Article Type: Special Issue Article

RESEARCH ARTICLE

INVITED SPECIAL ARTICLE

For the Special Issue: Using and Navigating the Plant Tree of Life

Short Title: Knowles et al.—Cause of gene tree discord

A matter of phylogenetic scale: Distinguishing incomplete lineage sorting from lateral gene transfer as the cause of gene tree discord in recent versus deep diversification histories

L. Lacey Knowles[1,2], Huateng Huang[1], Jeet Sukumaran[1], and Stephen A. Smith[1]

Manuscript received 31 October 2017; revision accepted 30 January 2018.

[1] Department of Ecology and Evolutionary Biology, 1109 Geddes Avenue, Museum of Zoology, University of Michigan, Ann Arbor, MI 48109-1079 USA

[2] Author for correspondence (e-mail: knowlesl@umich.edu)

Citation: Knowles, L. L., H. Huang, J. Sukumaran, and S. A. Smith. 2018. A matter of phylogenetic scale: Distinguishing incomplete lineage sorting from lateral gene transfer as the cause of gene tree discord in recent versus deep diversification histories. *American Journal of Botany* 105(3): XXX.

DOI: XXXX

PREMISE OF THE STUDY: Discordant gene trees are commonly encountered when sequences from thousands of loci are applied to estimate phylogenetic relationships. Several processes contribute to this discord. Yet, we have no methods that jointly model different sources of conflict when estimating phylogenies. An alternative to analyzing entire genomes or all the sequenced loci is to identify a subset of loci for phylogenetic analysis. If we can identify data partitions that are most likely to reflect descent from a common ancestor (i.e., discordant loci that indeed reflect incomplete lineage sorting [ILS], as opposed to some other process, such as lateral gene transfer [LGT]), we can analyze this subset using powerful coalescent-based species-tree approaches.

METHODS: Test data sets were simulated where discord among loci could arise from ILS and LGT. Data sets where analyzed using the newly developed program CLASSIPHY (Huang et al., 2018) to assess whether our ability to distinguish the cause of discord among loci varied when ILS and LGT occurred in the recent versus deep past and whether the accuracy of these inferences were affected by the mutational process.

KEY RESULTS: We show that accuracy of probabilistic classification of individual loci by the cause of discord differed when ILS and LGT events occurred more recently compared with the distant past and that the signal-to-noise ratio arising from the mutational process contributes to difficulties in inferring LGT data partitions.

CONCLUSIONS: We discuss our findings in terms of the promise and limitations of identifying subsets of loci for species-tree inference that will not violate the underlying coalescent model (i.e., data partitions in which ILS, and not LGT, contributes to discord). We also discuss the empirical implications of our work given the many recalcitrant nodes in the tree of life (e.g., origins of angiosperms, amniotes, or Neoaves), and recent arguments for concatenating loci.

KEY WORDS: CLASSIPHY; coalescence; gene-tree discord; incomplete lineage sorting; lateral gene transfer; species tree

When phylogenetic relationships among species are examined using genomic or transcriptomic scale data sets, the discord (e.g., incongruent branching patterns) among individual gene trees is clear. There are various processes that can result in this discord, such as incomplete lineage sorting (ILS), lateral gene transfer (LGT), hybridization (H), and gene duplication and loss (DL). In addition to these biological processes that generate discord, discord may be due to a lack of informative phylogenetic data or to errors in sequence assembly. Disagreement between gene trees has traditionally posed a challenge for phylogenetic analysis, especially when the only strategy for combining data from different genes was simple concatenation, which effectively treats discord as noise, rather than fundamental structure of systems. However, significant advances for estimating phylogenetic relationships despite gene tree discord have been made with species-tree methods, which explicitly model the coalescent process, turning one source of discord, ILS, into a source of information rather than noise (Edwards, 2009; Knowles and Kubatko, 2010).

There have also been recent developments for modeling sources of discord other than ILS when estimating phylogenetic relationships (e.g., gene duplication and loss: Boussau et al., 2013; hybrid origin of taxa: Bilschak et al., 2017; Meng and Kubatko, 2009; networks: Solís-Lemus and Ané, 2016; Solís-Lemus et al., 2017; Zhang et al., 2018; Wen and Nakleh, 2017). However, adequate methods do not exist that simultaneously estimate phylogenetic trees, model sources of conflict, model molecular substitution, and perform well when more than one cause of discord are considered (Boussau et al., 2013). As a result, slight changes to data set assembly and/or phylogenetic reconstruction methods often generate different species trees (Jarvis et al., 2014, Wickett et al., 2014; Xi et al., 2014).

These studies emphasize that the key to resolving relationships lies not just with more data, but also with decisions about which data to include and what analyses to apply (e.g., Smith et al., 2015; Brown and Thompson, 2017; Shen et al., 2017). Model misspecification, for example, of the model of molecular evolution (e.g., nonstationarity of composition, Foster, 2004; Morgan et al., 2013; Cox et al., 2014; Jarvis et al., 2014) or gene-tree evolution (e.g., ignoring coalescent-based variation among loci; Kubatko and Degnan, 2007), has been shown to dramatically reduce the accuracy of phylogenetic reconstruction. However, models cannot currently, nor are they likely soon to be, capable of accommodating all the heterogeneity and complexity in full genomes and transcriptomes. Consequently, approaches that focus on identifying subsets of data that conform to the assumptions of, or are otherwise optimized for, the particular models used in a given phylogenetic analysis, have been expanding (e.g., Huang et al., 2016, 2018; Brown and Thompson, 2017; Richards et al., 2017).

Different criteria might be applied to identify which loci from a larger pool might be included in a phylogenetic analysis. For example, loci may be chosen based on characterizations of their phylogenetic signal (Gori et al., 2016; Huang et al., 2016; Lewitus and Morlon, 2016). As an alternative to using a statistical criterion to reduce the heterogeneity in data that does not consider what processes underlie the discord, the biological basis of the discord might be considered explicitly when identifying data partitions. For example, data partitions might be based on whether discord is caused by ILS versus LGT using the recently developed program CLASSIPHY (see Huang et al., 2018). Such data partitions, like the characterizations based on statistical criteria that are agnostic to cause (see Gori et al., 2016), cannot only be used to avoid model misspecification (i.e., only loci for which discord arises from ILS might be included in a species-tree analysis of phylogenetic relationships), but they may also provide additional information that is of biological interest—the proportion of loci evolving under different evolutionary processes (see Huang et al., 2018). That is, discord among loci is more than just a statistical inconvenience, but can be usefully leveraged to inform and improve the analysis if the underlying process can correctly be identified and modeled.

A newly developed method (CLASSIPHY) provides for the identification of the processes that generate discord in a given locus (Huang et al., 2018), and apart from the fundamental utility of this (e.g., for phylogenetic inference as described above), this method also allows us to ask questions about the evolution of the discord itself. Here we use this method to ask two key questions about the utility of this concept in practice. First, how sensitive is the accurate classification of loci to the diversification history itself? Second, how is the accuracy of data partitions (i.e., the inferred subsets of loci with discord due to ILS) influenced by the mutational process? We answer these questions using simulated test data, so that we have a priori knowledge of the identities of the particular loci that are discordant due to ILS versus LGT, as well as the timing of ILS and LGT events themselves.

By assessing the effects of the timing of divergence, as well as the mutational process, on our ability to distinguish ILS from LGT as the cause of discord among loci (see Fig. 1) in the present study, we provide a critical context for empirical applications of the program CLASSIPHY (Huang et al., 2018) given that ILS and LGT events in practice may occur in the recent or more distant evolutionary past and that real data sets are comprised of DNA sequences. We discuss the relevance of our findings to strategies for resolving the recalcitrant nodes that have come to characterize many deep nodes in the history of divergence of different clades (e.g., birds and plants; Mirarab et al., 2014; Wickett et al., 2014), but also to debates on best practices (e.g., over concatenation versus species-tree estimation; de Queiroz and Gatesy, 2007; Kubatko and Degan, 2007; Zhong et al., 2013; Gatesy and Springer, 2014; Liu et al., 2014; Xi et al., 2014).

<H1>MATERIALS AND METHODS

We use simulation-based testing to examine the robustness of inferred data partitions to the timing of ILS and LGT events, as well as the mutational process (see Fig. 1), for which the contribution of ILS and LGT to discord and mutational effects on gene tree estimation are known. Because we simulated our test data, we know which loci have discord arising from ILS versus LGT events, and thus, we can evaluate whether we can accurately distinguish between discord due to ILS versus LGT by comparing the known contribution of ILS and LGT with the probabilistically inferred contribution of ILS and LGT for each of the test data sets (details below), using the program CLASSIPHY (Huang et al., 2018), which is freely available on GitHub (https://github.com/huatengh/Classiphy).

<h2>Simulated test data sets

Test data sets were simulated where discord among loci could arise from ILS and LGT. Only LGT events that induce a topological discord are considered, and hereafter are simply referred to as LGT loci (i.e., LGT events that do not alter the topology of a gene tree are not classified as LGT loci). The rest of the loci, which do not contain LGT events, but could contain discord due to ILS, are referred to as ILS loci.

To generate the test data sets, three separate steps are involved to simulate LGT, lineage sorting, and mutation, respectively (Fig. 1A). Specifically, starting with a species tree: (1) locus trees were simulated with random LGT events, (2) genealogies were simulated within the locus tree according to coalescent process (i.e., simulating random ILS events), and (3) nucleotide data sets were simulated on the genealogies with a substitution model (details see below). For isolating the effects of the timing of divergence on the ability to discriminate ILS from LGT loci, all test data sets were simulated under a single species tree in which the relative timing of divergence among taxa remained the same, but the total depth of the species tree was increased (Fig. 1B). Specifically, the test data were simulated under a 50 taxon species tree with either a total depth of 25N (and referred to as “shallow”), or one with a total depth of 100N (and referred to as “deep”) with the additional branch length added to the tips of the tree (see Fig. 1B), rather than rescaling all the branches. However, to avoid introducing additional LGT events on these extended terminal branches, the extra length was added to the simulated genealogies (i.e., after step 2 from above), rather than changing the species tree itself. As such, the distribution of ILS and LGT events was held constant; only the absolute timing of specific events shifted.

To examine the effect of mutation rates on our ability to distinguish ILS versus LGT loci, we compared the accuracy of the inferred data partitions defined by loci with discord due to ILS, but not LGT, under two conditions: coalescent gene genealogies versus estimated gene trees from nucleotides of individual tests data sets were analyzed in CLASSIPHY (Fig. 1A). That is, the phylogenetic estimate for a locus (i.e., the gene tree inferred from nucleotide data) may differ from the actual genealogy of that locus (i.e., the coalescent history of the locus) because of limited phylogenetic signal (for more details on the mismatch between estimated gene trees and coalescent genealogies due to mutational variance see Huang et al., 2009, 2014; Lanier et al., 2013). This potential mismatch between the gene genealogy and estimated gene tree on the accuracy of data partitions is relevant to classifying empirical data, and therefore is included here, given the distribution of homoplasy is dependent upon the diversity history, and its effects on the performance of CLASSIPHY have not yet been investigated.

<h2>Choice of parameters in simulated test data

All test data sets were simulated under the same species tree (Fig. 1B). This species tree was chosen to be representative of a history with an average amount of discord due to ILS. Specifically, the species tree was identified from a set of 100 species trees simulated under Yule birth-and-death model (with speciation rate = 2 × extinction rate), for which 500 genealogies were simulated, and the average RF distance between species tree and its genealogies (i.e., ILS-caused distance) was calculated. The chosen tree (Fig. 1B) was then identified by ranking all 100 species trees according to the species-tree-genealogy distance, selecting the species tree closest to the mean distance. Hence, the species tree used in this study is a random 50-taxon tree at 25N depth with an average level of expected discord due to ILS.

Based on this species tree, we simulated 1000 different data sets. Each data set consisted of 800 locus trees (i.e., 800 independent loci; see Fig. 1), but the data sets differed in their respective rates of LGT, which ranged from 2e-10 to 2e-9 LGT events per generation. Since LGT events were introduced at random, in each data set of 800 locus trees, there are locus trees that differ topologically with the species tree (i.e., LGT between nonsister lineage), those with the same topology but different branch lengths (i.e., LGT between sister lineages), and those having no LGT events. The range of LGT rate was chosen such that 90% of the locus-trees contain 2–15% LGT loci. Since estimating large numbers of gene trees is computationally intensive, only three data sets with different amounts of LGT loci were selected for simulating nucleotide sequences and estimating gene trees. These data sets were identified by ranking the data sets based on the proportion of LGT loci (i.e., ranking were established after step 2 of the simulation procedure; see Fig. 1A), and selecting the three data sets at the 25%, 50%, and 75% quantile. These test data sets, each comprising 800 loci, contained 4.9%, 7.5% and 11.5% LGT loci, respectively. Genealogies with one individual per species were simulated for each locus tree according to the coalescent model (i.e., a genealogy may differ from its locus tree because of ILS).

We used SimPhy (Mallo et al., 2016) for simulating the species trees, locus trees and genealogies above (a wrapper function to use Simphy is included in the CLASSIPHY R package), and used Seq-Gen (Rambaut and Grassly, 1997) to simulate nucleotide data sets on the genealogies. For each genealogy (under the shallow versus deep history of 25N versus 100N total depth, respectively), nucleotide data sets of 1000 bp were simulated with the program Seq-Gen (Rambaut and Grassly, 1997) under an HKY85 model of nucleotide substitution with a transition–transversion ratio of 3.0, a gamma mutation rate distribution with shape parameter of 0.8, and nucleotide frequencies of A = 0.3, C = 0.2, T = 0.3, and G = 0.2 for the ancestral sequence. From the simulated DNA sequences, gene trees were estimated using RAxML (Stamatakis, 2014). These estimated gene trees may differ from the actual genealogy because of limited phylogenetic signal (for more details on the mismatch between estimated gene trees and coalescent genealogies due to mutational variance see Huang et al., 2009, 2014; Lanier et al., 2013). An outgroup lineage (but with no LGT between the outgroup and in-group lineages) was used to root estimated gene tree (Fig. 1B). The outgroup is not included in any of the calculated summary statistics used in the CLASSIPHY analyses (i.e., the outgroup does not contribute to the classification of loci as ILS versus LGT loci).

In total, there were 12 test data sets: test data sets for each of 3 different LGT quantiles for a shallow versus deep divergence history, based on either the estimated gene trees or the genealogies themselves (after excluding them from the training data set). Each test data set was analyzed separately using CLASSIPHY (Huang et al. 2018), as described below. The test data sets, and the parameter file for the simulated training sets, are freely available from GitHub (https://github.com/huatengh/Classiphy).

<h2>Classification of loci by the cause of discord

CLASSIPHY is a simulation-trained (supervised) machine learning approach (Huang et al. 2018). Unlike traditional machine learning approaches, which typically use empirical data for both training and evaluation, in the CLASSIPHY approach, training data are generated under known processes of ILS and LGT, following Sukuumaran et al. (2016), and the entire process is described in detail below. The machine learning algorithm used in CLASSIPHY is discriminant analysis of principal components (DAPC) and has been described comprehensively (Jombart, 2008; Jombart et al., 2010). Briefly, this algorithm involves calculating a set of summary statistics on the training data, projecting these statistics onto principal component axes, and using the principal component axes scores as input to construct a discriminant analysis classifier, which in turn is applied to the target data to classify them with respect to the generating model. While there are many machine learning algorithms available, we have found that the DAPC performs well enough for applications such as this (Sukumaran et al, 2016) to base our analyses on it and has been borne out through our own assessments (Huateng et al., 2018).

The basic steps involved in this simulation-trained DAPC procedure are (1) simulation of gene trees under regimes corresponding to different processes that might contribute to discord—in this case, ILS and LGT, (2) calculation of summary statistics on simulated data sets to train a classification function, (3) construction of a discriminant analysis function based on principal components extracted from the training data set, (4) assessment of the performance of the summary statistics by inspection of posterior prediction of the training data set, and (5) application of the discriminant analysis function to the original data to classify it with respect to whether ILS or LGT underlie observed gene tree discord. The summary statistics are not used directly, but rather the principal components extracted from the summary statistics are used to construct the DAPC function in this machine learning approach. For a detailed description of CLASSIPHY, see Huang et al. (2018), which describes the concept of identifying data partitions by the biological cause of discord and demonstrates the validity of the statistical approach applied in CLASSIPHY. Here, we limit our focus to questions that can be answered statistically by application of this method to data sets that differ with respect to the timing of ILS and LGT events, and the effects of the mutational process on the accuracy of identifying data partitions (i.e., data subsets with ILS, but no LGT).

The training set applied in the CLASSIPHY analyses here comprised 1000 data sets with different LGT rates, each with 800 loci, where the rate of LGT was drawn from the distribution described above, simulated under the one species tree (Fig. 1B). For each of the 800 loci of each test data set (i.e., a total of 800 loci × 12 data sets), a probabilistic classification was generated using a standard posterior probabilities >0.5 threshold to classify a locus as either an ILS locus or LGT locus (see Huang et al. 2018 for other thresholds that might be applied using CLASSIPHY).

<h2>*Assessing accuracy of data partitions*

The inferred classification of individual loci from the CLASSIPHY analysis was compared to the actual history of each locus to evaluate the accuracy of distinguishing ILS and LGT loci. Accuracy of the data partitions are summarized separately for (1) each of the three LGT rates (i.e., a low, medium, and high LGT rate), and (2) the different depths of divergence (i.e., shallow versus deep divergence histories). For each of these separate scenarios, a receiver operating characteristic (ROC) curve analysis was performed using the *pROC* R package (Robin et al., 2011). Such analyses are commonly used to characterize and compare the results from machine learning approaches. ROC plots provide a visualization for assessing the performance of the classifier over its entire operating range, as opposed to relying just on the area under the curve (the AUC) to evaluate the classifier. In addition, linear regressions were used to test whether or not the posterior probability of LGT correlated with the degree of discord between a gene tree and the species tree (i.e., the species-to-locus Robinson-Fould’s distance).

To examine the effect of mutation on the ability to accurately distinguish ILS and LGT loci, the classification accuracy of data partitions was compared when the genealogy versus the estimated gene tree was analyzed with CLASSIPHY. These results are presented after standardizing by the overall accuracy of classification for each of the three LGT rates (i.e., a low, medium, and high LGT rate), and the different divergence depths (i.e., shallow versus deep divergence histories), to establish the effect of mutation on accurate classification (as opposed to inherent differences in the accurate classification of individual loci). Specifically, the differences in the percentage correct classification when based on genealogies versus estimated gene trees were calculated and presented.

<H1>**RESULTS**

There are two important observations about the performance as measured in terms of posterior probabilities for the true versus false model (Fig. 2). First, as visualized in the ROC curves (Fig. 2), for loci with high posterior probabilities, the method is sensitive to both (1) the rate of LGT and (2) whether these events occur in the recent versus distant past (see Fig. 1 for simulation design). For example, in all cases, irrespective of the rate of LGT, the accuracy of data partitions (i.e., classification of ILS and LGT loci) decreases when those events are in the more distant past, (Fig. 2). In these curves, the true positive rate (TP, representing sensitivity) is plotted against the false positive rate (FP, representing 1 – specificity) for the $p > 0.5$ threshold used here to classify ILS and LGT loci. This sensitivity to whether the events causing discord occurred in the recent versus distant past is also reflected in the difference in the summary provided by the AUC scores for each test data set with either low, medium, or high proportions of discord cause by LGT (see Fig. 2). More specifically, there is a drop in classification performance (i.e., lower AUC scores) with higher proportions of LGT events and when the events occur in the distant past. Note that because the additional branch lengths were added to the tips of the tree (see Fig. 1B), rather than rescaling all the branches, and these were added to the simulated genealogies after step 2 as described in the methods (Fig. 1A), the difference in performance can only arise from shifts in the absolute timing of specific events (i.e., no additional LGT events were introduced by the extended terminal branches of the deep history of species divergence; Fig. 1B).

We can see that the largest decrease in the percentage of loci classified accurately ranges from about 8% to 15% and is associated with the deep divergence histories (Table 1). Moreover, this analysis also shows how the drop in the accuracy of CLASSIPHY with the depth of the divergence events observed in the ROC analyses (Fig. 2) primarily reflects the decreased accurate classification of LGT loci, not ILS loci. This result highlights that, while phylogenetic scale matters (that is, whether the processes generating discord occurred in the recent versus deep past), the data partitions representing ILS loci tend to be more accurate relative to identifying data partitions of LGT loci (Table 1). This sensitivity in identifying data partitions of LGT loci when the events occurred in the more distant past can be visualized by the relationship between the posterior probability of LGT and the degree of discord between a gene tree and the species tree (i.e., the species-to-locus Robinson-Fould's distance). This relationship is clear when the LGT events occurred in the recent past, but it becomes degraded when those LGT events occur in the more distant past (Fig. 3).

<H1>**DISCUSSION**

Our examination into how the accuracy of distinguishing ILS and LGT depends on the timing of these events has immediate implications for applications using the program CLASSIPHY. However, our work also points to more general issues surrounding the decisions that researchers make about how to handle topological discordance across loci, as well as the lack of phylogenetic signal. We also acknowledge that much more work needs to be done before informed decisions about best practices might be made. Nevertheless, our work is an example of how the field can take steps toward characterizing the relative contributions of different sources of discord, and by doing so, potentially improve phylogenetic estimates.

<h2>Identifying the cause of discord and implications for what to do about discordant trees

The approach implemented in CLASSIPHY (Huang et al., 2018) and the analyses discussed here are important not only for reconstructing species trees, but also for exploring the processes that lead to discord in phylogenies. We are still in the early stages of analyzing and understanding large genomic and transcriptomic data sets. Significant technological and methodological challenges have already been overcome but more continue to arise.

Despite the relative newness of large genomic data sets, some major patterns are emerging. For example, it has become clear that simply adding more data is not going to confidently resolve all the recalcitrant nodes across the tree of life. As recent studies have demonstrated, gene tree discord is very common and can show diverse patterns (Jarvis et al., 2014; Smith et al., 2015; Brown and Thompson, 2016; Shen et al., 2017). Furthermore, these studies have demonstrated that a relatively small number of genes can dramatically alter species tree estimates and the probability of including "outlier" genes (i.e., genes contributing to model misspecification in phylogenetic inference) increases as we increase the amount of data because the inherent heterogeneity of sequence data can only increase with additional taxa and loci. This additional data complexity necessarily complicates our ability to reconstruct phylogenies.

Our results suggest that there is the potential to filter data sets for genes in which the conflict is due to evolutionary processes that can be correctly modeled so as to inform, rather than distort, the phylogeny—specifically, identifying data partitions of ILS loci (see Fig. 2 and Table 1). However, our results also suggest that the ability to accurately classify loci by the cause of discord depends on the diversification history itself. More specifically, the accuracy of the CLASSIPHY approach (Huang et al., 2018) is not strictly a function of the rate of LGT, but instead depends upon the timing of those LGT events, with events in the distant past being classified less accurately than those in the recent past (even for the same rate of LGT) (Fig. 2). This behavior presents challenges for studying the process and patterns of LGT and relates to our ability to test hypotheses about the role of LGT in the diversification of some groups (e.g., Xi et al., 2012).

Nevertheless, one of the most compelling aspects of our results is that discord due to ILS tends to be accurately identified, irrespective of whether the events took place in the recent or more distant past (Table 1). Moreover, this result is generally fairly robust to mutational variance (i.e., there is not much of a difference in the classification accuracy of ILS loci based an estimated gene tree versus the actual genealogy; Table 1). Limited phylogenetic signal that might contribute to differences between estimated gene trees and the actual genealogies, at least for the parameter space considered here, is not a significant problem. This finding has important implications for decisions researchers might make about phylogenetic analysis and, in particular, estimating species trees (Knowles and Kubatko, 2010). First, it dispels a common misconception that species tree approaches may not be appropriate when divergence occurs in the more distant past because, within any species, individuals will have coalesced to a common ancestor. As our results clearly show, ILS, whether it happens in the recent past or more distant past, can be detected, though with somewhat lower accuracy (Table 1), even with a single individual sequenced per species. In other words, the discord arising from the random sorting of gene lineages, which occurs irrespective of whether divergence is recent or in the distant past, should not be confused with the distinct concept of optimal sample design, and how sampling more individuals might or might not be useful to phylogenetic inference (e.g., when diversification occurs in the more distant past, sampling more individuals will not improve phylogenetic estimates; see McCormack et al., 2009; Knowles, 2010).

The second important implication of these results on the classification of ILS loci bears on whether the lack of phylogenetic signal of individual genes is necessarily a legitimate reason to concatenate data (e.g., Jarvis et al., 2014). Unlike the detection of LGT trees, where there is a fairly substantial effect of mutational variance for deeper histories (Table 1; see also Fig. 3), the detection of ILS trees is consistent through time (Table 1). Again, this suggests that at least for the parameter space studied here, the ability to detect ILS is generally robust even when those events occur in the recent or distant past. As such, the results bolster arguments that with improved model fit (i.e., accurate modeling of the nucleotide substitution process such that estimated gene trees match the underlying genealogies), and when ILS is the primary source of discord among the gene trees analyzed, species tree analyses can be accurate for phylogenetic inference. Whole genomes or transcriptomes likely have many processes that shape gene tree evolution and so instead of presuming that all the discord is the result of ILS, we may be able to identify and use those ILS loci for species tree construction. However, the difference between the classification accuracy of LGT loci based on the actual genealogical histories versus the gene trees themselves, as well as reduced accuracy of LGT loci for deeper species divergence times relative to more shallow histories (Table 1), suggests that limited phylogenetic signal may become problematic (see also Richards et al. 2017).

<h2>Accurately estimating species' phylogenies and moving beyond the species tree

Detailed interrogation of genomic data sets (e.g., Fontaine et al., 2015; Smith et al., 2015; Shen et al., 2017) has provided clear evidence that processes other than the coalescent (reviewed by Maddison, 1997) contribute significantly to gene-tree discord. The importance of proper modeling of data (e.g., Kubatko and Degnan, 2007; Ruprecht et al., 2017), and the impact of the different sources of discord on phylogenomic analyses, highlights that data abundance alone will not be sufficient to infer accurate inference of species relationships. Debates on best practices (e.g., over concatenation and coalescence, de Queiroz and Gatesy, 2007; Kubatko and Degnan, 2007; Edwards, 2009; Zhong et al., 2013; Gatesy and Springer, 2014; Liu et al., 2014; Xi et al., 2014) typically do not address the many sources of discord that contribute to conflict. Perhaps more importantly, these discussions often do not consider that, in addition to the better construction of species trees, analyses of the patterns of conflict also lead to a better understanding of the evolutionary processes and events that occurred within the lineages being analyzed. As such, our study represents an important step toward characterizing the relative contributions of different sources of discord. Instead of simply concatenating all the data, which violates our models of evolution, we might examine the data in more detail, and if we can identify those genes where discord is the result of ILS, we may have an opportunity to better resolve species relationships. However, even though we demonstrate that the identification of ILS genes is possible under the parameter space explored here, empirical data are more complex. For example, as the scope of a particular phylogenetic analysis increases, the probability of having multiple processes influence the evolution of a single gene tree also increases. And so our results, along with those of other genomic studies over the last few years, suggest that decisions about phylogenetic analyses will be more nuanced, contrary to debating which one method might be best (i.e., simply assuming that concatenation will avoid unwanted problems is not a justifiable position). As data sets continue to expand in taxonomic and genomic coverage, how we might achieve the most accurate phylogenetic estimates, while at the same time, extract information about the evolutionary processes structuring phylogenomic data, is a pressing question that deserves more attention.

What approach might researchers take to reach a balance between data content and model fit to achieve accurate phylogenetic inference? Despite compelling arguments for improved model fit to increase the accuracy of phylogenetic inference, and given the difficulties in analyzing big data, some researchers have shifted back to the use of concatenated data sets with only nucleotide evolution modeled in the inference procedure (e.g., Jarvis et al., 2014; Wickett et al., 2014; Prum et al., 2015; Yang et al., 2015), prompting others to attempt to argue for the superiority of concatenation specifically (e.g., Gatesy and Springer, 2017). Without arguing for or against the specific application of these methods to particular data sets, the lessons from genomics and transcriptomics over the last few years have demonstrated that this practice masks significant discord underlying the data. This underlying discord can result in researchers finding strong support for conflicting relationships with only minor modifications to the genes included between analyses. As such, the field of phylogenetics is at an interesting juncture. Big data are providing unprecedented opportunities to conduct phylogenetic analyses at a scale that encompasses entire genomes. However, such analyses face computational challenges and pose new challenges from their increased heterogeneity (i.e., larger data sets have a greater number of processes that might contribute to discordant gene trees).

We hope that our work here will draw attention to one potential avenue for potentially improving phylogenetic estimates by minimizing some model misspecification (in this case, excluding LGT trees from a set of discordant trees), while we learn something about the processes underlying the discord observed in phylogenomics, two goals that are certainly out of reach when researchers decide to concatenate. By embracing the heterogeneity in gene trees and exploring the sources of discord, we stand to gain a better understanding of how the resulting phylogenies may or may not be distorted by gene tree discord (e.g., Huang et al., 2014). Moreover, even if we do not currently have methods that can infer phylogenies under models that account for multiple discord-generating processes, identifying the processes informing the data is still useful for applications beyond a focus on species tree inference per se (e.g., how does the contribution of LGT vary across clades or whether LGT is associated with ecological shifts). In the future, understanding which model features are important to provide a realistic framework for inferring species phylogenies when these data sets contain multiple discord-generating processes will be mutually beneficial to both endeavors, and it is in this spirit of exploration that we present our results.

<H1>**ACKNOWLEDGEMENTS**

We thank the organizers of the special issue for the opportunity to contribute our work and for their patience and the anonymous referees for helpful suggestions that improved the manuscript.

<H1>**LITERATURE CITED**

Ané, C., B. Larget, D.A., Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* 24: 1575–1575.

Blischak, P., J. Chifman, A. D. Wolfe, and L. S. Kubatko. 2017. HyDe: a Python package for genome-scale hybridization detection, submitted, available on *bioRXiv*.

Boussau, B., G. J. Szollosi, L. Duret, M. Gouy, E. Tannier, and V. Daubin. 2013. Genome-scale coestimation of species and gene trees. *Genome Research* 23: 323-330.

Brown, J.M., and R.C. Thomson. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Systematic Biology* 66: 517–530.

Cox, C. J., B. Li, P. G. Foster, T. M. Embley, and P. Civan. 2014. Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. *Systematic Biology* 63: 272-279. doi:10.1093/sysbio/syt109

de Queiroz, A., and J. Gatesy (2007) The supermatrix approach to systematics. *Trends in Ecology & Evolution* 22: 34–41. doi:10.1016/j.tree.2006.10.002

Fontaine, M.C., J.B. Pease, A. Steele, R.M. Waterhouse, D.E. Neafsey, I.V. Sharakhov, X. Jiang, et al. 2015. Mosquito genomics. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* 347: 1258524–1258524.

Foster, P. G. 2004. Modeling compositional heterogeneity. *Systematic Biology* 53: 485-495. doi:10.1080/10635150490445779

Gatesy, J., and M. Springer. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Molcular Phylogenetics and Evolution* 80: 231–266. doi:10.1016/j.ympev.2014.08.013

Gori, K., T. Suchan, N. Alvarez, N. Goldman, and C. Dessimoz. 2016. Clustering genes of common evolutionary history. *Molecular Biology and Evolution* 33: 1590–1605. doi.org/10.1093/molbev/msw038

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

Huang, H., J. Sukumaran, S. A. Smith, and L. L. Knowles. 2017. Cause of gene tree discord? CLASSIPHY, a program for distinguishing incomplete lineage sorting and lateral gene transfer in phylogenetics. *Peer J*, in review. Available at https://peerj.com/preprints/3489/

Huang, H., L. Tran, and L. L. Knowles. 2014. Do estimated and actual species phylogenies match? Evaluation of African cichlid radiations. *Molecular Phylogenetics and Evolution* 78: 56–65.

Huang, W., G. Zhou, M. Marchand, J. Ash**,** D. Morris, P. Van Dooren, J. M. Brown, et al. 2016. TreeScaper: visualizing and extracting phylogenetic signal from sets of trees. *Molecular Biology and Evolution* 33: 3314–3316.

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

Jombart, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24:1403–1405.

Jombart, T., S. Devillard, and F. Balloux. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* 11:94-109.

Knowles, L. L. 2009. Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. *Systematic Biology* 58: 463–467.

Knowles, L. L. 2010. Sampling strategies for species-tree estimation. *In* L. L. Knowles and L. S. Kubatko [eds.], Estimating species trees: practical and theoretical aspects, 163–172. Wiley-Blackwell, Hoboken, NJ, USA.

Knowles, L. L., and L. S. Kubatko [eds.]. 2010. Estimating species trees: practical and theoretical aspects. Wiley-Blackwell, Hoboken, NJ, USA.

Kubatko, L. S. 2009. Identifying hybridization events in the presence of coalescence via model selection, *Systematic Biology* 58: 478-488.

Kubatko, L. S., and J. H. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology* 56: 17–24.

Lanier, H. C., H. Huang, and L. L. Knowles. 2013. How low can you go? The effects of mutation rate on the accuracy of species-tree reconstruction. *Molecular Phylogenetics and Evolution* 70: 112–119.

Liu, L., Z. Xi, and C. C. Davis. 2014. Coalescent methods are robust to the simultaneous effects of long branches and incomplete lineage sorting. *Molecular Biology and Evolution* 32: 791–805. doi:10.1093/molbev/msu331

Maddison, W. P. 1997 . Gene trees in species trees. *Systematic Biology* 46: 523–536.

Mallo, D., L. D. Martins, and D. Posada. 2016. SimPhy: Phylogenomic simulation of gene, locus, and species trees. *Systematic Biology* 65: 334–344.

McCormack, J. E., H. Huang, and L. L. Knowles. 2009. Maximum-likelihood estimates of species trees: How accuracy of phylogenetic inference depends upon the divergence history and sampling design. *Systematic Biology* 58: 501–508.

Meng, C., and L. S. Kubatko. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theoretical Population Biology* 75: 35–45.

Moore, M. J., C. D. Bell, P. S. Soltis, and D. S. Soltis. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences, USA* 104: 19363–19368. doi:10.1073/pnas.0708072104

Morgan, C. C., P. G. Foster, A. E. Webb, D. Pisani, J. O. McInerney, and M. J. O'Connell. 2013. Heterogeneous models place the root of the placental mammal phylogeny. *Molecular Biology and Evolution* 30: 2145–2156. doi:10.1093/molbev/mst117

Mirarab, S., M. S. Bayzid, B. Boussau, and T. Warnow. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* 346:1337.

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

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. C. Sanchez, and M. Müller. 2011. pROC: an open-source package for R and S plus to analyze and compare ROC curves. *BMC Bioinformatics* 12: 77.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53: 131–147.

Rambaut, A., and N. C. Grassly. 1997. SEQ-GEN: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences* 13:235–238.

Richards, E. J., J. M. Brown, A. J. Barley, R. A. Chong, and R. C. Thomson. 2017. Unexpected variation across mitochondrial gene trees and evidence for systematic error: How much gene tree variation is biological? *BioRxiv* https://doi.org/10.1101/171413.

Ruprecht, C., R. Lohaus, K. Vanneste, M. Mutwil, Z. Nikoloski, Y. Van de Peer, and S. Persson. 2017. Revisiting ancestral polyploidy in plants. *Science Advances* 3: e1603195.

Shen, X.-X., C. T. Hittinger, and A. Rokas. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology and Evolution* 1: 126.

Solís-Lemus, C., and C. Ané. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLOS Genetics* 12: e1005896.

Solís-Lemus, C., P. Bastide, and C. Ané. 2017. PhyloNetworks: a package for phylogenetic networks. *Molecular Biology and Evolution* 34: 3292–3298.

Smith, S. A., M. J. Moore, J. W. Brown, and Y. Yang. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* 15: 150.

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

Sukumaran, J., E. P. Economo, and L. L. Knowles. 2016. Machine learning biogeographic processes from biotic patterns: a new trait-dependent dispersal and diversification model with model choice by simulation-trained discriminant analysis. *Systematic Biology* 65: 525-545.

Wen, D., and L. Nakhleh. 2017. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Systematic Biology*, in press. doi.org/10.1093/sysbio/syx085.

Wickett, N. J., S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences, USA* 111: E4859–4868.

Xi, Z., R. K. Bradley, K. J. Wurdack, K. M. Wong, M. Sugumaran, K. Bomblies, J. S. Rest, and C. C. Davis. 2012. Horizontal transfer of expressed genes in a parasitic flowering plant. *BMC Genomics* 13: 227. doi:10.1186/1471-2164-13-227

Xi, Z. X., L. Liu, J. S. Rest, and C. C. Davis. 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. *Systematic Biology* 63: 919–932.

Yang, Y., M. J. Moore, S. F. Brockington, D. E. Soltis, G. K.-S. Wong, E. J. Carpenter, Y. Zhang, et al. 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution* 32: 2001–2014.

Zhang, C., H. A. Oglivie, A. J. Drummond, and T. Stadler. 2018. Bayesian inference of species networks from multilocus sequence data. M*olecular Biology and Evolution* 35: 504–517. doi:10.1093/molbev/msx307

Zhong, B., L. Liu, Z. Yan, and D. Penny. 2013. Origin of land plants using the multispecies coalescent model. *Trends in Plant Science* 18: 492–495. doi:10.1016/j.tplants.2013.04.009

FIGURE 1. (A) Schematic of the three steps to simulate lateral gene transfer (LGT), lineage sorting, and mutation, respectively, and (B) the topology of the species tree and different tree depths (i.e., shallow and deep) used to simulate test data sets. Specifically, starting with a species tree (Fig. 1B): (1) locus trees were simulated with random LGT events, (2) genealogies were simulated within the locus tree according to a coalescent process (i.e., simulating random incomplete lineage sorting [ILS] events), and (3) nucleotide data sets were simulated on the genealogies under a model of nucleotide substitution; note the subscripts identify the steps in the simulation process that were carried out for each independent locus, for example, for locus *i* to locus *j*. Either the gene genealogies or estimated gene trees from the nucleotide data sets were analyzed with CLASSIPHY to examine the impact of homoplasy on inferred data partitions (i.e., groups of ILS loci versus LGT loci).

FIGURE 2. Classification performance when species divergence is relatively recent (i.e., shallow; shown in the solid line) compared to deeper divergence times (show in the dotted line) for different contributions of lateral gene transfer (LGT) to gene tree discord, ranging from low, medium, and high relative proportions of LGT loci, as characterized by the receiver operating characteristic (ROC) curve. A classifier with no power will sit on the diagonal (i.e., essentially random guessing, 0.5, whether a locus is a incomplete lineage sorting [ILS] versus LGT locus). The area under the curve (AUC) scores (with a maximum value of 1) are also shown for comparison of the accuracy of the classifier in distinguishing ILS versus LGT loci (presented next to the solid and dashed lines) for shallow versus deep histories.

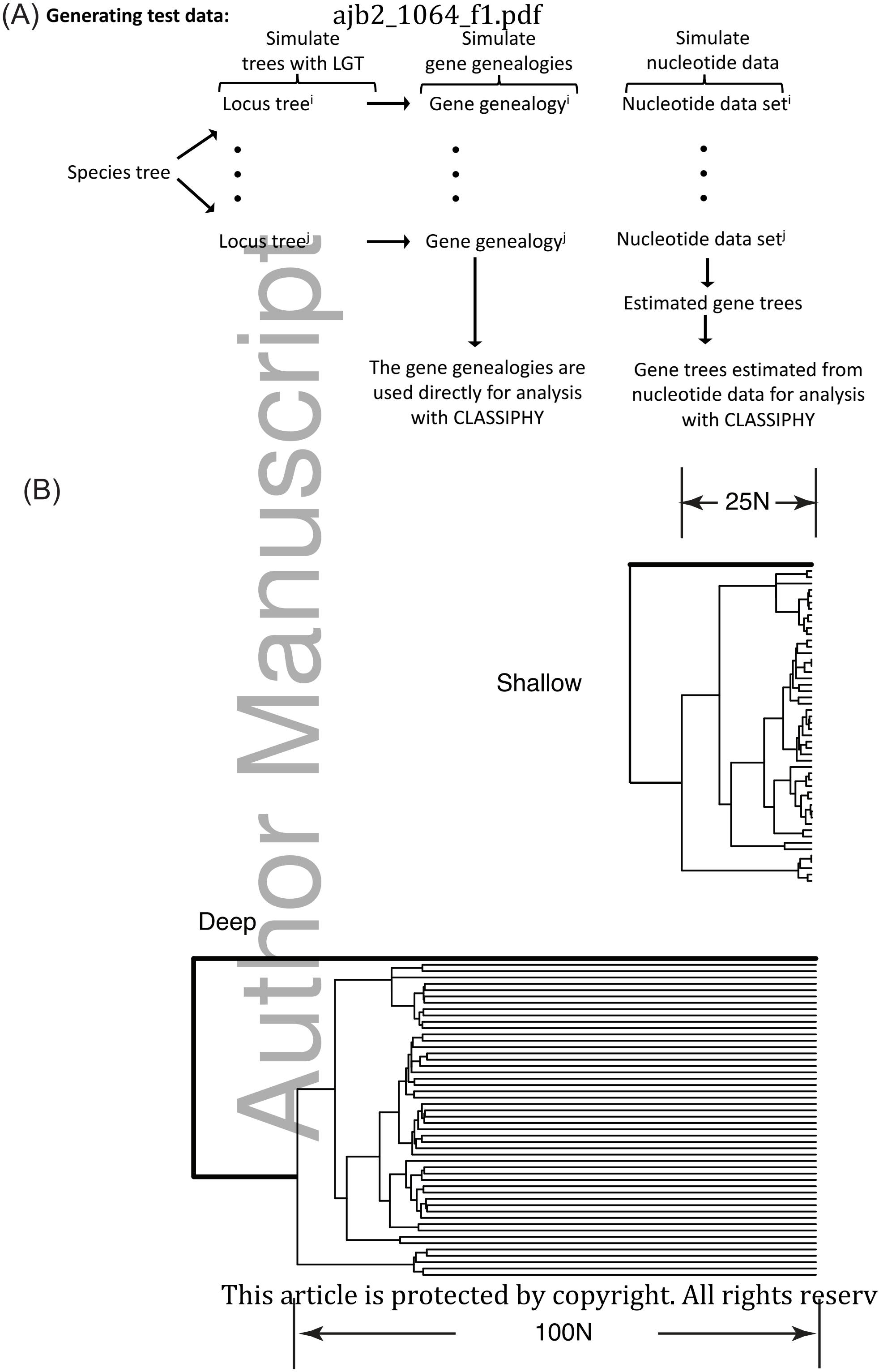
FIGURE 3. Variation in the classification performance among gene trees as a function of the Robinson-Fould’s distance between each gene tree and the species tree (i.e., each dot shows the posterior probability of lateral gene transfer [LGT]) when species divergence is relatively recent

(shown on the left) compared to deeper divergence times (show on the right), for (A) low, (B) medium, and (C) high relative proportions of LGT loci.

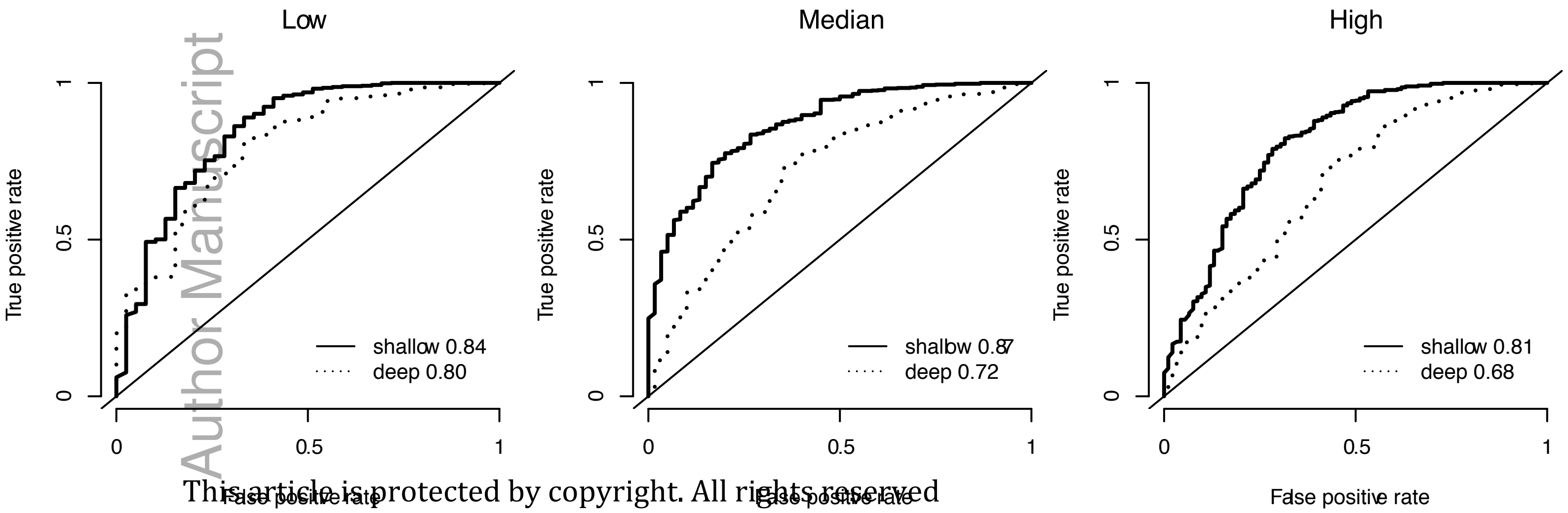
TABLE 1. Comparison between the classification accuracy of incomplete lineage sorting (ILS) and lateral gene transfer (LGT) loci based on the genealogy versus estimated gene trees highlights the impact of mutational variance (i.e., the mismatch between the actual genealogical history of a locus and the estimated gene tree of a particular locus; see Huang et al., 2009). Because the results are standardized (i.e., the difference in the percentage of correct classification of ILS and LGT loci when based on the genealogy versus the estimated gene tree), the effect of mutation separate from any inherent differences in the accuracy of classification of a locus is clear.

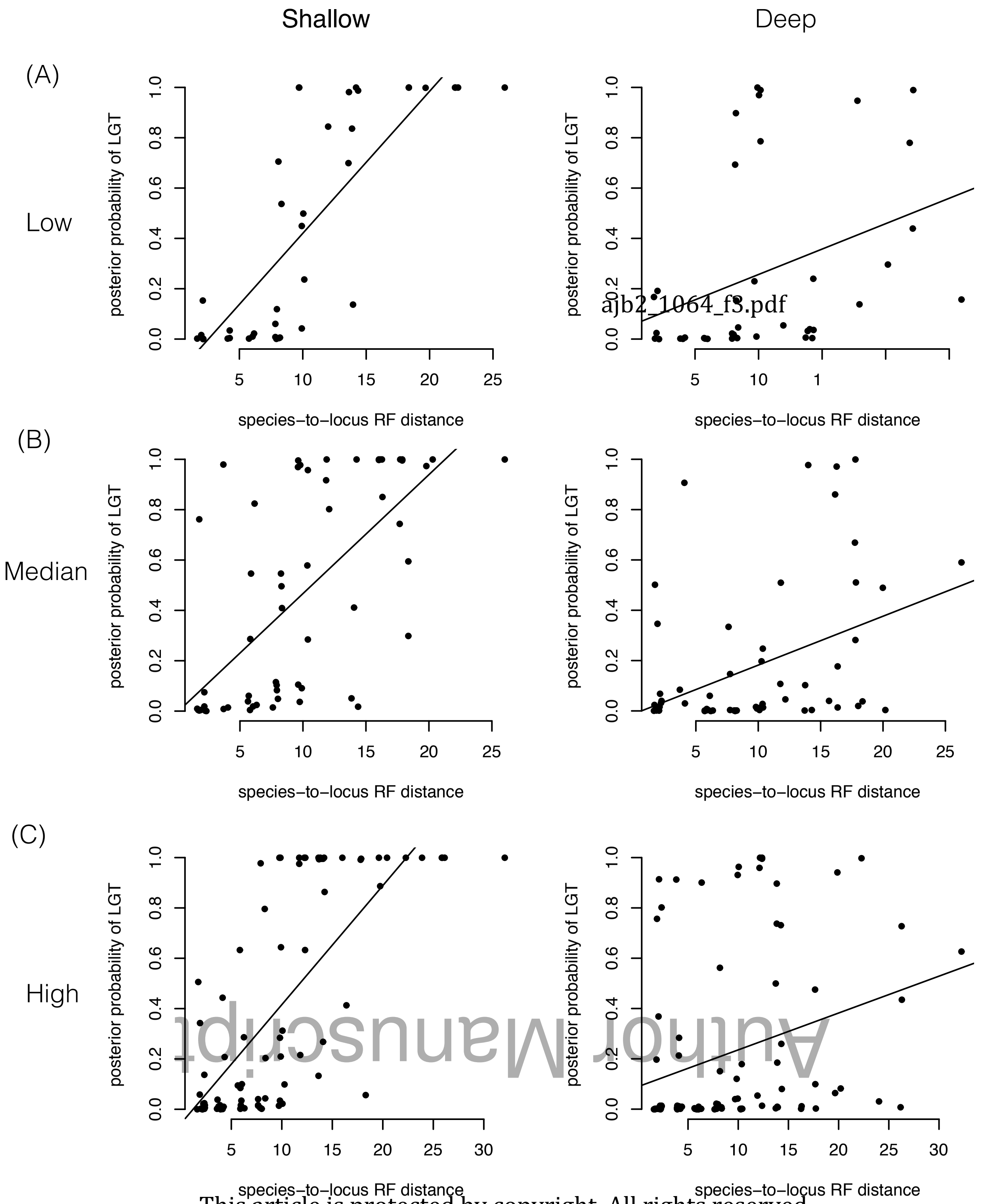
	% Decrease in classification accuracy due to mutation			
	Shallow divergence history		**Deep divergence history**	
LGT rate	**ILS**	**LGT**	**ILS**	**LGT**
Low	0	8%	0	15%
Medium	3%	0	2%	18%
High	8%	0	0	31%

ajb2_1064_f1.pdf



ajb2_1064_f2.pdf




Shallow
Deep
(A)
Low
(B)
Median
(C)
High
posterior probability of LGT
0.0
0.2
0.4
0.6
0.8
1.0
species−to−locus RF distance
5
10
15
20
25
30
1