Syst. Biol. 58(5):527–536, 2009
© The Author(s) 2009. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oxfordjournals.org
DOI:10.1093/sysbio/syp047
Advance Access publication on August 26, 2009

What Is the Danger of the Anomaly Zone for Empirical Phylogenetics?

HUATENG HUANG AND L. LACEY KNOWLES*

Department of Ecology and Evolutionary Biology, Museum of Zoology, University of Michigan, 1109 Geddes Avenue, Ann Arbor, MI 48109-1079, USA; E-mail: huatengh@umich.edu;
**Correspondence to be sent to: Department of Ecology and Evolutionary Biology, Museum of Zoology, University of Michigan, 1109 Geddes Avenue, Ann Arbor, MI 48109-1079, USA; E-mail: knowlesl@umich.edu.*

Abstract.—The increasing number of observations of gene trees with discordant topologies in phylogenetic studies has raised awareness about the problems of incongruence between species trees and gene trees. Moreover, theoretical treatments focusing on the impact of coalescent variance on phylogenetic study have also identified situations where the most probable gene trees are ones that do not match the underlying species tree (i.e., anomalous gene trees [AGTs]). However, although the theoretical proof of the existence of AGTs is alarming, the actual risk that AGTs pose to empirical phylogenetic study is far from clear. Establishing the conditions (i.e., the branch lengths in a species tree) for which AGTs are possible does not address the critical issue of how prevalent they might be. Furthermore, theoretical characterization of the species trees for which AGTs may pose a problem (i.e., the anomaly zone or the species histories for which AGTs are theoretically possible) is based on consideration of just one source of variance that contributes to species tree and gene tree discord—gene lineage coalescence. Yet, empirical data contain another important stochastic component—mutational variance. Estimated gene trees will differ from the underlying gene trees (i.e., the actual genealogy) because of the random process of mutation. Here, we take a simulation approach to investigate the prevalence of AGTs, among estimated gene trees, thereby characterizing the boundaries of the anomaly zone taking into account both coalescent and mutational variances. We also determine the frequency of realized AGTs, which is critical to putting the theoretical work on AGTs into a realistic biological context. Two salient results emerge from this investigation. First, our results show that mutational variance can indeed expand the parameter space (i.e., the relative branch lengths in a species tree) where AGTs might be observed in empirical data. By exploring the underlying cause for the expanded anomaly zone, we identify aspects of empirical data relevant to avoiding the problems that AGTs pose for species tree inference from multilocus data. Second, for the empirical species histories where AGTs are possible, unresolved trees—not AGTs—predominate the pool of estimated gene trees. This result suggests that the risk of AGTs, while they exist in theory, may rarely be realized in practice. By considering the biological realities of both mutational and coalescent variances, the study has refined, and redefined, what the actual challenges are for empirical phylogenetic study of recently diverged taxa that have speciated rapidly—AGTs themselves are unlikely to pose a significant danger to empirical phylogenetic study. [Anomaly zone; coalescence; gene tree; lineage sorting; mutation; phylogenetics; species tree.]

Incongruence between gene trees and species trees has long been acknowledged as a serious challenge for phylogenetic studies (Pamilo and Nei 1988; Takahata 1989; Maddison 1997). However, it is only recently that the potential magnitude of the problem has become apparent. Discordant gene trees are routinely encountered in multilocus studies (e.g., Jennings and Edwards 2005; Carstens and Knowles 2007; Knowles and Carstens 2007; Wong et al. 2007; Carling and Brumfield 2008; Sanderson 2008). Moreover, recent theoretical work has also identified a very ominous situation in which the most probable gene trees do not match the underlying species tree—anomalous gene trees [AGTs] (Degnan and Rosenberg 2006; Rosenberg and Tao 2008). Under such species histories, the gene trees can lead to incorrect conclusions about the history of species divergence with current methodologies. Moreover, within the anomaly zone (i.e., species histories for which AGTs are theoretically possible) increased sampling of loci, by itself, will not increase the accuracy of phylogenetic inference because the most frequent gene trees will provide positively misleading information about species relationships.

Although the theoretical proof of the existence of AGTs is alarming, the actual risk that AGTs pose to empirical phylogenetic study is far from clear. First, establishing the conditions (i.e., the branch lengths in a species tree) for which AGTs are possible (Degnan and Rosenberg 2006; Rosenberg and Tao 2008) does not address the critical issue of how prevalent they might be. For example, if AGTs are possible, but not probable, then even for those species histories where AGTs can theoretically occur (i.e., species trees within the anomaly zone), they may not represent a significant danger. On the other hand, if the frequency of AGTs for a given history of species divergence is high, then they may very well result in misleading phylogenetic inferences. Second, theoretical characterization of the species trees for which AGTs may pose a problem is based on consideration of just one source of variance that contributes to species tree and gene tree discordance—gene lineage coalescence. Yet, empirical data contain another inherent stochastic component—mutational variance. Estimated gene trees will differ from the underlying gene tree produced by the coalescent process because of the random process of mutation (Fig. 1). The impact of this mutational variance on the zone (i.e., species tree branch lengths) for which AGTs will be realized in empirical data remains to be investigated; therefore, unlike previous theoretical investigations (Degnan and Rosenberg 2006; Rosenberg and Tao 2008), our study focuses on estimated gene trees that are AGTs—that is, the most frequent estimated gene tree does not match the species tree.

Here, we take a simulation approach to investigate the prevalence of AGTs to determine how significant

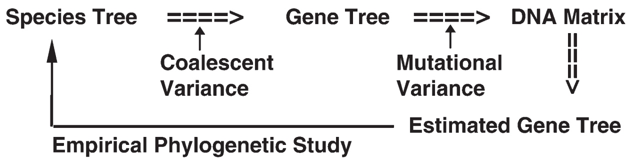

FIGURE 1. The schematic shows the concepts and terminology relevant to investigating the genetic sources of variance in an empirical phylogenetic study. A species tree represents the actual history of divergence among species. A gene tree in this context refers to the actual evolutionary history of orthologous genes across species, where the coalescent process introduces a variance in the gene trees across loci that evolve within the branches of different species. Estimated gene tree refers to the evolutionary history estimated from DNA sequences; because of the mutational process, and errors with estimating the gene tree, the topology of the estimated gene tree may not match the underlying gene tree (i.e., the coalescent gene tree) for a locus. In contrast to previous theoretical treatments of the anomaly zone (i.e., Degnan and Rosenberg 2006; Rosenberg and Tao 2008), this study considers both coalescent and mutational sources of variances for empirical phylogenetic study, focusing on AGTs for estimated gene trees rather than coalescent gene trees.

a threat they actually represent for empirical phylogenetic investigation. We focus on a 4-taxon species tree with an asymmetric topology (Fig. 2), which is the simplest tree that can produce AGTs. Moreover, the relative branch lengths of the species tree defining the boundary of the anomaly zone have also been solved analytically (Degnan and Rosenberg 2006). In this study, we focus on both (1) the frequency of AGTs across the zone where AGTs are possible (i.e., the prevalence of AGTs for species trees with differing branch lengths) and (2) the impact of mutational variance on the boundary of the anomaly zone (i.e., what are the relative branch lengths in a species tree where AGTs are possible, and does this differ depending on the mutation rate).

Unlike the coalescent variance (Kingman 1982), the complex properties of mutational variance make investigating its effect on species tree estimation difficult. Coalescent variance, under assumptions of a Fisher–Wright population, can be expressed via analytical equations (Takahata and Nei 1985; Pamilo and Nei 1988). That is, given a species history, the frequency spectrum of different gene tree topologies can be calculated (Degnan and Salter 2005). In contrast, no such treatment of mutational variance exists, which no doubt reflects the difficulties associated with characterizing the multifarious effects of mutation. The difference between a single estimated gene tree (i.e., topology and branch lengths) from its underlying gene tree (e.g., Saitou and Nei 1987) cannot simply be ascribed to the nucleotide substitution process (i.e., the model of molecular evolution, which includes parameters such as the transition–transversion ratio, frequencies of nucleotides,

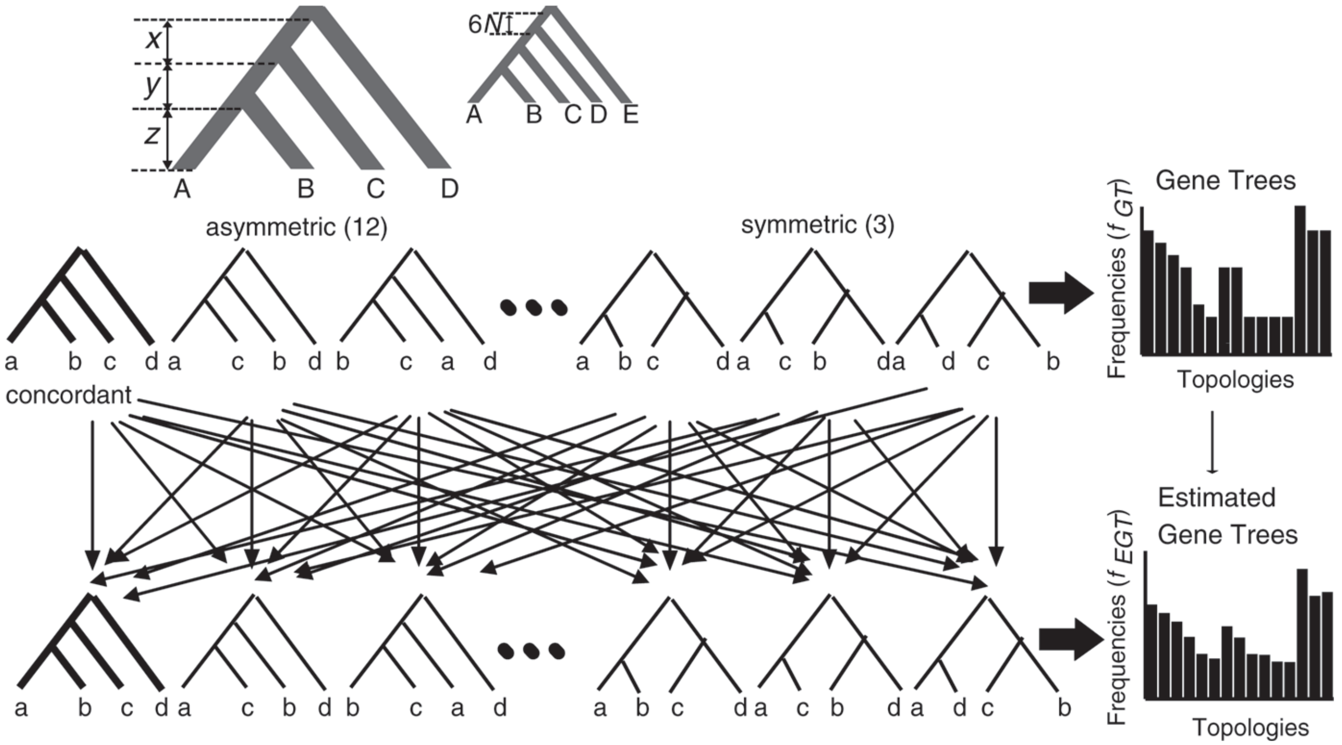

FIGURE 2. For the 4-taxon (Species A, B, C, and D), asymmetric species tree used in the study, there are 15 possible tree topologies: 12 asymmetric trees (which are not all shown, as indicated by the dotted lines) and 3 symmetric trees; an out-group, Species E, was used as the root. Variation in the species tree branch lengths x and y was explored to characterize the anomaly zone for estimated gene trees (i.e., species tree where the most frequent estimated gene tree topology does not match the species tree). The frequency distribution of gene tree topologies (f_{GT}) may differ from the frequency distribution of estimated gene tree topologies (f_{EGT}), as illustrated with the 2 histograms, because of the mutation process (i.e., the estimated gene tree topology may differ from the actual underlying gene tree topology). Hence, the boundaries of the anomaly zone characterized with coalescent gene trees may differ when estimated gene trees are analyzed.

and heterogeneity in mutation rate across sites; Ripplinger and Sullivan 2008). Differences between the actual genealogical history of a locus and the estimated gene tree (Fig. 1) can also arise from the criteria used for evaluating trees and tree space-searching algorithms (e.g., Zhou et al. 2007). Lastly, when considered in the context of multiple independent loci, because estimated gene trees may differ from their underlying gene trees, mutation can also cause a deviation from the expected frequency spectrum of topologies for any given species tree (Degnan and Salter 2005; Kubatko and Degnan 2007). Given that AGTs are defined by the frequency spectrum of gene tree topologies (i.e., a gene tree is called anomalous whenever it has a higher probability under the coalescent model than the gene tree with a topology that matches the species tree), any mutational-induced shift in the frequency spectrum of topologies needs to be examined. In terms of evaluating the threat that AGTs pose for empirical phylogenetic study, shifts in the frequency spectrum of particular interest would be those that could produce an expanded zone (i.e., species tree branch lengths) where AGTs are possible (i.e., a greater range of species histories might be subject to AGTs).

Our results show that mutational variance can indeed expand the parameter space (i.e., the relative branch lengths in a species tree) where AGTs are possible. The cause of the expansion is discussed along with detailed analysis of several divergence scenarios. Yet, when we examine the frequency of AGTs among the estimated gene trees for those species histories where AGTs are possible, we show that they are improbable for biologically realistic values of mutation rate (i.e., $\theta = 0.01$–0.001; Drost and Lee 1995). For these conditions, a polytomy (i.e., an unresolved internode) is more likely than an AGT. Therefore, although theoretically possible, there is insufficient mutation for AGTs to be realized in practice. As discussed, the minimum species tree branch length wherein estimated gene trees faithfully reflect the topology of the underlying gene tree actually exceeds the zone where AGTs are theoretically possible, meaning that AGTs themselves are unlikely to pose a significant danger to empirical phylogenetic study.

Methods

General Simulation Procedure

When considering coalescent variance, 2 internal branch lengths on a 4-taxon, asymmetric bifurcating species tree with no migration after speciation (Species A, B, C, and D), x and y (see Fig. 2), determine whether there are AGTs, as well as the number of AGTs (i.e., the number of topologies that occur with higher frequency than gene trees matching the species tree topology; see Degnan and Rosenberg 2006). To characterize the impact of mutational variance on the prevalence of AGTs for estimated gene tree, genealogies for diploid loci were simulated using the program MS (Hudson 2002) under a neutral coalescent model, with 1 individual per species, for different species trees with specific x and y branch lengths. For each gene tree, 1 set of DNA sequences with a length of 1000 bp was simulated with the program Seq-gen (Rambaut and Grassly 1997), using a HKY85 mutation model (discrete gamma distribution with a shape parameter of 0.8 and 4 categories, transition–transversion ratio of 0.3 with nucleotide probabilities set to 0.3 A, 0.2 C, 0.2 G, and 0.3 T). A gene tree was estimated for each set of sequences with PAUP* version 4.0b10 (Swofford 2000), and 1 gene tree was selected by the maximum likelihood (ML) criteria from an exhaustive search. The frequency of the 16 tree topologies (i.e., 12 asymmetric topologies, 3 symmetric topologies, and any estimated gene tree with a polytomy, which was considered as 1 topological category) was calculated from the replicated data sets per species tree (see below for details about the number of replicates used).

An out-group species (E) that diverged $6N$ generations prior to the common ancestor of Species A, B, C, and D (where $N =$ effective population size) was used to root the estimated gene trees (Fig. 2). With this branch length, Species A, B, C, and D formed a monophyletic group in most of the gene trees (i.e., in about 95% of the replicate data sets). In those few cases where a deep coalescence occurred between Species A, B, C, and D and the out-group, Species E, the gene tree was excluded from further analysis to avoid additional coalescent variance being mistaken as mutational variance (i.e., the estimated gene tree differed from its underlying gene tree because Species E was not an out-group because of coalescent stochasticity); this does not affect the conclusions about the prevalence of AGTs (i.e., the anomaly zone for the simulated data and based on the coalescent variance is comparable to that defined by Degnan and Rosenberg 2006).

Characterization of the Frequency Spectrum of Estimated Gene Trees for Different Species Trees

Only symmetric gene tree topologies can have a higher probability than the topology that is concordant with a 4-taxon species tree (Degnan and Rosenberg 2006). The frequency of each of the 3 topologies (((a, b), (c, d)), ((a, c), (b, d)), and ((a, d), (b, c))), relative to the concordant topology (((a, b), c), d), was tallied for each species tree, under 2 simulation strategies designed to characterize the realized anomaly zone with mutational variance (i.e., the relative branch lengths in a species tree where AGTs are possible for estimated gene trees, specifically the branches x and y in the species tree; Fig. 2).

In principle, the effect of mutational variance on the anomaly zones could be examined by simulating replicate data sets for each branch length combination of x and y in the species tree. However, this approach has 2 potential problems. The number of replicate data sets needed for accurately assessing the existence of AGTs is not clear. For values of x and y near the boundaries of the anomaly zone, the probabilities of the different estimated gene tree topologies are so similar that the number of AGTs (i.e., 0 or >0) may not be inferred accurately with a limited number of simulations. To combat this problem, we used a method of incrementally increasing the number of simulated data sets for each set of x and y branch lengths. For each species tree, the number of AGTs was assessed for every additional set of 1000 replicate data sets. When there was no change in the number of observed AGTs with an additional 2 sets of replicates (in increments of 1000), no additional data were simulated. In other words, the number of AGTs calculated was invariant across a minimum range of 3000 estimated gene trees, and therefore it is unlikely that the estimated frequency of AGTs would change with additional data. To deal with the second problem of computational inefficiency in finding the boundaries of the anomaly zone, we used a bisection approach. For a given x, the y value starts from $0.01N$ and was continually updated to $y'(=y+1N)$ until reaching a point where (x, y') have 0 AGTs, thereby defining a parameter space where the boundary of the anomaly zone was crossed. The next set of simulations were conducted based on the average value $y''\left(=\frac{y+y'}{2}\right)$, and depending on whether the number of AGTs at (x, y'') is bigger than 0, the next set of branch lengths explored in the parameter space was $0.25N$ distance either above or below y''. This bisecting step was repeated until the resolution (i.e., minimal distance between parameter values) was $\pm 0.015625N$ (or ±1562 years for a species with a population size of 10^5 and 1 generation per year), which is a very small increment for phylogenetic study. This fine scale mapping of the boundary zone was conducted across a range of x from 0.11 to $1.91N$ and 0.01 to $11N$ for y (these values span almost the entire anomaly zone, see Results for details).

This procedure was carried out on data sets simulated under 3 different mutation rates: θ ($4N\mu$, μ = mutation rate per site) of 0.005, 0.01, and 0.05, which not only span the range observed in empirical data (i.e., $\theta = 0.01$–0.001) but also include an artificially inflated mutation rate (i.e., $\theta = 0.05$). All these simulations were also explored under 2 differing external branch lengths of the species tree (i.e., z in Fig. 2): $z = 5N$ and $15N$. Although the external branch lengths have no effect on the frequency spectrum of gene tree resulting from the coalescent process (i.e., f_{GT} in Fig. 2), z might have an effect on the frequency spectrum of estimated gene trees (i.e., f_{EGT} in Fig. 2) through its effect on the absolute genetic distances between taxa.

Exploring the Cause for an Expansion of the Anomaly Zone

Although the expansion of the anomaly zone with gene trees estimated from the simulated nucleotide data sets indicates that mutational variance has induced a shift in the frequency distribution of tree topologies (Fig. 2), the cause for this shift is not intuitively obvious. Two different factors could contribute to this shift: (1) the proportion of estimated gene trees that match the topologies of the underlying gene trees, herein referred to as the proportion of correct reconstruction (P_{C}), may differ across the 15 possible topologies and (2) the estimated gene trees that do not match the underlying gene trees, herein referred to as the proportion of misidentification (P_{M}), may not be equally distributed among the 15 possible topologies. To investigate the relative contributions of these 2 factors in causing a deviation from the expected frequency distribution of tree topologies, P_{C} and P_{P} were calculated for each of the 15 possible topologies. For example, for a gene tree with the ith topology, P^i_{C} is the proportion of replicate data sets with an estimated gene tree with the ith topology, whereas P^i_{M} is the proportion of estimated gene trees with the ith topology representing misidentified topologies. Therefore, the frequency of the ith topology among the estimated gene trees (f^i_{EGT}) can be calculated as

$$f^i_{\mathrm{EGT}} = f^i_{\mathrm{GT}} \times P^i_{\mathrm{C}} + \left\{ \sum_{i=1}^{15} \left[f^i_{\mathrm{GT}} \times (1 - P^i_{\mathrm{C}}) \right] \right\} \times P^i_{\mathrm{M}},$$

where f^i_{GT} denotes the frequency of the ith topology in gene trees.

These analyses are performed on 8 species trees located along the boundaries of the anomaly zone, namely for species trees with an x branch length of 0.05, 0.10, 0.15, and 0.20, and the corresponding y values, as calculated according to equations (4) and (5) in Degnan and Rosenberg (2006), where z was set to $5N$. On average, 4729 replicate estimated gene trees were examined for each species tree from simulated data with $\theta = 0.01$; the number of replicates for each species tree differed slightly because only those gene trees for which Species A, B, C, and D formed a monophyletic group relative to the out-group, Species E, were considered from the 5000 simulated data sets.

Results

Impact of Mutational Variance on the Size of the Anomaly Zone

Compared with species trees with AGTs when just the coalescent variance is considered, there is an obvious expansion of the anomaly zone based on analysis of estimated gene trees, which also incorporate mutational variance (Fig. 3). The expansion is apparent at all mutation rates and for different lengths of the external species tree branch, z, although the magnitude of the effect differs. It is worth noting that the degree of expansion shown here should also be considered conservative given that the mutational model of evolution was known, whereas the substitution model for an empirical study would have to be estimated.

Cause for the Expansion of the Anomaly Zone

Expansion of the species tree branch lengths defining the boundary of the anomaly zone (Fig. 3) reflects an increased number of estimated gene trees with anomalous topologies caused by mutational variance. This

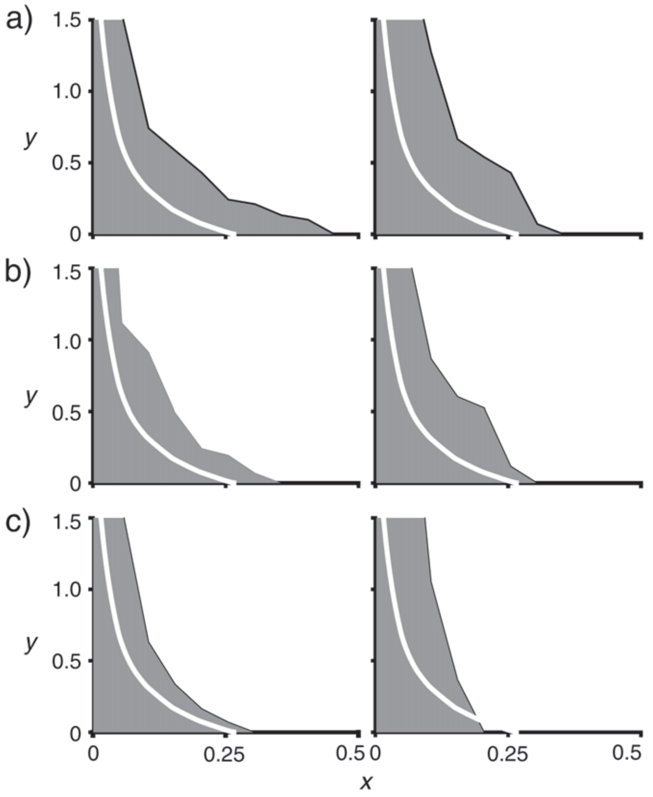

FIGURE 3. Distribution of species tree branch lengths (i.e., x and y, in units of $2N$) that define the boundary of the anomaly zone, which shows an expansion of species trees characterized by AGTs when mutational variance is considered (shaded area). The anomaly zone below the white line delimits the area with AGTs based on consideration of just the stochasticity of the coalescent (i.e., characterization of the anomaly zone based on coalescent gene trees rather than estimated gene trees, Fig. 1). Results from the different simulation conditions are shown, with plots on the left versus right reflecting species trees with $z = 5N$ and $z = 15N$, respectively, under 3 different mutation rates a) $\theta = 0.005$, b) $\theta = 0.01$, and c) $\theta = 0.05$.

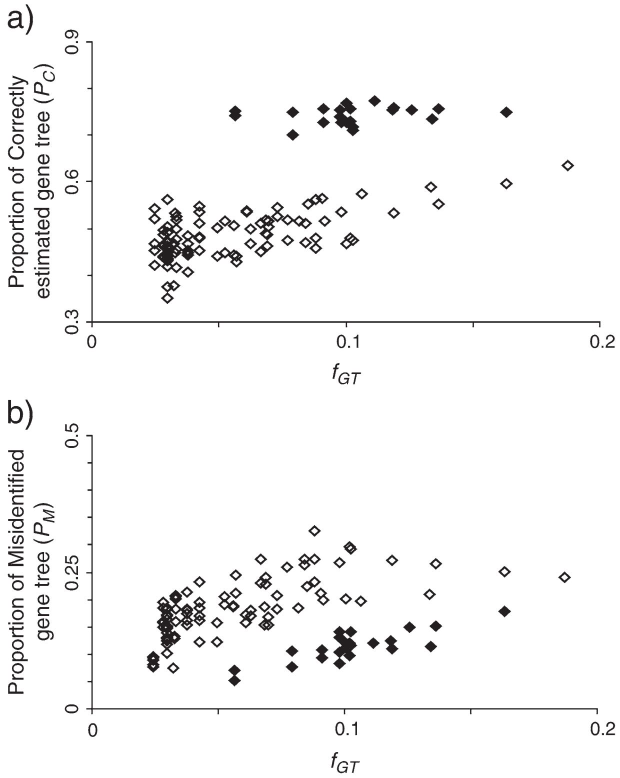

FIGURE 4. Comparison of the proportion of a) correctly estimated (P_{C}) and b) misidentified (P_{M}) gene trees for symmetric (shown in solid diamonds) and asymmetric topologies (shown as open diamonds) for species trees along the anomaly zone boundary; the frequency of the respective gene tree topologies (f_{GT}) is shown along the x-axis.

shift in the expected frequency distribution of topologies appears to be caused by a concomitant increase in the frequencies of estimated gene trees with symmetric topologies ($f_{\mathrm{GT}} < f_{\mathrm{EGT}}$) and a decrease in the frequency of the concordant tree topology ($f_{\mathrm{GT}} > f_{\mathrm{EGT}}$). The proportion of gene trees that are correctly reconstructed (P_{C}) reveals 2 distinct groups (Fig. 4a), with asymmetric gene trees showing a significantly lower frequency of correct estimation compared with symmetric gene trees. This pattern identifies one mechanism underlying the shift in the anomaly zone—a deficit of correctly estimated asymmetric gene trees, relative to symmetric ones. When the estimated gene trees do not match the underlying gene tree, examination of the proportion of the misidentified gene trees (P_{M}) shows that both asymmetric and symmetric topologies are represented in similar proportions among the misidentified gene trees (Fig. 4b). Although there is a slightly lower average P_{M} for symmetric gene trees, which is consistent with the inherent bias associated with phylogenetic estimation procedures (e.g., Huelsenbeck and Kirkpatrick 1996), the relatively small effect suggests a minimal contribution to the shift in the frequency distribution of tree topologies. This suggests that the expanding anomaly zone is not caused by an inflation of the frequency of symmetric trees from misidentified gene trees.

Prevalence of AGTs based on Estimated Gene Trees

The expansion of the anomaly zone (Fig. 3) when estimated gene trees are analyzed, relative to their underlying gene trees (Fig. 1), suggests that AGTs might represent a bigger problem for empirical phylogenetic studies than initially identified (Degnan and Rosenberg 2006). In fact, the increase in the anomaly zone caused by mutational variance is comparable to or greater than the amount of expansion observed with increasing the number of taxa (Rosenberg and Tao 2008). However, with estimated gene trees, there is another important class of topologies that does not exist with coalescent gene trees—estimated gene trees with polytomies (i.e., estimated gene trees with unresolved branches, as identified by an ML tree with zero branch length). This polytomy zone actually predominates species trees with very short internodes (Fig. 5). For this region

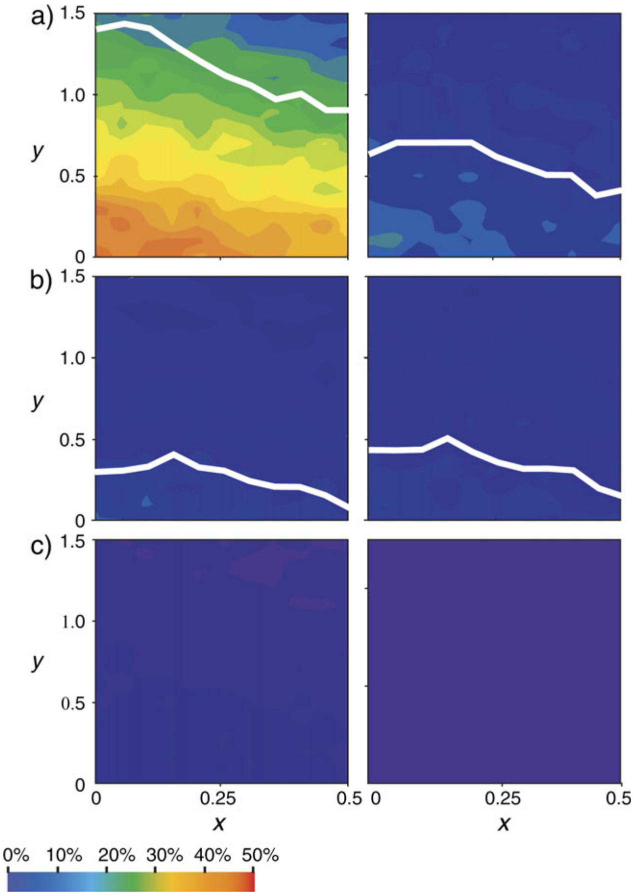

FIGURE 5. Frequencies of estimated gene trees with a polytomy for the different species trees (i.e., differing x and y branch lengths, in units of $2N$) for different simulation conditions: specifically, for 3 different mutation rates a) $\theta = 0.005$, b) $\theta = 0.01$, and c) $\theta = 0.05$, and the branch length of $z = 5N$ (shown on the left) and $z = 15N$ (shown on the right). The white lines demark the boundary of the polytomy zone—for all species trees under the white line, a polytomy is the most frequent topology of the estimated gene trees.

of species tree parameter space, the polytomy zone overlaps broadly with most of the region where AGTs are possible (Fig. 3). These analyses indicate that the most probable topology is a polytomy (Fig. 5a,b), not an AGT; AGTs are only realized for a confined parameter space (i.e., species trees with very short x branch lengths and a y greater than 0.5), with $\theta = 0.01$ (Fig. 5b). As the branch lengths (i.e., x and y) in the species tree increase, the polytomy zone is replaced by a region where estimated gene trees are resolved (i.e., area above the white line; Fig. 5a,b). However, at these species tree branch lengths (i.e., areas where resolved estimated gene trees, not polytomies, predominate), the most frequent topology observed in the estimated gene tree is likely to be the one that is concordant with the species tree (i.e., the species tree branch lengths fall outside the anomaly zone—above the grey area in Fig. 3). Only with artificially high mutation rates (i.e., $\theta = 0.05$; Fig. 5c), does the frequency of AGTs exceed the frequency of polytomies. However, loci with this high of a mutation rate, or nonrecombining DNA fragments greater than the 1000 bp used here, generally are rarely seen in phylogeny studies.

DISCUSSION

Given a sample of estimated gene trees from multiple independent loci, one might intuit that the actual species tree could be accurately identified using a democratic consensus procedure (e.g., Jennings and Edwards 2005). However, the discovery of AGTs (Degnan and Rosenberg 2006) indicated that even with unlimited data, the democratic consensus would not identify the correct species tree, which is alarming for phylogenetic studies. Moreover, recent study on a 5-taxon species tree (Rosenberg and Tao 2008) revealed that the anomaly zone expanded with the addition of taxa, again signaling an inherent danger for estimating species relationships in groups that have radiated recently (i.e., short internal branch lengths in the species tree). Not withstanding the virtues of these theoretical studies, for empirical studies the question is how frequently will AGTs be represented among a set of estimated gene trees. Our analysis of the anomaly zone based on estimated gene trees (Fig. 1), in contrast to previous treatments based on coalescent gene trees (Degnan and Rosenberg 2006; Rosenberg and Tao 2008), provides this much needed context. Two salient results emerge from this investigation into the effects of mutational variance on the anomaly zone. First, the documented expansion of the anomaly zone with estimated gene trees (as opposed to gene trees), and its underlying cause, identifies aspects of empirical data relevant to avoiding the problems that AGTs pose for species tree inference from multilocus data. Second, with realistic mutation rates (i.e., $\theta \leqslant 0.01$) the predominance of unresolved estimated gene trees, rather than AGTs, within the anomaly zone suggests that the risk of AGTs, while they exist in theory, may rarely be realized in practice.

Why Are Asymmetric Gene Trees Less Likely to be Correctly Estimated?

The analyses suggest that the cause of the expanded anomaly zone with estimated gene trees (Fig. 3) is a deficit of correctly estimated asymmetric gene trees (Fig. 4a), as opposed to a significant increase in the representation of symmetric topologies among the misidentified gene trees (Fig. 4b). Why would asymmetric gene trees be less likely to be correctly reconstructed for a given species tree and mutation rate?

The answer appears to lie in the differing branch lengths of asymmetric and symmetric gene trees, with the shortest branch length in asymmetric gene trees being considerably shorter than the shortest branch length in a symmetric gene tree (Fig. 6). Because the length of a gene tree branch determines the probability

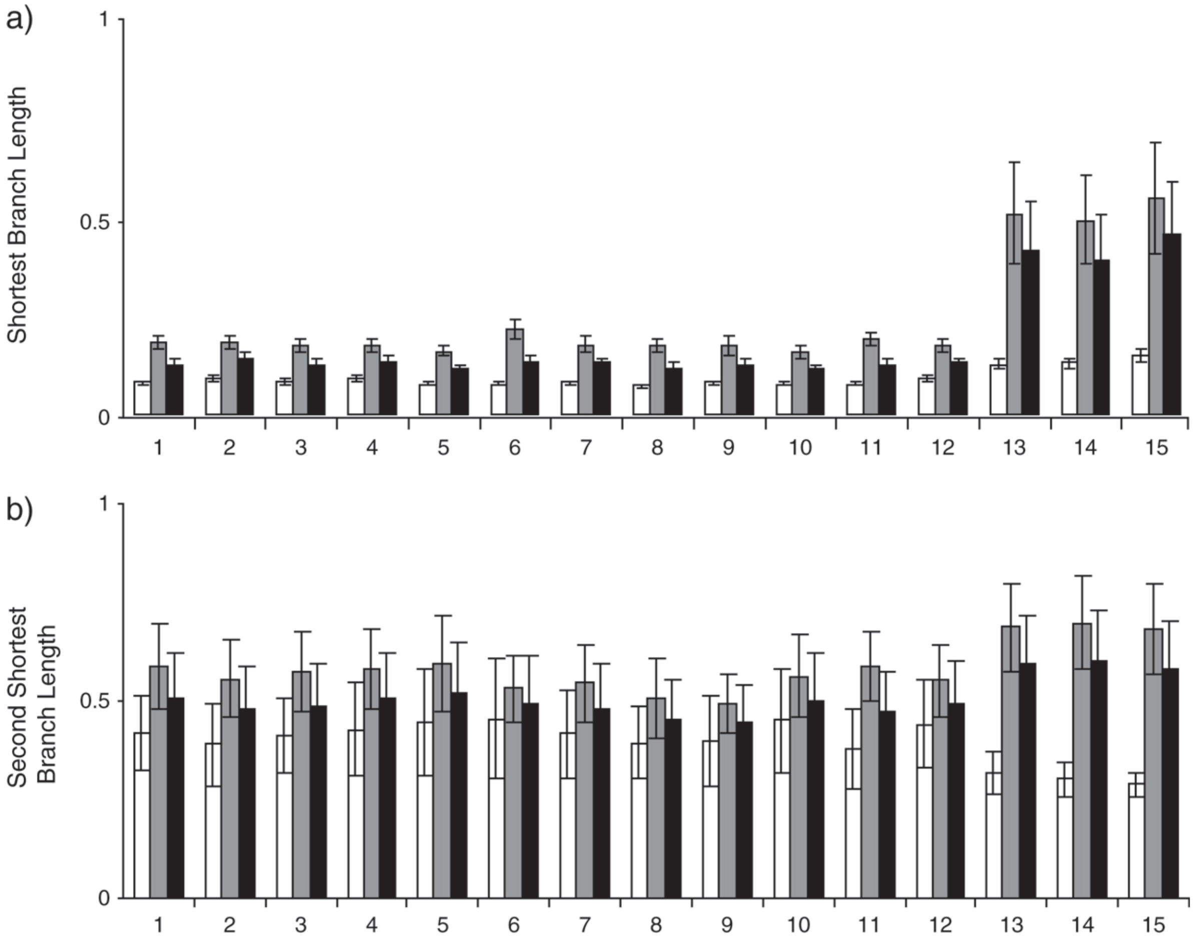

FIGURE 6. An example of how the average lengths (in units of 2N) of the a) shortest and b) second-shortest branch in estimated gene trees differ between asymmetric and symmetric topologies, based on 10 000 simulated data sets (standard errors are shown). The 15 different gene tree topologies are shown along the x-axis, identifying the average length of the shortest and second-shortest branch (shown in black), as well as the average length of the shortest and second-shortest branch for incorrectly estimated topologies (shown in white) and correctly estimated topologies (shown in grey). Asymmetric topologies are shown as 1 through 12, with 1 representing the one concordant with the species tree, and 13 through 15 are symmetric trees. The species tree is characterized by branch length $x = 0.10$ and $y = 0.088$.

density function for the number of mutations, the greatest effects of mutational variance will be manifest with the shortest branches. A similar pattern is observed when the second-shortest branch of the gene tree is considered—that is, the average branch length is shorter for asymmetric compared with symmetric topologies, though the effect is much less dramatic (Fig. 6b).

To confirm that branch length, and not some other factor related to topology per se, is responsible for the shift in the frequency distribution of topologies of estimated gene trees (relative to the frequency distribution of gene trees; Fig. 2), we examined the proportion of correctly estimated gene trees (P_C) for asymmetric and symmetric gene trees with the same average shortest branch (Table 1). When controlling for the differences in branch lengths (i.e., selecting species trees where asymmetric and symmetric gene trees had the same average shortest branch), there was no difference in the proportion of correctly estimated gene trees between asymmetric and symmetric topologies. This confirms that it is not topology per se (Table 1), but the length of the shortest branch, and specifically the relatively shorter branches of asymmetric compared with symmetric gene trees (Fig. 6), that results in a deficit of correctly estimated

TABLE 1. The proportion of correctly estimated gene trees (P_C) for the asymmetric topology that is concordant with the species tree and the symmetric coalescent gene trees when they have similar average shortest branch lengths; results are shown for 4 different species trees

Branch lengths[a]		P_C	
x	y	Concordant[b]	Symmetric[b]
0.075	6.297	0.645	0.653
0.125	4.828	0.695	0.675
0.175	4.860	0.689	0.664
0.225	4.328	0.698	0.677

[a] Branch lengths for x and y are shown in units of 2N.
[b] Based on 10 000 simulated replicate data sets for each species tree.

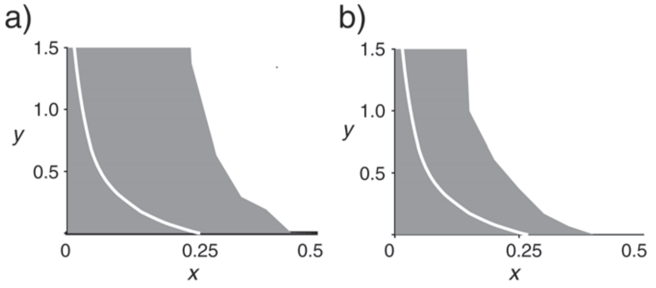

FIGURE 7. Distribution of species tree branch lengths (x and y are shown in units of $2N$) that define the boundary of the anomaly zone for midpoint rooted estimated gene trees (shaded area) with $z = 5N$ a) and $z = 15N$ b), under mutation rate $\theta = 0.01$. The white line marks the boundary for anomaly zones without mutational variance (i.e., based on analysis of gene trees).

asymmetric gene trees and hence an expansion of the anomaly zone.

By identifying the underlying cause for the expansion of the anomaly zone that occurs with estimated gene trees, there are several strategies empiricists can apply to avoid potential problems with AGTs. The first is a simple one—choose loci with fast mutation rates. This will minimize the effects of mutational variance, which leads to a smaller realized anomaly zone (see Fig. 3). Second, use out-groups to estimate species relationships not a distance-based procedure like midpoint rooting, where the effects of mutational variance may be further amplified. Indeed, a preliminary analysis shows a dramatic expansion of the anomaly zone with midpoint rooted trees (Fig. 7).

Evaluating the Danger of AGTs for Empirical Phylogenetic Study

Despite the occurrence of AGTs, and an expansion of the anomaly zone when mutational variance is considered (Fig. 3), our results also indicate that the danger AGTs actually pose for empirical phylogenetic study is limited. In contrast to coalescent genealogies, where the probability that more than 2 lineages will coalesce in the same generation is extremely small (Hudson 1990), polytomies dominate the estimated gene trees for the class of species trees located within the anomaly zone (Fig. 5a,b). Therefore, instead of encountering AGTs, the most probable estimated gene tree an empiricist is likely to recover (at least for typical mutation rates) is one that is uninformative about the species tree. In other words, by focusing on the estimated gene trees, as opposed to coalescent gene trees, the results show that when resolved estimated gene trees are likely (and hence there is the potential for AGTs), AGTs are no longer a threat because the species trees are not located within the anomaly zone (i.e., the species tree branch lengths are too long to generate AGTs). Moreover, the flatness of the frequency distribution of estimated gene trees within the anomaly zone, as revealed by the low maximum frequency of estimated gene tree topologies (Fig. 8), also lessens any real danger of AGTs for empirical phylogenetic study. For example, even if the frequency of one anomalous topology is 15%, and other topologies have equal frequencies (i.e., 6% for the other 14 possible topologies), the chance that the anomalous topology will be the most frequent one in a sample of 20 loci is only 50%. Moreover, because the frequency of AGTs is actually much lower than 15% across most of the anomaly zone (typically slightly less than 7%; Fig. 8), the chance that the most frequent topology among 20 loci will be the anomalous topology is indeed very, very low. Placing the danger posed by AGTs in an empirical context is important for highlighting the true challenges for phylogenetic study. For example, recently developed methods based on triplets offer one way that AGTs might be overcome (e.g., Ewing et al. 2008; Degnan et al. 2009). Nevertheless, if the actual problem with the anomaly zone is the lack of resolution, not AGTs (as shown here; also see Ewing et al. 2008), then these methods will do little to

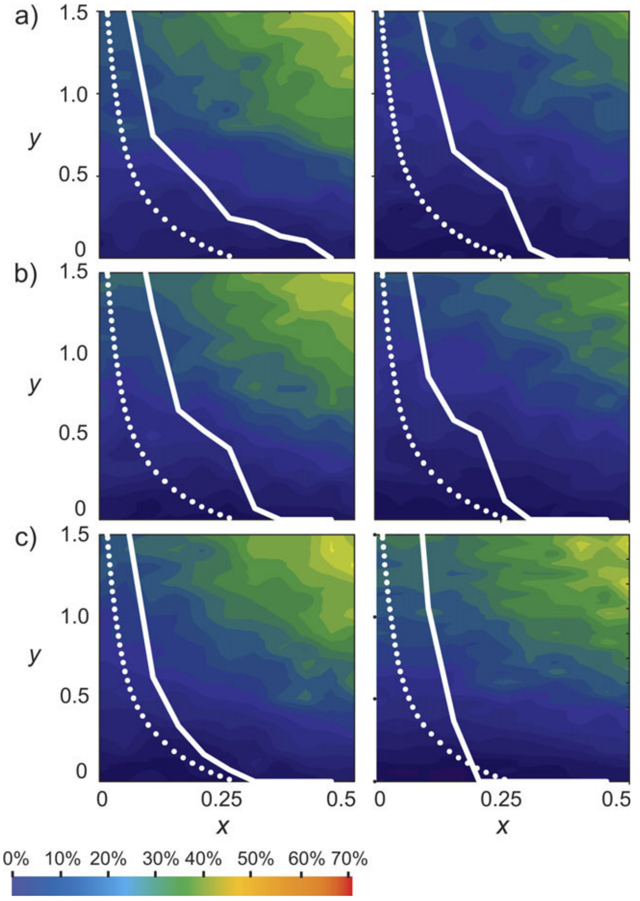

FIGURE 8. Highest frequency of resolved topologies in estimated gene trees for the different species trees (x and y branch lengths both are expressed in units of $2N$) with $z = 5N$ (shown on the left) and $z = 15N$ (shown on the right), under the 3 different mutation rates: a) $\theta = 0.005$, b) $\theta = 0.01$, and c) $\theta = 0.05$. The white solid and dashed lines mark the boundary of the anomaly zone based on estimated gene trees and coalescent gene trees, respectively.

address the challenges facing empirical phylogenetic study.

General Lessons from the Impact of Mutational Variance on Estimated Gene Trees

Investigation of the anomaly zone was motivated by the suggestion that the inherent mismatch between the most frequent estimated gene tree and the actual history of species divergence would pose significant (and perhaps insurmountable) challenges to obtaining an accurate estimate of species relationships (Degnan and Rosenberg 2006; Rosenberg and Tao 2008). However, our study indicates that the danger of AGTs in practice is not what it is in theory, once mutational variance is taken into account. This finding does not mean that the difficulties with estimating species relationships (i.e., the underlying species tree) have gone away. The predominance of the polytomy zone coupled with the low frequency of estimated gene trees with a topology matching the species tree (Fig. 8) in adjoining regions of parameter space that define the polytomy (Fig. 5) and anomaly zone (Fig. 3) indicates that the primary focus should be developing a method that accurately extracts the gene tree signal to infer the species tree. By considering the biological realities of both mutational and coalescent variances, the study has refined, and perhaps redefined, the problem by identifying what those challenges actually are for empirical phylogenetic study. Therefore, it is informative to consider the implications of our results for the procedures we might use to estimate species trees (Maddison 1997).

With respect to estimating a species tree, the available methodological procedures differ in how they extract information about the underlying history of species divergence, as well as the type of information they utilize. Studies have shown that depending on the approach, the accuracy of the estimated species trees can be similar across methods for some species histories (namely, older species divergences) but may differ considerably, especially for recently diverged species (e.g., Brumfield et al. 2008; McCormack et al. 2009). Such studies highlight the potential gains that complex procedures, which extract more information from the estimated gene tree data, can offer for estimating species trees. Although it may indeed be desirable to fully utilize the information contained in the estimated gene trees, this study suggests that some caution is warranted. More information will be better than less, but only as long as the information being extracted from estimated gene trees is accurate. For example, in the case of AGTs, more loci will not necessarily provide a more accurate estimate of the species tree—the most probable gene trees (and therefore, the most frequent topology) will not match the underlying species tree (Degnan and Rosenberg 2006). Likewise, for recent species divergence where the effects of mutational variance are exaggerated, an estimated gene tree will not faithfully reflect the genealogical history, differing not only in branch lengths but also in topology (Fig. 4). Even for species trees outside the anomaly zone, the proportion of correctly estimated gene trees (P_C) was below 75% (Table 1). These estimates should also be considered conservative, given that in this case the DNA substitution model used to obtain the estimated gene trees was known (i.e., it matched exactly the conditions under which the data were generated) and that an exhaustive search was performed to estimate the gene trees. In other words, the effects of mutational variance may be greater with empirical data than what is documented here. Consequently, it may not be possible to accurately estimate species relationships for such histories (i.e., those characterized by rapid and recent speciation), even with recently developed methods for directly inferring the underlying species tree (e.g., Maddison and Knowles 2006; Carling and Brumfield 2007; Liu and Pearl 2007; Ewing et al. 2008; Knowles and Chan 2008; Kubatko et al. 2009; McCormack et al. 2009).

Another important issue and open question is how many loci and how much variation in the loci are needed to obtain an accurate estimate of the species tree, irrespective of what approach might be used for estimating species trees (e.g., minimizing the number of deep coalescences, estimating the most likely species tree, or a Bayesian analysis for a species tree; Maddison and Knowles 2006; Liu and Pearl 2007; Kubatko et al. 2009). Studies have shown that sampling effort and design have a significant impact (e.g., Maddison and Knowles 2006; Edwards et al. 2007; McCormack et al. 2009); however, systematic investigation is lacking for recently developed methods, especially with regard to the mutational process. As shown here, mutational variance can cause not only a mismatch between an estimated gene tree and the underlying gene tree for any single locus but also results in a flatter frequency distribution of tree topologies than expected (Degnan and Salter 2005). This is expected to increase the number of sampled loci needed to obtain an accurate characterization of the underlying probability distribution of estimated gene tree topologies for a given species tree. It remains to be determined if the realized anomaly zone is as intractable as the classic Felsenstein zone, another example where phylogenetic accuracy is compromised by the mutational process. Mutational variance, as an inevitable part in empirical sequence data, obviously needs to be investigated in the context of species tree estimation.

Funding

National Science Foundation (DEB 0447224 and DEB 0715487 to L.L.K.).

Acknowledgments

We would like to thank James Degnan for helpful discussion and comments on previous versions of the manuscript, as well as suggestions by Laura Kubatko, Greg Ewing, and an anonymous reviewer.

References

Brumfield R.T., Liu L., Lum D.E., Edwards S.V. 2008. Comparison of species tree methods for reconstructing the phylogeny of bearded manakins (Aves: Pipridae, Manacus) from multilocus sequence data. Syst. Biol. 57:719–731.

Carling M.D., Brumfield R.T. 2007. Gene sampling strategies for multi-locus population estimates of genetic diversity (theta). PLoS ONE. 2:e160.

Carling M.D., Brumfield R.T. 2008. Integrating phylogenetic and population genetic analyses of multiple loci to test species divergence hypotheses in Passerina buntings. Genetics. 178:363–377.

Carstens B.C., Knowles L.L. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from Melanoplus grasshoppers. Syst. Biol. 56:400–411.

Degnan J.H., DeGiorigio M., Bryant D., Rosenberg N.A. 2009. Properties of consensus methods for inferring species trees from gene trees. Syst. Biol. 58:35–64.

Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.

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

Drost J.B., Lee W.R. 1995. Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among drosophila, mouse, and human. Environ. Mol. Mutagen. 25(Suppl 26):48–64.

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

Ewing G.B., Ebersberger I., Schmidt H.A., von Haeseler A. 2008. Rooted triple consensus and anomalous gene trees. BMC Evol. Biol. 8:118.

Hudson R.R. 1990. Gene genealogies and the coalescent process. In: Futuyma D., Antonovics J., editors. Oxford surveys in evolutionary biology. Vol. 7. New York: Oxford University Press. p. 1–44.

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

Huelsenbeck J.P., Kirkpatrick M. 1996. Do phylogenetic methods produce trees with biased shapes? Evolution. 50:1418–1424.

Jennings W.B., Edwards S.V. 2005. Speciational history of Australian grass finches (Poephila) inferred from thirty gene trees. Evolution. 59:2033–2047.

Kingman J.F.C. 1982. The coalescent. Stoch. Process. Appl. 13:235–248.

Knowles L.L., Carstens B.C. 2007. Estimating a geographically explicit model of population divergence. Evolution. 61:477–493.

Knowles L.L., Chan Y.-H. 2008. Resolving species phylogenies of recent evolutionary radiations. Ann. Mo. Bot. Gard. 95:224–231.

Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics. 25:971–973.

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

Liu L., Pearl D.K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.

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

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

McCormack J.E., Huang H., Knowles L.L. 2009. Maximum-likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. Syst. Biol (in press).

Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

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

Ripplinger J., Sullivan J. 2008. Does choice in model selection affect maximum likelihood analysis? Syst. Biol. 57:76–85.

Rosenberg N.A., Tao R. 2008. Discordance of species trees with their most likely gene trees: the case of five taxa. Syst. Biol. 57:131–140.

Saitou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Sanderson M.J. 2008. Phylogenetic signal in the eukaryotic tree of life. Science. 321:121–123.

Swofford D.L. 2000. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4.0b. Sunderland (MA): Sinauer Associates.

Takahata N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics. 122:957–966.

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

Wong A., Jensen J.D., Pool J.E., Aquadro C.F. 2007. Phylogenetic incongruence in the Drosophila melanogaster species group. Mol. Phylogenet. Evol. 43:1138–1150.

Zhou H., Gu J., Lamont S.J., Gu X. 2007. Evolutionary analysis for functional divergence of the toll-like receptor gene family and altered functional constraints. J. Mol. Evol. 65:119–123.

Received 9 December 2008; reviews returned 8 April 2009; accepted 26 July 2009
Associate Editor: Laura Kubatko