3 , and 100 equally spaced values from 0.0 to 5.0 are used for T_1 and T_2 , similar to the four-taxon case. The probability of getting two topologically identical gene trees for each choice of branch lengths is calculated as described in the last section.

Distribution of pairwise RF distance among gene trees

It is also of interest to measure the pairwise distance between a pair of randomly sampled gene trees. We use the RF distance to measure this and consider the distribution of the pairwise RF distances among gene trees under the same species tree. The pairwise RF distances between pairs of gene trees is measured for five-taxon species trees and for seven-taxon species trees with varying internal branch lengths, to provide an example of how this distribution changes as both internal branch lengths and number of taxa vary.

The five-taxon species trees have the format $(((((A : T_0, B : T_0) : T_1, C : (T_0 + T_1)) :$

$T_2, D : (T_0 + T_1 + T_2) : T_3, E : (T_0 + T_1 + T_2 + T_3)$. One of the three internal branch lengths
 T_1, T_2 , and T_3 is varied from 0.5 to 5.0, while the other two branch lengths, as well as T_0 , are
 fixed at 1.0. Both rooted and unrooted RF distances are calculated by the RF.dist function
 given in the phangorn (Steel and Penny, 1993) package in R (R Core Team, 2016). The
 pairwise RF distance distribution is computed by summing up the probabilities (weighted
 by the gene tree probabilities) of getting a particular RF distance.

The seven-taxon species trees have an increased computational complexity, because there
 are 135,135 possible gene trees in total. Thus, we select the seven-taxa yeast species tree
 reported by Rokas et al. (2003) as the model species tree. To test the species tree with
 varying branch lengths, the estimated branch lengths from Fan and Kubatko (2011) are
 scaled by 0.8-, 1.0-, and 1.2-fold. The distribution of pairwise RF distances is **computed** in
 the same way as in the five-taxon case.

RESULTS

Gene tree incongruence increases with an increasing number of taxa

The probability of independently drawing two gene trees that match from asymmetric species
 trees with varying numbers of taxa is shown in Figure 1(a), **while Figure 1(b) represents the**
probability of getting matching gene trees from symmetric species trees. In both panels of
Figure 1, three series of species trees with long internal branch lengths (2.0, the dotted line),
medium internal branch lengths (1.0, the solid line) and short internal branch lengths (0.5,
the dashed line) are used to calculate the gene tree incongruence. The results are consistent
with our expectation that the probability of getting two topologically identical gene trees
decreases as the number of taxa increases, and it is clear that this probability will become
even smaller when there are more than eight taxa in the species tree. Notably, the probability
that two gene trees are topologically identical is surprisingly low for only eight taxa with

both medium and short internal branch lengths, illustrating that two gene trees will rarely agree with each other for a large species tree (with more than eight taxa) when the internal branch lengths are not long enough. Even for the species trees with long internal branch lengths (2.0), the probability of sampling two topologically-matched gene trees is still below 40%. The results shown in Figure 1 also demonstrate that species trees with shorter branch lengths are more likely to generate incongruent gene trees, when the number of taxa is fixed. Even for a three-taxon species tree, it is still quite possible that two topologically different gene trees will be sampled (e.g., this happens with 56.4% of the time for the three-taxon species tree with “short” internal branch lengths (0.5)).

Gene tree incongruence decreases with increasing branch lengths of the species tree

To further explore the effect of branch lengths on the extent of gene tree incongruence, we use four-taxon and five-taxon asymmetric species trees with varying branch lengths. There are two internal branch lengths in the four-taxon species trees, and each is examined at 100 equally spaced values from 0.0 to 5.0 (Figure 2(a)). Unless both internal branches, T_1 and T_2 , are long enough (larger than 2.0), it is very likely (prob. > 40%) that two topologically different gene trees will be sampled. It is highly possible (prob. > 80%) that the two gene trees match each other when both T_1 and T_2 exceed 3.0, which is a considerably long interval between speciation times. Notably, short speciation times will create more evenly distributed gene tree topologies, therefore increasing gene tree incongruence. Similar gene tree matching probability patterns appear in the five-taxon species tree cases (Figure 2(b)-(d)). We examine three different levels of T_3 (0.1, 1.0 and 5.0), and the overall probability that two sampled gene trees match increases with an increasing value of T_3 . It is not surprising that with a short branch length of T_3 (less than 0.5), it is not possible for the probability that two gene trees match to exceed 60%. From Figure 2, it is clear that there are positive correlations between the extent of gene tree congruence and any internal branch lengths.

Figure 1: Gene tree congruence for species trees with different numbers of taxa. The x-axis shows the number of taxa in the completely asymmetric species tree, and the y-axis shows the probability of sampling two **topologically** identical gene trees. **Species trees with long internal branch lengths (2.0) are denoted with the dotted line. Species trees with medium internal branch lengths (1.0) are denoted with the solid line. The dashed line denotes the species trees with short internal branch lengths (0.5).** (a) **Completely asymmetric species trees with species number varying from three to eight.** (b) **Selected symmetric species trees.** Let x denotes the internal branch lengths (0.5, 1.0, or 2.0), the symmetric species trees are: Three taxa: $((A:x, B:x):x, C:2x)$; Four taxa: $((A:x, B:x):x, (C:x, D:x):x)$; Five taxa: $((A:x, B:x):x, C:2x):x, (D:2x, E:2x):x)$; Six taxa: $((A:x, B:x):x, C:2x):x, ((D:x, E:x):x, F:2x):x)$; Seven taxa: $((A:x, B:x):x, (C:x, D:x)):x, ((E:x, F:x):x, G:2x):x)$; Eight taxa: $((A:x, B:x):x, (C:x, D:x)):x, ((E:x, F:x):x, (G:x, H:x)):x)$.

Figure 2: Gene tree congruence for four-taxon and five-taxon species trees with varying internal branch lengths. The internal branch lengths T_1 and T_2 are denoted on the x-axis and the y-axis, respectively. The gene tree matching probability (varying from 0 to 1) is divided into 40 levels as shown in the color legend. (a) Four-taxon species tree $((A : T_0, B : T_0) : T_1, C : (T_0 + T_1)) : T_2, D : (T_0 + T_1 + T_2))$, where 100 equally spaced values from 0.0 to 5.0 for T_1 and T_2 are examined. (b) - (d) Five-taxon species tree $((((A : T_0, B : T_0) : T_1, C : (T_0 + T_1)) : T_2, D : (T_0 + T_1 + T_2)) : T_3, E : (T_0 + T_1 + T_2 + T_3))$. Three levels of T_3 (0.1, 1.0, and 5.0) are examined. The other two branch lengths T_1 and T_2 are set as in the four-taxon case.

215 The rooted and unrooted pairwise RF distance distributions for gene trees sampled from
 216 several five-taxon species trees are shown in Figure 3. For a five-taxon species tree, the
 217 largest rooted RF distance is 6 and the largest unrooted RF distance is 4. When the RF
 218 distance is 0, the two gene trees have identical topologies. The effect of each internal branch
 219 length is examined by fixing the other branch lengths. It is quite reasonable that rooted
 220 gene trees (Figure 3 (a), (c), (e)) have a lower probability of matching each other or having
 221 similar topologies than the unrooted gene trees (Figure 3 (b), (d), (f)). However, even for
 222 the unrooted gene trees, the probability of **sampling** different gene trees exceeds 40% in all
 223 of the species trees we examined. The distributions of RF distance shown in Figure 3 also
 224 clearly indicate that two gene trees are more likely to agree with each other when the interval
 225 between speciation events is longer (Figure 3 (a)-(e)). The only exception is the unrooted RF
 226 distance of the species tree with fixed T_1 and T_2 (Figure 3 (f)). When T_3 varies from 0.5 to
 227 5.0, the distribution of unrooted RF distance maintains the same. This phenomenon is not
 228 difficult to explain. As the most ancient internal branch, the length of T_3 only affects the
 229 root position of the gene trees, thus it does not change the topology of the unrooted gene
 230 trees.

231 Pairwise RF distances among gene trees are also measured in a seven-taxa yeast species
 232 tree (Rokas *et al.*, 2003; Fan and Kubatko, 2011). We scaled the estimated branch lengths
 233 of this yeast species tree (Fan and Kubatko, 2011) by 0.8-, 1.0-, and 1.2-fold to assess the
 234 effect of branch length on the distribution of RF distances (Table 1). Again, the branch
 235 lengths play an important role in the distribution of pairwise RF distances. Considering RF
 236 distances less than or equal to 4 as “similar”, and RF distances larger than 6 as “different”,
 237 the probability of getting “similar” gene trees is consistently decreasing with decreasing
 238 internal branch lengths. Conversely, it is more likely to get “different” gene trees when the
 239 branch lengths are shorter. Notably, in all three of the species trees, the probability of getting

Figure 3: Distribution of pairwise RF distances among gene trees for the five-taxon species tree: $((((A : T_0, B : T_0) : T_1, C : (T_0 + T_1)) : T_2, D : (T_0 + T_1 + T_2)) : T_3, E : (T_0 + T_1 + T_2 + T_3))$. One of the three internal branch lengths T_1 ((a), (b)), T_2 ((c), (d)), and T_3 ((e), (f)) varies from 0.5 to 5.0, while the other two branch lengths, as well as T_0 , are fixed at 1.0. The varying internal branch length is denoted on the x-axis, and the probabilities of observing the RF distance for two sampled gene trees are denoted on the y-axis with different colors, as shown in the color legend. The left column ((a), (c), (e)) shows the rooted RF distances, while the right column ((b), (d), (f)) shows the unrooted RF distances.

240 two “similar” gene trees that have a rooted RF distance less or equal than 4, or have an
 241 unrooted RF distance less or equal than 2 is smaller than or around 50%. Furthermore, the
 242 exact matching probability of two gene trees is considerably small (below 5% for the rooted
 243 gene trees and around 10% for the unrooted gene trees), which indicates that conflicting
 244 gene trees due to incomplete lineage sorting are common .

Rooted			
RF	0.8	1	1.2
0	0.0180	0.0310	0.0473
2	0.0882	0.1354	0.1847
4	0.2232	0.2868	0.3316
6	0.3193	0.3213	0.2970
8	0.2659	0.1873	0.1229
10	0.0854	0.0383	0.0165
Unrooted			
RF	0.8	1	1.2
0	0.0505	0.0789	0.1111
2	0.1771	0.2408	0.2985
4	0.3547	0.3933	0.4004
6	0.3192	0.2403	0.1686
8	0.0985	0.0467	0.0214

Table 1: Probability distribution of pairwise RF distances between gene trees in a seven-taxon species tree. The branch lengths are scaled by 0.8-, 1.0-, and 1.2-fold (headings of column 2 – 4) to calculate the probability of the RF distance. The headings “Rooted” and “Unrooted” refer to the RF distance between rooted gene trees and unrooted gene trees, respectively. Entries under the title “RF” denote the possible values of the RF distance.

245

DISCUSSION

246 As mentioned in the Introduction, coalescent theory is now extensively used in phylogenetic
 247 species tree inference, and the potential for discordance between gene trees and the species

tree is well-studied (Degnan and Salter, 2005; Degnan and Rosenberg, 2006; Pollard *et al.*, 2006; Liu and Pearl, 2007; Degnan and Rosenberg, 2009; Edwards *et al.*, 2016). However, the importance of considering incongruent gene trees in studies that use phylogenies, but for which phylogeny estimation is not the ultimate goal, is rarely highlighted. In the settings of character mapping and phylogenetic comparative studies, for example, failure to consider gene tree incongruence generated by ILS has the potential to cause systematic biases in the results. To examine this issue in more detail, we quantified the extent of incongruence expected between pairs of gene trees sampled from a fixed species tree in two ways. First, we compute the probability that two sampled gene trees match topologically for species trees with three to eight taxa and varying branch lengths, and second, we compute the probability distribution of RF distances for specific species trees with four, five, and seven taxa. Our major finding is to demonstrate that the probability that two sampled gene trees match can be appreciably low, even when the trees under consideration are not large. Our results highlight that this issue should be of concern in the use of methods based on gene trees, as it cannot be expected, even in studies with limited taxon sampling, that the phylogenies underlying the traits of interest will be topologically congruent.

A high level of gene-tree conflict is widely observed in empirical studies (Spinks and Shaffer, 2009; Betancur-R *et al.*, 2013; Salichos and Rokas, 2013; Pyron *et al.*, 2014), though causes other than ILS are sometimes also hypothesized, such as rate variation among genes, gene flow/hybridization, natural selection or potential estimation error in gene tree inference (Avice *et al.*, 1987; Rand, 2001; Sanderson and Shaffer, 2002; Funk and Omland, 2003; Ballard and Whitlock, 2004; Ballard and Rand, 2005; Spinks and Shaffer, 2009; Simmons *et al.*, 2016). These causes can be difficult to distinguish from the ILS based on the estimated gene trees (Yang, 2002; Leigh *et al.*, 2008; Betancur-R *et al.*, 2013). Our study suggests one way in which this could be assessed, at least for small problems, as follows. First, pairwise RF distances among gene trees in empirical studies can be measured by the method described in Simmons *et al.* (2016). Then, by contrasting the observed pair-

wise RF distances among gene trees with those expected for certain parameter choices, it may be possible to assess whether ILS is the dominant cause of the gene tree incongruence. To aid in this, we provide a set of scripts that can be used with COAL (Degnan and Salter, 2005) and R to compute gene tree probabilities and RF distance distributions (<https://github.com/tian52/2016Match/tree/master/Scripts>). We also provide files that will compute the probability of sampling two matching gene trees from asymmetric species trees with three to six taxa for a set of user-specified branch lengths.

If the observed and expected pairwise RF distances among gene trees are similar, ILS is very likely to be the cause of conflict among observed gene trees. However, if the gene trees are more congruent than the expected, it is possible that the genes cannot be assumed to freely recombine, or that the genes are experiencing selection or convergent evolution (Hahn and Nakhleh, 2015). On the other hand, when the gene trees are less congruent than the expected, rate variation among genes could be important in the phylogenetic inference. Gene flow/hybridization and potential estimation error may affect the gene tree incongruence in either direction, depending on the specific case. Overall, we suggest that researchers should use the expected gene tree incongruence to assess whether ILS is a reasonable explanation for the extent of observed gene tree incongruence before explaining the gene tree incongruence by other causes.

Given the expectation of high levels of gene tree incongruence for even small taxon samples, the question becomes how to adequately deal with such incongruence in procedures for which estimated gene trees provide a necessary framework for inference, such as the phylogenetic comparative method. Clearly, correctly dealing with the possibility of gene tree incongruence is an extremely difficult task, and more research in this area is necessary, as stated in Hahn and Nakhleh (2015). A potential solution in the comparative method framework is to incorporate phylogenetic uncertainty in the underlying trees for all traits under consideration by integrating over all possible gene trees under a particular evolutionary model. However, this approach would incur a substantial computational cost, and it is likely

that more efficient approaches would need to be developed for problems that are large in either the number of traits or the number of taxa (or both).

Despite the complications that arise when dealing with the possibility of gene tree incongruence due to ILS, some of which are highlighted here, we argue that overall gene tree incongruence should be considered as additional phylogenetic information, rather than as “error” or “insufficient data”. When gene tree incongruence arises from ILS alone, it is actually an “intrinsic part of phylogeny’s nature”, as stated by Maddison (1997).

ACKNOWLEDGEMENTS

We thank E. Gluck Thaler, A. Otero Bravo, and P. Blischak for helpful discussion. We thank anonymous reviewers who provided helpful comments on earlier drafts of the manuscript. Both authors were partially supported by NSF DEB-1455399 to L.K. and Dr. Andrea Wolfe.

- 314 Avise, J. C., Reeb, C. A., and Saunders, N. C. 1987. Geographic population structure and
315 species differences in mitochondrial DNA of mouthbrooding marine catfishes (Ariidae) and
316 demersal spawning toadfishes (Batrachoididae). *Evolution*, pages 991–1002.
- 317 Ballard, J. W. O. and Rand, D. M. 2005. The population biology of mitochondrial DNA
318 and its phylogenetic implications. *Annual Review of Ecology, Evolution, and Systematics*,
319 pages 621–642.
- 320 Ballard, J. W. O. and Whitlock, M. C. 2004. The incomplete natural history of mitochondria.
321 *Molecular ecology*, 13(4): 729–744.
- 322 Bayzid, M. S. and Warnow, T. 2012. Estimating optimal species trees from incomplete gene
323 trees under deep coalescence. *Journal of Computational Biology*, 19(6): 591–605.
- 324 Bell, N. E. and Hyvönen, J. 2010. Phylogeny of the moss class Polytrichopsida (Bryophyta):
325 generic-level structure and incongruent gene trees. *Molecular phylogenetics and evolution*,
326 55(2): 381–398.
- 327 Betancur-R, R., Broughton, R. E., Wiley, E. O., Carpenter, K., López, J. A., Li, C., Holcroft,
328 N. I., Arcila, D., Sanciangco, M., Cureton, J., *et al.* 2013. The tree of life and a new
329 classification of bony fishes. *PLoS Curr*, 5: 1–33.
- 330 Cranston, K. A., Hurwitz, B., Ware, D., Stein, L., and Wing, R. A. 2009. Species trees from
331 highly incongruent gene trees in rice. *Systematic Biology*, page syp054.
- 332 Degnan, J. and Salter, L. 2005. Gene tree distributions under the coalescent process. *Evo-*
333 *lution*, 59: 24–37.
- 334 Degnan, J. H. and Rosenberg, N. A. 2006. Discordance of species trees with their most likely
335 gene trees. *PLoS Genetics*, 3: 762–768.

336 Degnan, J. H. and Rosenberg, N. A. 2009. Gene tree discordance, phylogenetic inference
337 and the multispecies coalescent. *Trends Ecol. Evol.*, 24(6): 332–340.

338 Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong,
339 B., Wu, S., Lemmon, E. M., Lemmon, A. R., *et al.* 2016. Implementing and testing
340 the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular*
341 *Phylogenetics and Evolution*, 94: 447–462.

342 Fan, H. H. and Kubatko, L. S. 2011. Estimating species trees using approximate Bayesian
343 computation. *Molecular Phylogenetics and Evolution*, 59(2): 354–363.

344 Ferris, S. D., Sage, R. D., Huang, C.-M., Nielsen, J. T., Ritte, U., and Wilson, A. C.
345 1983. Flow of mitochondrial DNA across a species boundary. *Proceedings of the National*
346 *Academy of Sciences*, 80(8): 2290–2294.

347 Funk, D. J. and Omland, K. E. 2003. Species-level paraphyly and polyphyly: frequency,
348 causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review*
349 *of Ecology, Evolution, and Systematics*, pages 397–423.

350 Good, J. M., Hird, S., Reid, N., Demboski, J. R., Stepan, S. J., MARTIN-NIMS, T. R., and
351 Sullivan, J. 2008. Ancient hybridization and mitochondrial capture between two species
352 of chipmunks. *Molecular Ecology*, 17(5): 1313–1327.

353 Hahn, M. W. and Nakhleh, L. 2015. Irrational exuberance for resolved species trees. *Evolu-*
354 *tion*.

355 Hein, J. 1993. A heuristic method to reconstruct the history of sequences subject to recom-
356 bination. *J. Mol. Evol.*, 36: 396–405.

357 Huelsenbeck, J. P. and Rannala, B. 2003. Detecting correlation between characters in a
358 comparative analysis with uncertain phylogeny. *Evolution*, 57(6): 1237–1247.

- 359 Huelsenbeck, J. P., Rannala, B., and Masly, J. P. 2000. Accommodating phylogenetic un-
360 certainty in evolutionary studies. *Science*, 288(5475): 2349–2350.
- 361 Kingman, J. F. C. 1982a. On the genealogy of large populations. *J. Appl. Prob.*, 19A: 27–43.
- 362 Kingman, J. F. C. 1982b. The coalescent. *Stoch. Proc. Appl.*, 13: 235–248.
- 363 Kubatko, L. S. 2009. Identifying hybridization events in the presence of coalescence via
364 model selection. *Syst. Biol.*, 58(5): 478–488.
- 365 Leaché, A. D. and McGuire, J. A. 2006. Phylogenetic relationships of horned lizards (Phryno-
366 soma) based on nuclear and mitochondrial data: Evidence for a misleading mitochondrial
367 gene tree. *Molecular phylogenetics and evolution*, 39(3): 628–644.
- 368 Leigh, J. W., Susko, E., Baumgartner, M., and Roger, A. J. 2008. Testing congruence in
369 phylogenomic analysis. *Systematic Biology*, 57(1): 104–115.
- 370 Liu, L. and Pearl, D. 2007. Species trees from gene trees: reconstructing Bayesian posterior
371 distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.*,
372 56: 504–514.
- 373 Liu, L., Yu, L., and Edwards, S. V. 2010. A maximum pseudo-likelihood approach for
374 estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(1):
375 302.
- 376 Lutzoni, F., Pagel, M., and Reeb, V. 2001. Major fungal lineages are derived from lichen
377 symbiotic ancestors. *Nature*, 411(6840): 937–940.
- 378 Maddison, W. P. 1997. Gene trees in species trees. *Syst. Biol.*, 46: 523–536.
- 379 McCracken, K. G. and Sorenson, M. D. 2005. Is homoplasy or lineage sorting the source of
380 incongruent mtDNA and nuclear gene trees in the stiff-tailed ducks (Nomonyx-Oxyura)?
381 *Systematic Biology*, 54(1): 35–55.

382 Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., Frandsen, P. B.,
383 Ware, J., Flouri, T., Beutel, R. G., *et al.* 2014. Phylogenomics resolves the timing and
384 pattern of insect evolution. *Science*, 346(6210): 763–767.

385 Moore, W. S. 1995. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees
386 versus nuclear-gene trees. *Evolution*, pages 718–726.

387 Moyer, G., Remington, R., and Turner, T. 2009. Incongruent gene trees, complex evolution-
388 ary processes, and the phylogeny of a group of North American minnows (*hybognathus*
389 *agassiz* 1855). *Molecular phylogenetics and evolution*, 50(3): 514–525.

390 Pagel, M. and Lutzoni, F. 2002. Accounting for phylogenetic uncertainty in comparative
391 studies of evolution and adaptation. In *Biological evolution and statistical physics*, pages
392 148–161. Springer.

393 Pagel, M., Meade, A., and Barker, D. 2004. Bayesian estimation of ancestral character states
394 on phylogenies. *Systematic biology*, 53(5): 673–684.

395 Pamilo, P. and Nei, M. 1988. Relationships between gene trees and species trees. *Molecular*
396 *Biology and Evolution*, 5(5): 568–583.

397 Peters, J. L., Zhuravlev, Y., Fefelov, I., Logie, A., and Omland, K. E. 2007. Nuclear loci
398 and coalescent methods support ancient hybridization as cause of mitochondrial paraphyly
399 between gadwall and falcated duck (*Anas* spp.). *Evolution*, 61(8): 1992–2006.

400 Pollard, D. A., Iyer, V. N., Moses, A. M., and Eisen, M. B. 2006. Widespread discordance of
401 gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS*
402 *Genet*, 2(10): e173.

403 Pyron, R. A., Hendry, C. R., Chou, V. M., Lemmon, E. M., Lemmon, A. R., and Burbrink,
404 F. T. 2014. Effectiveness of phylogenomic data and coalescent species-tree methods for

- resolving difficult nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia).
Molecular phylogenetics and evolution, 81: 221–231.
- R Core Team 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rand, D. M. 2001. The units of selection of mitochondrial dna. *Annual Review of Ecology and Systematics*, pages 415–448.
- Rannala, B. and Yang, Z. 2003. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 164: 1645–1656.
- Richman, A. D. and Price, T. 1992. Evolution of ecological differences in the Old world leaf warblers. *Nature*, 355(6363): 817–821.
- Robertson, C. L., Soane, L., Siegel, Z. T., and Fiskum, G. 2006. The potential role of mitochondria in pediatric traumatic brain injury. *Developmental neuroscience*, 28(4-5): 432–446.
- Robinson, D. F. and Foulds, L. R. 1979. Comparison of weighted labelled trees. In *Combinatorial mathematics VI*, pages 119–126. Springer.
- Rokas, A., Williams, B. L., King, N., and Carroll, S. B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960): 798–804.
- Roos, C., Zinner, D., Kubatko, L. S., Schwarz, C., Yang, M., Meyer, D., Nash, S. D., Xing, J., Batzer, M. A., Brameier, M., Leendertz, F. H., Ziegler, T., Perwitasari-Farajallah, D., Nadler, T., Walter, L., and Osterholz, M. 2011. Nuclear versus mitochondrial DNA: Evidence for hybridization in colobine monkeys. *BMC Evolutionary Biology*, 11: 77.
- Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. *Theor. Popul. Biol.*, 61: 225–247.

- 428 Salichos, L. and Rokas, A. 2013. Inferring ancient divergences requires genes with strong
429 phylogenetic signals. *Nature*, 497(7449): 327–331.
- 430 Sanderson, M. J. and Shaffer, H. B. 2002. Troubleshooting molecular phylogenetic analyses.
431 *Annual review of ecology and Systematics*, pages 49–72.
- 432 Sang, T. and Zhong, Y. 2001. Testing hybridization hypotheses based on incongruent gene
433 trees. *Syst. Biol.*, 49: 422–434.
- 434 Shaw, K. L. 2002. Conflict between nuclear and mitochondrial DNA phylogenies of a re-
435 cent species radiation: What mtDNA reveals and conceals about modes of speciation in
436 Hawaiian crickets. *Proc. Nat. Acad. Sci. USA*, 99: 16122–16127.
- 437 Simmons, M. P., Sloan, D. B., and Gatesy, J. 2016. The effects of subsampling gene trees on
438 coalescent methods applied to ancient divergences. *Molecular phylogenetics and evolution*,
439 97: 76–89.
- 440 Sota, T. and Vogler, A. P. 2001. Incongruence of mitochondrial and nuclear gene trees in
441 the carabid beetles Ohomopterus. *Systematic biology*, 50(1): 39–59.
- 442 Spinks, P. Q. and Shaffer, H. B. 2009. Conflicting mitochondrial and nuclear phylogenies for
443 the widely disjunct *Emys* (Testudines: Emydidae) species complex, and what they tell us
444 about biogeography and hybridization. *Systematic biology*, 58(1): 1–20.
- 445 Steel, M. A. and Penny, D. 1993. Distributions of tree comparison metrics some new results.
446 *Systematic Biology*, 42(2): 126–141.
- 447 Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics*,
448 105: 437–460.
- 449 Takahata, N. 1989. Gene genealogy in three related populations: consistency probability
450 between gene and population trees. *Genetics*, 122: 957–966.

- 451 Takahata, N. and Nei, M. 1985. Gene genealogy and variance of interpopulational nucleotide
452 differences. *Genetics*, 110: 325–344.
- 453 Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in popu-
454 lation genetics models. *Theor. Popul. Biol.*, 26: 119–164.
- 455 Wiegmann, B. M., Trautwein, M. D., Kim, J.-W., Cassel, B. K., Bertone, M. A., Winterton,
456 S. L., and Yeates, D. K. 2009. Single-copy nuclear genes resolve the phylogeny of the
457 holometabolous insects. *BMC biology*, 7(1): 1.
- 458 Yang, Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids
459 using data from multiple loci. *Genetics*, 162: 1811–1823.