Systematic Biology Advance Access published March 28, 2011

Point of View

DOI:10.1093/sysbio/syr025

Missing Data in Phylogenetic Analysis: Reconciling Results from Simulations and Empirical Data

JOHN J. WIENS* AND MATTHEW C. MORRILL

Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245, USA;
**Correspondence to be sent to: Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245, USA;*
E-mail: wiensj@life.bio.sunysb.edu.

Received 22 May 2010; reviews returned 1 September 2010; accepted 21 February 2011
Associate Editor: Karl Kjer

This paper will attempt to resolve some controversies about the effects of missing data on phylogenetic analysis. Whether missing data are generally problematic is a critical issue in modern phylogenetics, especially as wildly different amounts of molecular data become available for different taxa, ranging from entire genomes, to single genes, to none (e.g., fossils). Our perception of the impact of missing data (or lack thereof) may strongly influence which taxa and characters we include in a phylogenetic analysis (Wiens 2006) and may lead to a diversity of serious errors. For example, if we think that missing data are problematic when they are not, then we may exclude taxa and characters that would otherwise benefit our analyses, given the abundant evidence that increasing numbers of both taxa and characters can potentially improve the accuracy of phylogenetic analyses (e.g., Huelsenbeck 1995; Rannala et al. 1998; Poe 2003), where accuracy is generally defined as the similarity between the estimated tree and the correct, known phylogeny. In contrast, if missing data cells are themselves intrinsically problematic (e.g., Huelsenbeck 1991), including taxa or characters with many missing data cells may lead to inaccurate phylogenetic estimates.

Several studies have explored how missing data may impact phylogenetic analyses, using both empirical and simulated data. Many simulation and empirical studies now suggest that it is often possible to include taxa that have large amounts of missing data without ill effects (e.g., Wiens 2003b; Driskell et al. 2004; Philippe et al. 2004; Wiens et al. 2005; Wiens and Moen 2008; Lynch and Wagner 2010; Thomson and Shaffer 2010; Wiens, Kuczynski, Townsend, et al. 2010). However, a recent simulation study (Lemmon et al. 2009) suggested instead that missing ("ambiguous") data are generally problematic for phylogenetic analysis and implied that these previous simulation and empirical studies are therefore incorrect. They justified their study based on the grounds that previous studies were supposedly in conflict about the impacts of missing data (p. 131).

In this paper, we will show that the paper by Lemmon et al. (2009; LEA hereafter) is problematic for several reasons. First, despite their statement that previous studies are in conflict, most simulation and empirical results on missing data can be easily explained within an existing theoretical framework (Wiens 2003b). Furthermore, many contradictory studies suggesting that missing data are not generally problematic for Bayesian and likelihood analyses (given some assumptions) were not addressed by LEA. Second, the sweeping negative conclusions of LEA are not necessarily supported by their results. LEA find missing data to be problematic primarily when using sets of invariant or saturated characters and/or when obvious rate heterogeneity is ignored. Their results do not support the idea that missing data generally lead to incorrect inferences about topology when informative data are analyzed with appropriate methods. We conduct new simulations under more realistic conditions, and these results show no evidence that missing data generally lead to inaccurate Bayesian estimates of phylogeny. In fact, we show that the practice of excluding characters simply because they contain missing data cells may itself reduce accuracy. We reanalyze the "manipulated" empirical example from LEA and find that, without these artificial "manipulations" of the data, their conclusions are not supported. We also analyze eight empirical data sets, each containing many taxa with extensive missing data. We show that these incomplete taxa are consistently placed into the expected higher taxa, often with very strong support. Overall, our results confirm previous simulation and empirical studies showing that taxa with extensive missing data can be accurately placed in phylogenetic analyses and that adding characters with missing data can be beneficial (at least under some conditions). We conclude by pointing out important areas for future research on the topic of missing data and phylogenetic analysis.

A General Framework for Interpreting Simulation and Empirical Results

There are two main ways that missing data might be added to a phylogenetic analysis, either through the

addition of incomplete characters or incomplete taxa. For example, imagine having data from two genes for a given genus of organisms, in which the first gene has been sequenced for all 10 species and the second gene has been sequenced for only 5 species. Given this situation, one might (a) analyze only the first gene for all 10 species and then decide whether or not to (b) add the second gene (adding incomplete characters that are missing data for 5 of the species). Or, one might (c) analyze only the 5 species having data for both genes, and then decide whether or not to (d) add the 5 species lacking data for the second gene (adding incomplete taxa that are lacking data for the first gene). Note that (b) and (d) are effectively identical. Most of the literature on missing data has focused on whether to include taxa lacking data for some characters (c versus d). LEA did not actually address this question, but they imply that their results overturn earlier studies that did (e.g., p. 141). Below, we briefly review previous studies on incomplete taxa (treating fossil studies collectively rather than individually), and address incomplete characters (a vs. b) in our simulations and under "Areas for Future Research."

Rather than being in conflict, we argue that most of the diverse empirical and simulation studies on missing data are largely consistent when viewed in light of the hypothesis that highly incomplete taxa can potentially be accurately placed if enough informative characters are sampled overall (Wiens 2003b). Thus, the apparent impacts of extensive missing data in these studies fall along a continuum (from negative to inconsequential) based on the overall number of characters in the analysis.

The issue of missing data first became prominent in association with parsimony analyses of morphological data for fossil taxa (e.g., Donoghue et al. 1989; Platnick et al. 1991). These studies have found incomplete taxa to be problematic in some cases (e.g., generating many equally parsimonious trees and poorly resolved consensus trees; Novacek 1992; Wilkinson 1995; Anderson 2001) but not others (e.g., Kearney 2002; Cobbett et al. 2007). However, these studies included relatively few characters (up to a few hundred, but often <100). The simulations of Huelsenbeck (1991) included only 100 characters and found highly incomplete taxa (75% missing data) were often problematic. Wiens and Reeder (1995) found that including highly incomplete taxa (75%) reduced accuracy somewhat in parsimony analyses of known viral phylogenies, but their two data sets (sequence and restriction site) each included less than 100 parsimony informative characters. Dragoo and Honeycutt (1997) showed that their parsimony analyses were largely insensitive to missing data with $\sim$1500 characters (three mitochondrial genes for carnivoran mammals), with no effect on topology when one or two of the three genes were replaced with missing data cells, but that some resolution was lost when some taxa had $\sim$87% missing data. The simulations of Wiens (2003b; see also 2003a) showed that highly incomplete taxa (e.g., 90% missing data) can be accurately placed given enough characters in parsimony, likelihood, and neighbor-joining analyses, but the exact level of character sampling needed depends on the phylogenetic method, distribution of missing data among characters, and branch lengths (e.g., accurate placement is more difficult with neighbor joining and parsimony, when missing data are randomly distributed among taxa, and when overall branch lengths are long and/or characters evolve rapidly).

Dunn et al. (2003) performed a limited set of simulations based on their data for 2293 rapidly evolving mitochondrial DNA characters for mylobatiform fish and found that the impact of incomplete taxa varied depending on the method, from negative (parsimony) to none (likelihood), relative to simulations in which all taxa were complete. Philippe et al. (2004) included 30,399 characters from 129 protein sequences among eukaryotes and found that highly incomplete taxa were placed with strong support in their empirical likelihood analyses (i.e., the four most incomplete taxa had 56%, 60%, 61%, and 76% missing data, and the likelihood bootstrap values placing them with their sister taxa are respectively 100%, 92%, 98%, and 95%). They also found high accuracy in their simulations based on those data (e.g., 100% accuracy for all nodes when 50% of the data were missing, and 89% mean accuracy across nodes when 90% were missing). Driskell et al. (2004) examined DNA data sets with very large numbers of characters (469,497 for metazoans and 96,698 for green plants) and extensive missing data (92% and 84%, respectively) and found that highly incomplete taxa were placed in clades expected from previous taxonomy with strong support based on parsimony bootstrapping. Wiens et al. (2005) included 3519 (mostly DNA) characters for treefrogs and showed that highly incomplete taxa were placed in the expected clades with very strong support by parsimony and Bayesian analyses (e.g., 10 species with $>90\%$ missing data each were all placed in the expected clades, with monophyly of these clades each supported with a Bayesian posterior probability (Pp) of 1.00). Other recent empirical studies (described below) have also shown that highly incomplete taxa are placed in the expected clades with strong support, and most of these studies included >4000 characters each (e.g., Lynch and Wagner 2010; Thomson and Shaffer 2010; Wiens, Kuczynski, Townsend, et al. 2010).

Wiens (2005) used simulations to show that adding highly incomplete taxa (i.e., 90% missing data) could be as effective as complete taxa in rescuing likelihood and Bayesian analyses from long-branch attraction, even when the models utilized in these analyses were misspecified (i.e., among-site rate heterogeneity and transition–transversion bias were simulated but not included in the estimation models), given a sample of 1000 characters. The simulations of amino acid data by Hartmann and Vision (2008) showed reduced accuracy with extensive missing data for parsimony, likelihood, and neighbor-joining analyses, but only included 500 characters. Wiens and Moen (2008; their fig. 2) used simulations to show that highly incomplete taxa could be accurately placed in Bayesian analyses given enough

characters (e.g., 2000), even when rate heterogeneity and substitution bias were simulated but not included in the Bayesian model.

In summary, all of these simulation and empirical studies seem to fit into this common framework, with highly incomplete taxa being potentially problematic when the overall number of characters is small and generally unproblematic when the number is large. This common framework seems to apply to all phylogenetic methods, not simply likelihood and Bayesian analysis.

The results of many of these studies contradict the conclusions of LEA but were not mentioned by them, including the ones that addressed the impact of missing data on likelihood and Bayesian analyses (e.g., the results of Philippe et al. 2004; Wiens 2005; Wiens et al. 2005 were not mentioned, and the latter two studies were not cited). For example, LEA conclude that "in both ML and Bayesian frameworks, among-site rate variation can interact with ambiguous data to produce misleading estimates of topology" (p. 130) and that estimation becomes problematic "when rate variation across sites is not properly modeled" (p. 141). But the simulation studies by Wiens (2005) and Wiens and Moen (2008) showed accurate estimation of topologies with incomplete taxa by Bayesian and likelihood methods when rate variation was simulated but completely ignored.

Problematic Simulations and Conclusions of Lemmon et al. (2009)

LEA analyzed a very limited set of simulated conditions and found results that were seemingly discordant with those of other simulation studies of the same topic. Yet, they make sweeping conclusions from their results (e.g., that their results have implications for "all analyses that rely on accurate estimates of topology or branch lengths", p. 130). They also imply that their results overturn those of previous studies. It is therefore important to look at what they did and found more closely.

LEA examined the four-taxon case, with the simplest model of sequence evolution (Jukes–Cantor), and equal branch lengths (given that an unrooted tree is estimated). For each set of conditions, they simulated two data sets (Gene A and Gene B), one with complete data for all taxa and characters, and another lacking data for all characters for two taxa. These genes were simulated under either the same or different rates of change. They then evaluated Bayesian Pp for the single internal node for Gene A alone and for Gene A and B combined. For maximum likelihood, they evaluated the frequency with which this clade was correctly reconstructed. They assumed that the combined data would give the same results as Gene A alone because data were only present in two of the four taxa for Gene B (making Gene B uninformative under the parsimony criterion, but note that this is not necessarily true for likelihood or Bayesian analysis, see below).

They found that Bayesian Pp for the combined data sometimes differed from the observed values based on Gene A alone (but for maximum likelihood a comparable result only occurred when branch lengths were arbitrarily fixed to nonzero values). They refer to these differences as "bias." In some cases, these biases appear to be problematic, as when Pp approaches zero for the true tree (such that the true phylogeny is not estimated). Similarly, they found some cases where Pp was very high for the true tree, even though both data sets were effectively invariable. They suggested that these biases were related to the Bayesian star-tree paradox (e.g., Lewis et al. 2005), the tendency for Bayesian analysis to strongly favor one tree when there is little information with which to choose among trees (regardless of missing data). However, they found virtually none of these extreme biases unless the characters were effectively invariant or "saturated" (i.e., used by LEA as meaning so variable so as to be effectively uninformative), and unless rate heterogeneity between genes was simulated and then ignored by failing to partition by genes. The only exception we find is in their fig. 4, for one set of conditions with very high rates in both genes and data missing in sister taxa. Thus, their results do not support their sweeping generalizations about the negative impact of missing data, especially for conditions likely to be encountered by most empirical systematists. This presumably explains why so many previous simulation and empirical studies contradict their conclusions about the negative impact of missing data on Bayesian and likelihood analyses (see above).

In some cases, they show that combined data Pp differ moderately from those for Gene A alone (e.g., their fig. 4, when rate of Gene A is low). However, arguing that these Pp are "biased" assumes that Gene B has no influence on topology estimation whatsoever (i.e., Pp for the combined analysis should be the same as for Gene A alone). Although this is true for parsimony, it is not necessarily true for Bayesian or likelihood analyses. For example, if Gene B has no influence on Bayesian estimates of topology (which are based on Pp), then how can it influence Pp at all? Clearly, the initial assumption that incomplete characters in Gene B have no impact cannot be fully correct. Furthermore, finding that combined data Pp differ from Pp for Gene A is not direct evidence that combined data Pp yield biased estimates of accuracy (i.e., in this context, the probability that the clade is correctly reconstructed by the method under a given set of conditions). Demonstrating bias would require directly examining the relationship between accuracy (the probability that the clade is correctly reconstructed by Bayesian analysis of Genes A and B combined) and the combined-data Pp for these conditions (for studies examining the relationship between Bayesian Pp and accuracy see, e.g., Wilcox et al. 2002; Alfaro et al. 2003; Huelsenbeck and Rannala 2004). LEA did not directly examine the relationship between accuracy and Pp when missing data are added, and so for these nonextreme conditions, their statements about "bias" caused by missing data are not actually based on any direct evidence.

In summary, the results of LEA suggest that missing data are primarily problematic when utilizing

uninformative characters and/or when failing to partition clearly heterogeneous data sets, conditions that may not be routinely encountered by most systematists. This is not to say that we think that data sets with missing data always yield accurate phylogenies with unbiased support values, but rather that the simulation results of LEA can have a very different interpretation from their sweeping negative conclusions, if one simply considers which of their results are relevant to what phylogeneticists actually do.

Another critical issue is the addition of sets of characters with data for only two species, which are expected to have little impact on the analysis (and which presumably would not be used by empirical systematists). It is unclear if their results are specific to adding only two species or if they also apply to larger numbers of species. We have therefore performed new simulations to address the relevance of the results of LEA to larger numbers of taxa.

New Simulations

Methods

We addressed how adding data from a gene with incomplete taxon sampling to the one with complete taxon sampling influences the accuracy of Bayesian phylogenetics. Simulation methods generally followed Wiens and Moen (2008). We simulated 16-taxon phylogenies that were either fully asymmetric or symmetric. Following LEA, we simulated DNA data with the Jukes–Cantor model with equal branch lengths across the tree. We generally used 500 characters per gene, but also simulated 100. Data set 1 had all characters for all 16 taxa. For Data set 2, we simulated the complete data, and then a set of taxa was randomly selected in each replicate to have all their characters replaced with missing data cells. In one set of simulations, eight taxa were incomplete in Data set 2 (one way of generalizing the design of LEA to larger numbers of taxa). In another set of simulations, 14 taxa were incomplete, leaving only 2 taxa in Data set 2 with nonmissing data (an alternate way of generalizing the design of LEA to more taxa). Data were simulated under a broad range of rates of change in each data set, from very low (probability of change in a given character along a given branch of 0.0001) to relatively high (0.50), and six intermediate rates (0.001, 0.01, 0.10, 0.20, 0.30, 0.40). Initial analyses on the asymmetric tree showed that the most extreme rates gave relatively low accuracy for these conditions (30% of tree or less resolved correctly). We simulated both equal rates in each data set and many unequal rates (Fig. 1), but not every possible combination of rates. We analyzed 100 replicates for each set of conditions. We analyzed Data set 1 alone and then analyzed data sets 1 and 2 combined. Data sets were analyzed using MrBayes v3.1.2 (Huelsenbeck and Ronquist 2001), assuming a Jukes–Cantor model with a parameter for unequal rates of change among sites (gamma), and other options set to default values. Importantly, combined analyses were partitioned, allowing a different value for gamma in each data set. Analyses were run for 50,000 generations each, sampling every 100 generations, and excluding the first 10,000 generations as burn-in. These settings provide adequate tree searches for these conditions (Wiens and Moen 2008). We evaluated accuracy for each replicate as the proportion of resolved nodes in the majority-rule Bayesian tree that are shared with the known, true topology, and overall accuracy (for a given set of conditions) is based on the mean for 100 replicates. This measure of accuracy is used in many previous simulation studies (e.g., Wiens 2003b; Wiens and Moen 2008); other measures are certainly possible, but they should also reflect the similarity between the true and estimated trees averaged across replicates. We did not directly evaluate Pp support for individual clades, but a clade will not be resolved unless its Pp is >0.50, and LEA did not directly examine the relationship between accuracy and Pp either.

Results

We find that across a broad range of conditions (Fig. 1), adding the data set consisting of 50% missing data (8 of 16 taxa incomplete) either increases or has no effect on accuracy, relative to analyzing the complete data set alone. Although the increases are typically small, under some conditions, the relative increase can be $>20\%$ (e.g., 0.49 vs. 60). These increases in accuracy may occur when the rates of change in the two data sets are equal, or when they are very unequal as well. When the added data set has only two complete taxa (as in the simulations of LEA), accuracy may be slightly higher or slightly lower than Data set 1 alone, and is consistently within 0.05. These latter results suggest that adding sets of characters with only two species has little influence on the overall accuracy of analyses with larger numbers of taxa, and that the design and results of LEA do not generalize to more realistic conditions.

Discussion

Contrary to the conclusions of LEA, we find no evidence that adding sets of characters with extensive missing data leads to misleading estimates of Bayesian phylogeny or support values (i.e., only clades with Pp > 0.50 are supported). Importantly, our results suggest that under some conditions, failing to add characters with missing data can lead to reduced phylogenetic accuracy. Thus, being overly cautious about excluding characters simply because they have missing data can lead to reduced phylogenetic accuracy. This is a critical point that LEA do not discuss.

Our simulation methods are not identical to those of LEA. For example, we assume that researchers will not choose to analyze data sets that completely lack phylogenetic information due to rates that are too fast or slow (and so we do not simulate these conditions, but we do simulate branches that are both extremely long and

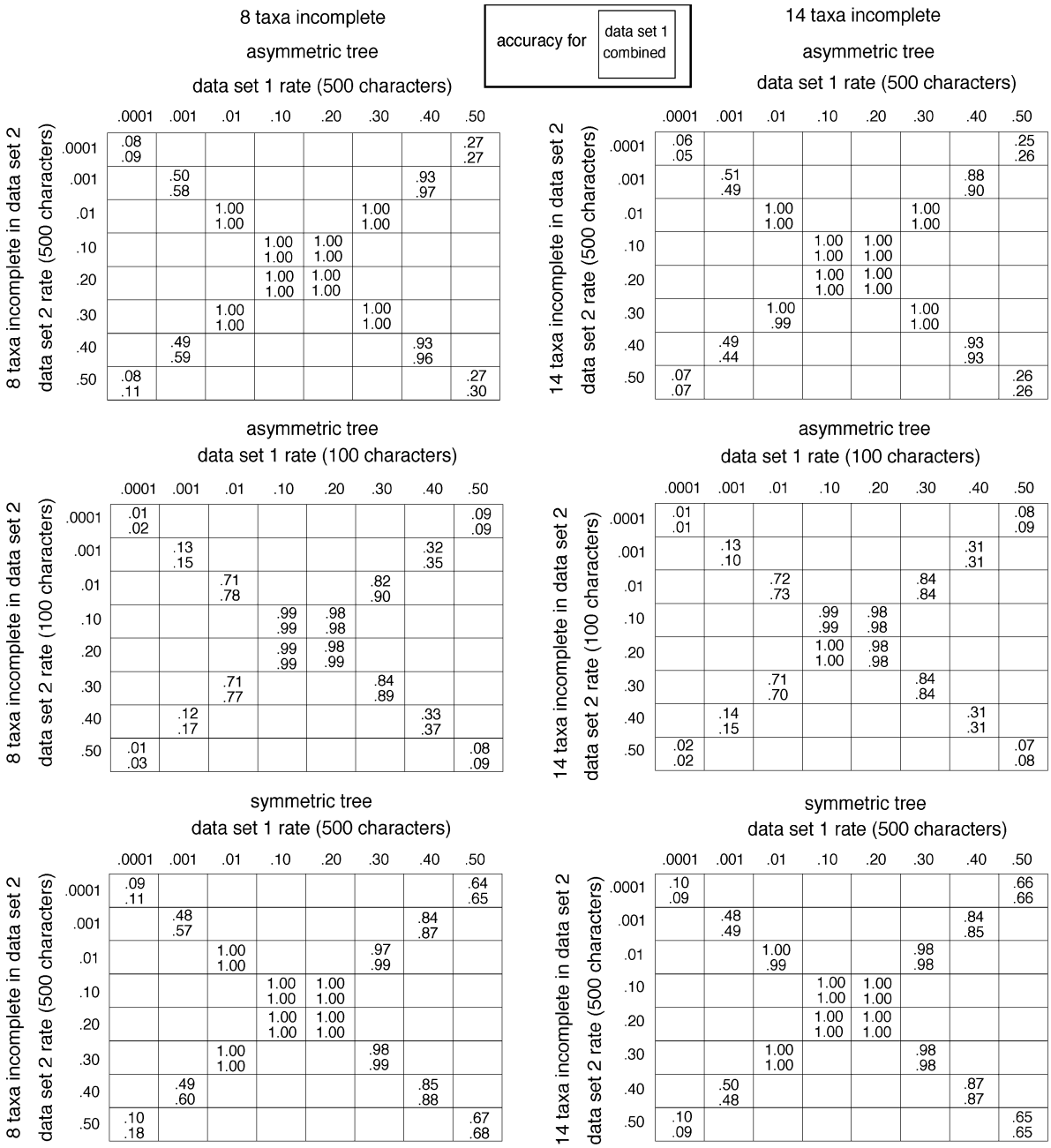

FIGURE 1. Results of simulations showing the impact of adding sets of characters with missing data on the accuracy of Bayesian phylogenetic analysis. Within each square, the top number is accuracy for Data set 1 alone, whereas the bottom number is for combined analysis of Data sets 1 and 2. Results on the left are for simulations in which Data set 2 has eight complete taxa and the right shows results in which Data set 2 has only two complete taxa.

extremely short). We also assume that most researchers will partition data sets evolving at different rates. But most importantly, our results suggest that the misleading Bayesian estimates noted by LEA do not necessarily occur under slightly more realistic conditions (e.g., more taxa, partitioned data, and use of variable characters). As one example, LEA suggest that Bayesian Pp may be strongly influenced by whether the taxa with nonmissing data are sister or nonsister taxa, but this simple division becomes unclear when additional taxa are included. For example, given a five-taxon tree (A, B) (C (D, E), with missing data in species C and D, nonmissing data are simultaneously present in both sister taxa (A, B) and nonsister taxa (A, E).

These simulations are also very limited and still very far from realistic. Many parameters that could have been varied were not (e.g., more complex substitution models, variation in rates within genes), in order to make the

results more comparable with those of LEA (see instead Wiens 2005; Wiens and Moen 2008). Perhaps the most important oversimplification is the use of equal branch lengths throughout the tree. In order to address how missing data influence Bayesian and likelihood analyses under fully realistic conditions, we also perform analyses of eight empirical data sets. Before we do that, we briefly address the empirical example offered by LEA.

Reexamining the Manipulated Empirical Example of LEA

LEA analyzed an empirical data set but again made many methodological choices that make these data very different from those analyzed by empirical systematists. They analyzed data from a single mitochondrial gene from eight species of plethodontid salamanders (but for which entire mitochondrial genomes were available; Mueller et al. 2004), deliberately analyzing a very small number of characters. We could not find an explanation for why these particular characters and taxa were chosen. They then added a second data set consisting of missing data for six of the eight species and "manipulated empirical data" for the other two. These added, nonmissing data consist of resampled sites from the same gene for the same species, selected to be either all invariant or all variable between the two species. Although they found that adding this second set of characters influenced Bayesian and likelihood estimates of topology, support, and branch lengths, this analysis raises many questions about its design. Why not use actual data (e.g., another gene) instead of resampling sites from the same gene? Why only variable and invariant sites? To what extent are their results an artifact of these methodological choices?

We addressed this latter question using very similar empirical data, and our results offer a dramatic contrast to those of LEA (Fig. 2). We downloaded the same 16S data, but instead of adding only invariant or variant sets of characters from the same gene, we added unmanipulated data from another gene (the widely used cytochrome *b*; again from Mueller et al. 2004) to the same species to which LEA added data. Clearly, adding another gene is more relevant to what empirical systematists actually do. Instead of finding that "ambiguous characters can strongly bias estimates of topological support and branch lengths" (p. 139) we find that Bayesian and likelihood estimates of topology, support, and branch lengths are almost identical after adding cytochrome *b* with data missing in six species (see Fig. 2 legend for methods). As in their simulations, it appears that the results of LEA reflect artifacts of adding invariant and saturated characters (and failing to partition data sets), and therefore may have limited relevance to most empirical studies.

Apart from their example involving "manipulated" empirical data, LEA do not show any empirical studies in which missing data seem to be problematic for Bayesian or likelihood methods. Note that on p. 142, LEA state "One of us (K.S.-H.) has come across such an example of discordance among gene trees in empirical data from North American fireflies. Once ambiguous sites were excluded from the analysis, gene tree congruence increased substantially (Stanger-Hall et al. 2007)." However, the only references to missing data in that paper were the following quotes: "However, due to stretches of missing data in individual taxa (due to differences in primer binding and sequencing success) and the possibility that these unduly influence the phylogenetic analysis (Lemmon et al. unpublished data), the final alignment was reduced to 1906 bp." (p. 36) and also (p. 42) "it seems to have a significant effect on the outcome of a ML and/or Bayesian analysis (Lemmon et al., unpublished data). This led us to exclude DNA segments with missing data for more than one taxon from our final alignment." Thus, the Stanger-Hall et al. (2007) paper does not contain the empirical results that LEA state that it does, only references to LEA.

New Results from Empirical Data Sets

The problem of missing data is something that empirical phylogeneticists may encounter every day. LEA state that the supposed negative impacts of missing data on phylogenetic analysis are relevant to "all studies" that estimate and use phylogenies (p. 130). If this were true, we would expect to see widespread negative impacts in empirical analyses that include extensive missing data. We have previously described empirical studies that showed evidence that such impacts can be small or nonexistent (e.g., Philippe et al. 2004; Wiens et al. 2005), specifically for likelihood and Bayesian analyses. Here we present analyses of eight additional empirical data sets that show similar patterns.

Obviously, the true phylogeny is unknown in most empirical data sets. However, one can make predictions about how methods will perform with real data given the results of simulations. It is not immediately clear what specific empirical predictions can be derived from the simulations of LEA. Nevertheless, they state that extensive missing data may "positively mislead" (p. 143) estimated topologies in likelihood and Bayesian analyses. If this were the case, then we predict that highly incomplete taxa will be placed in clades that appear to be incorrect based on previous taxonomy and systematic research (i.e., assessing accuracy based on congruence). In contrast, if the hypothesis of Wiens (2003b) is correct, and if sufficient characters have been sampled, then we expect that incomplete taxa will be placed in the expected higher taxa (e.g., genera, families), and with strong support. In addition, there should be strong support for the localized placement of these species within these higher level taxa (if sufficient characters were sampled). Following Wiens et al. (2005), we test for a negative relationship between localized clade support and incompleteness of individual taxa.

Methods

We selected eight published empirical data sets (Table 1), all involving Bayesian or likelihood analyses

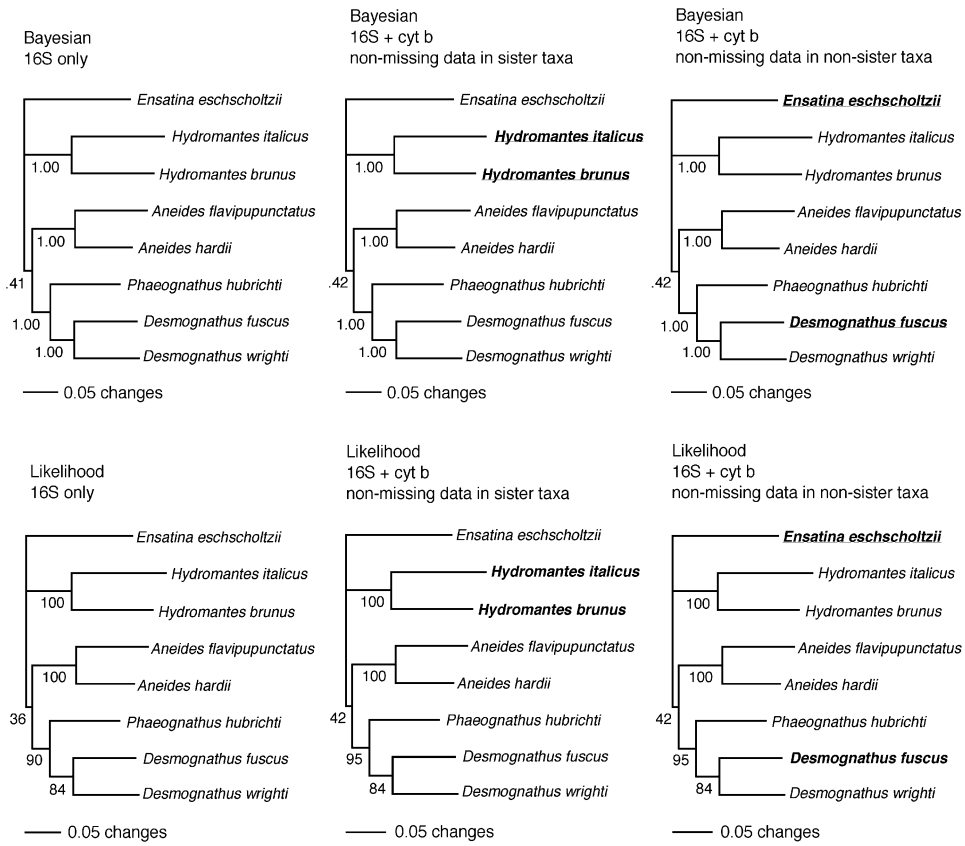

FIGURE 2. Analyses of mitochondrial DNA sequence data in plethodontid salamanders show that adding sets of characters with extensive missing data may have negligible impacts on topology, support, and branch lengths (contrast with fig. 7 of LEA). We obtained the same 16S data for the same species as LEA (aligned using MUSCLE; Edgar 2004), but instead of adding resampled invariant and variable sets of characters from 16S, we added data from another gene (cytochrome *b*; cyt b). Data are added for the same pairs of sister and nonsister species (boldfaced) used by LEA; all other species have missing data cells for cyt b. Bayesian analyses used MrBayes v3.1.2 (Huelsenbeck and Ronquist 2001) with the GTR + I + Γ model and 1,000,000 generations. Likelihood analyses used RAxML version 7.2 (Stamatakis 2006) with the recommended GTR + Γ model, with 100 integrated bootstrapping and heuristic search replicates. Both analyses were partitioned by gene. Numbers adjacent to branches indicate posterior probabilities (Bayesian) or bootstrap values (likelihood). Trees are unrooted (but see Kozak et al. 2009 for justification for rooting near *Ensatina* and *Hydromantes*).

of mostly nonoverlapping vertebrate clades. All eight have at least one species with >70% missing data, and four data sets have at least one species with >90% missing data. Given that all eight studies are from our laboratory, we know that taxa were not simply excluded because of incompleteness, or because of how incompleteness influenced the results (this is less clear for other studies). All eight studies include multiple genes, and six of eight include both nuclear and mitochondrial genes. Thus, there is considerable rate heterogeneity among genes and other data partitions.

For each data set, we quantified the percentage of missing data cells in each species. Most missing data originate from the complete absence of data from one or more genes (or parts of genes) in combined analyses, but a small fraction also comes from gaps in alignments. Levels of branch support were based on whichever model-based method was used in the original study (i.e., Bayesian vs. likelihood); we arbitrarily selected likelihood for Hua et al. (2009), which used both. For the ranid and phrynosomatid data sets, we used 49% as the lowest bootstrap, as specific values <50% were unavailable; however, relatively few nodes had values <50% (14.6% for ranids, 7.3% for phrynosomatids) and using reasonable alternate values (e.g., 25%) gave identical correlation results. Detailed methods are described in the original studies. However, given that effects of missing data may depend on how a software package treats these cells, we note that maximum likelihood analyses used RAxML (Stamatakis 2006) and Bayesian analyses used MrBayes (Huelsenbeck and Ronquist 2001).

We first evaluated whether highly incomplete taxa are placed in the clades expected based on previous taxonomy, and whether they are placed in these clades with strong support. If highly incomplete taxa are generally problematic, then they should not be consistently placed in the clades predicted by previous taxonomy, or if they are, the support for these clades should be weak. For each data set, we identified a set of nonnested clades from previous taxonomy. These mostly consisted of genera, as the generic-level assignment of most of these

species was previously established based on nonmolecular data. However, for higher-level snake phylogeny, with only one species per sampled genus, we used families (and well-established subfamilies for Colubridae). For ranids, we used subfamilies, given that the taxonomy for these clades is relatively stable (e.g., Bossuyt et al. 2006; Frost et al. 2006) whereas generic-level taxonomy is not (e.g., Frost et al. 2006 vs. Wiens et al. 2009). We then tallied the support for each clade (likelihood bootstrap or Bayesian Pp) and the species with the maximum amount of missing data in that clade (i.e., the taxon that should be most difficult to accurately place). We acknowledge the possibility that these higher taxa may be associated with longer branches than a random sampling of internal branches within the tree, but this potential source of bias should not overturn our results or conclusions.

Following Wiens et al. (2005), we then quantified the level of branch support for the specific placement of each individual species based on either 1) the support for the node placing them with their sister species (for species that are sister to a single species), or 2) the average of the support for the clade uniting them with their sister group, and the support for the clade excluding the species from that sister group (for species that are sister to >1 species). Given the level of incompleteness and branch support for each species, we then examined the relationship between these variables using nonparametric Spearman's rank correlation analysis, implemented in Statview.

Note that a negative relationship between support and completeness is not inconsistent with the mechanism proposed by Wiens (2003b). If there are too little data to accurately place a taxon on the tree, then the support for its placement should be weak. However, the simulations of Wiens (2003b) suggest that, given enough characters, even highly incomplete taxa will be accurately placed in the phylogeny with high consistency. This should be reflected with high support values. Finally, we note that this analysis does not necessarily address whether support values are biased by missing data, unless they are strongly biased to be consistently positive or negative (but LEA do not address moderate biases either because they did not directly test how accuracy and support values are related).

Results

The eight data sets collectively include >1000 species and >60 higher taxa, and almost all of these species are placed in the expected higher taxa, despite many having extensive missing data (Tables 1 and 2). Furthermore, the monophyly of most of these clades is strongly supported (Bayesian Pp = 1.00; likelihood bootstrap = 78%–100%, but most >90%). In the three cases in which genera are not monophyletic in these data sets, there are other factors besides missing data that are involved. In bolitoglossine salamanders, *Pseudoeurycea* and *Lineotriton* both appear to be nonmonophyletic (Wiens, Parra-Olea, et al. 2007), but previous phylogenetic studies with little missing data suggest that this reflects parallel evolution and misleading taxonomy (Parra-Olea and Wake 2001). The nonmonophyly of *Trachemys* seems to reflect conflict between mitochondrial and nuclear genes, not missing data per se (Wiens, Kuczynski, Arif, et al. 2010). In summary, if missing data are generally problematic as LEA suggest, there does not seem to be any evidence for it in these eight data sets (unless the previous nonmolecular taxonomies in these groups have been misled in a way that is consistent with the misleading effects of missing data on likelihood and Bayesian analyses of DNA sequence data).

Only two of the eight studies show significant negative relationships between branch support and incompleteness (Fig. 3). These results suggest that missing data have little consistent negative impact on levels of branch support, and there is sometimes strong support for the localized phylogenetic placement of taxa with >90% missing data (Fig. 3), within these expected higher taxa. Interestingly, the two data sets with significant relationships between support and completeness (plethodontids, ranids) have the largest numbers of taxa but only modest numbers of characters (Table 1). Again, we note that when too few informative characters have been sampled in a taxon, we expect only weak support for its placement in the tree.

Other Studies

In addition to these eight studies and others mentioned previously (e.g., Driskell et al. 2004; Philippe et al. 2004; Wiens et al. 2005), other recent studies have also shown similar patterns (e.g., Lynch and Wagner 2010; Thomson and Shaffer 2010; Wiens, Kuczynski, Townsend, et al. 2010; Pyron et al. 2011). For example, Lynch and Wagner (2010) examined boid snake relationships with a Bayesian analysis of 14,417 molecular characters, with some taxa 98% incomplete and each taxon having an average of 70% missing data. Yet, their phylogeny is generally strongly supported and congruent with previous hypotheses and taxonomy (e.g., of six genera with >1 species, five are strongly supported as monophyletic with Pp > 0.98). Wiens, Kuczynski, Townsend, et al. (2010) showed that addition of >15,000 molecular characters to a data set of 363 morphological characters for squamate reptiles did not change the placement of most fossil taxa in a combined Bayesian analysis (despite the fossils having >98% missing data in this analysis) and caused no significant change in Bayesian Pp for fossil taxa. Furthermore, the placement of fossil taxa was consistent with previous taxonomy (e.g., fossil snakes placed in snakes), both before and after addition of molecular data.

AREAS FOR FUTURE RESEARCH

There are now many studies showing concordant support for the idea that highly incomplete taxa can be

TABLE 1. Basic information on the eight data sets used in analyses of incomplete taxa and clade support

Clade	Character data	Number characters	Number taxa
Plethodontid salamanders	3 mitochondrial, 3 nuclear genes	5590	182
Bolitoglossine salmanders	2 mitochondrial genes	1823	157
Treefrogs (*Hyla*)	4 mitochondrial, 6 nuclear genes	7083	35
Hemiphractid frogs	2 mitochondrial, 2 nuclear genes	4370	53
Ranid frogs	1 mitochondrial, 3 nuclear genes	5307	389
Emydid turtles	2 mitochondrial, 6 nuclear genes	5264	38
Phrynosomatid lizards	5 mitochondrial, 6 nuclear genes	8582	122
Snakes	20 nuclear genes	13,332	50

Clade	Phylogenetic method	Range missing data (%) per species	Mean missing data (%) per species	References
Plethodontid salamanders	ML	1.6–93.5	43.9	Kozak et al. (2009)
Bolitoglossine salmanders	BA	1.6–75.9	36.4	Wiens, Parra-Olea, et al. (2007)
Treefrogs (*Hyla*)	ML	4–96	39.7	Hua et al. (2009)
Hemiphractid frogs	BA	1.2–77.4	19.9	Wiens, Kuczynski, Duellman, et al. (2007b)
Ranid frogs	ML	1.1–90.8	52.5	Wiens et al. (2009)
Emydid turtles	BA	2.7–72.6	20.3	Wiens, Kuczynski, et al. (2010)
Phrynosomatid lizards	ML	1.6–92.3	56.2	Wiens, Kuczynski, Arif, et al. (2010)
Snakes	ML	2.5–72.4	17.5	Wiens et al. (2008)

Notes: BA = Bayesian analysis; ML = maximum likelihood.

accurately placed in model-based analyses, and sweeping statements about the negative impacts of missing data are not substantiated. Nevertheless, many other aspects of the potential impact of missing data on phylogenetic analysis are still in need of further research.

Adding Characters with Missing Data

In addition to the effects of incomplete taxa, another major question is: given a complete set of characters for a set of taxa, is it useful to add a second set of characters that are incomplete (because they include data for only some of the taxa)? In other words, when do the benefits of adding more characters outweigh the potential disadvantages of increasing missing data in the matrix? Superficially, it might seem that the simulations of LEA addressed this question. However, their results may be of limited relevance to empirical studies because only two species were added. Our simulations here (Fig. 1) suggest that adding a set of characters with data for 50% of the species is generally either beneficial or harmless for Bayesian analysis. However, these simulations were not comprehensive either, and additional analyses are needed (e.g., exploring unequal branch lengths, different numbers of characters, and different levels of taxon sampling).

Other simulation and empirical studies have also found results suggesting that incomplete characters can be beneficial, but with some caveats. Wiens (1998) found that adding sets of incomplete characters can potentially increase accuracy for parsimony, but that accuracy was increased more by distributing the same amount of added data among fewer taxa and more characters (and with less missing data). He also found potential problems of long-branch attraction when a set of highly incomplete characters is added.

Wiens et al. (2005) showed that adding a set of slow-evolving characters (nuclear genes) available for only some taxa (and with much missing data) seemed to improve results relative to those from analyzing only fast-evolving characters (mitochondrial genes) for a larger number of taxa. Specifically, some taxa are apparently misplaced in the analysis of fast-evolving characters alone (based on previous taxonomy), but not in the combined analysis.

The simulations of Gouveia-Oliveira et al. (2007) showed that accuracy of likelihood analyses was much higher when sequences with gaps (i.e., missing data) are included rather than excluded. Similarly, Wiens (2009) used simulations to address whether adding molecular data improves phylogenetic accuracy for fossil taxa, in a combined analysis of molecular and morphological data, with parsimony and Bayesian analysis (where the molecular data are missing in the fossil taxa). These simulations showed that under many conditions, adding molecular data improved accuracy for fossil taxa. A review of empirical studies (Wiens 2009) showed that adding molecular data can improve resolution (i.e., resolve polytomies in consensus trees) for the placement of fossil taxa, at least in some parsimony analyses (e.g. Manos et al. 2007). An analysis of squamate reptiles (Wiens, Kuczynski, Townsend, et al. 2010) confirmed that molecular data can change the placement of some fossil taxa, in addition to increasing resolution.

Estimating Divergence Times

It would also be worthwhile to investigate the effects of missing data on estimation of divergence times. LEA state that their results on branch length estimation are relevant to this issue, but they acknowledge that their results may be an artifact of not including rate heterogeneity in the likelihood model (p. 139), and this latter hypothesis is supported by our analyses also (Fig. 2). Furthermore, their study contains no actual estimation of divergence dates. We have conducted several

TABLE 2. Summary of support for previously recognized higher taxa (genera, families, subfamilies) within the eight data sets, showing that almost all higher taxa are strongly supported as monophyletic, despite many of them containing one or more taxa with extensive missing data

Clade	Higher taxon	Method	Support	Maximum incompleteness (%)
Plethodontid salamanders	*Batrachoseps*	ML	84	93.5
	Gyrinophilus	ML	100	49.2
	Pseudotriton	ML	78	47.1
	Eurycea	ML	100	80.5
	Plethodon	ML	100	80.2
	Hydromantes	ML	100	55.0
	Ensatina	ML	100	50.3
	Aneides	ML	100	41.6
	Desmognathus	ML	100	80.4
Bolitoglossine salamanders	*Cryptotriton*	BA	1.00	75.6
	Dendrotriton	BA	1.00	46.4
	Nototriton	BA	1.00	75.6
	Oedipina	BA	1.00	48.7
	Thorius	BA	1.00	4.7
	Chiropterotriton	BA	1.00	71.9
	Pseudoeurycea	BA	Not supported	75.9
	Ixalotriton (nested inside *Pseudoeurycea)*	BA	1.00	49.7
	Lineotriton (nested inside *Pseudoeurycea)*	BA	Not supported	4.8
	Bolitoglossa	BA	1.00	70.8
Treefrogs (*Hyla*)	*Tlalocohyla*	ML	100	52
	Isthmohyla	ML	100	56
	Smilisca	ML	100	22
	Hyla	ML	79	96
Hemiphractid frogs	*Flectonotus*	BA	1.00	14.8
	Hemiphractus	BA	1.00	68.7
	Stefania	BA	1.00	56.9
	Gastrotheca	BA	1.00	77.4
Ranid frogs	*Ptychadeninae*	ML	96	90.8
	Phrynobatrachinae	ML	94	54.0
	Conrauinae	ML	100	44.8
	Petropetidinae	ML	100	54.2
	Pyxicephalinae	ML	86	89.6
	Micrixalinae	ML	100	55.2
	Dicroglossinae	ML	96	89.8
	Ranixalinae	ML	88	87.1
	Ceratobatrachinae	ML	100	46.9
	Nyctibatrachinae	ML	85	26.7
	Mantellinae	ML	99	52.2
	Rhacophorinae	ML	94	90.3
	Raninae	ML	88	90.7
Emydid turtles	*Glyptemys*	BA	1.00	22.3
	Terrapene	BA	1.00	42.6
	Pseudemys	BA	1.00	33.1
	Trachemys	BA	Not supported	72.6
	Malaclemys	BA	1.00	15.2
	Graptemys	BA	1.00	57.6
Phrynosomatid lizards	*Holbrookia*	ML	100	77.6
	Uma	ML	100	81.5
	Phrynosoma	ML	100	92.3
	Uta	ML	100	53.2
	Petrosaurus	ML	100	37.1
	Urosaurus	ML	100	52.6
	Sceloporus (including *Sator*)	ML	84	91.9
Snakes	Tropidophiidae	ML	100	16.9
	Pythonidae	ML	100	51.4
	Uropeltidae	ML	100	72.4
	Boidae	ML	100	47.7
	Viperidae	ML	100	23.2
	Atractaspididae	ML	100	68.9
	Boodontidae	ML	100	12.9
	Elapidae	ML	100	12.8
	Colubridae-Xenodontinae	ML	100	31.4
	Colubridae-Colubrinae	ML	100	15.8
	Colubridae-Natricinae	ML	100	58.3

Notes: BA = Bayesian analysis; ML = maximum likelihood.

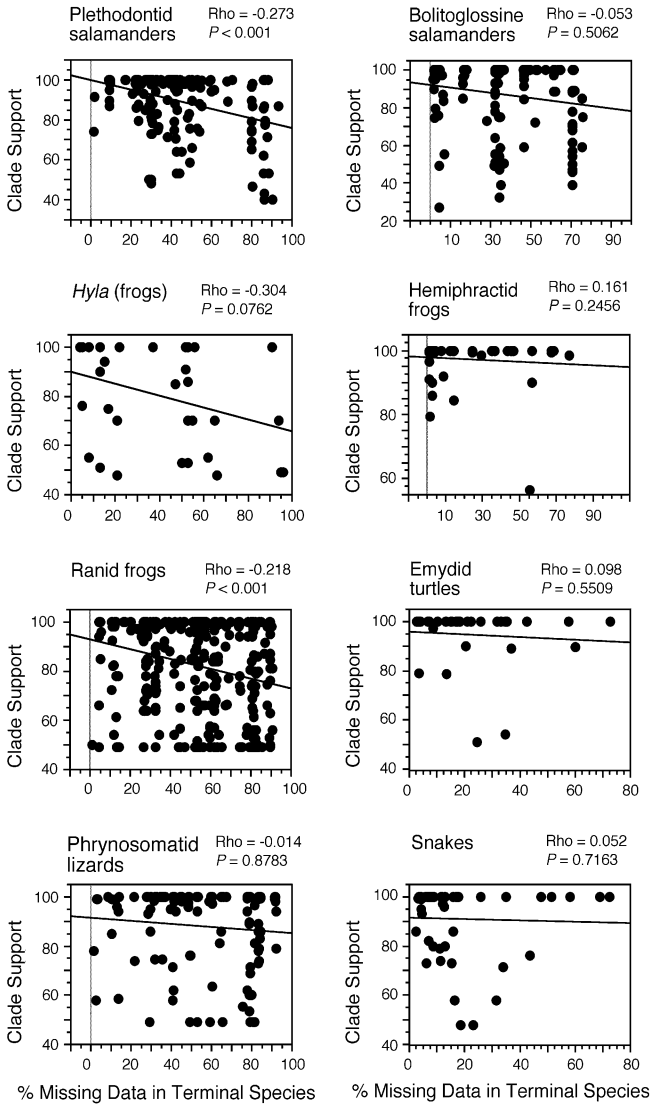

FIGURE 3. The relationship between the incompleteness of a taxon (% missing data) and the support for its localized placement in phylogenetic analyses using likelihood and Bayesian analysis.

divergence-dating analyses using matrices that contain extensive missing data (e.g., Wiens, Parra-Olea, et al. 2007; Kozak et al. 2009; Wiens et al. 2009), using both penalized likelihood and Bayesian approaches (Sanderson 2002; Drummond et al. 2006). Yet, we have found no evidence to suggest that these estimates are generally misled by missing data. Instead, these estimates are generally similar to those for the same groups based on smaller data sets with fewer missing data cells (e.g., Bossuyt et al. 2006 vs. Wiens et al. 2009 for ranid frogs; Wiens 2007 vs. Kozak et al. 2009 for plethodontid salamanders). But again, this is an area in need of further investigation.

Other Areas

Many other areas remain to be investigated. For example, it is unclear how congruence among gene trees may interact with missing data to impact phylogenetic accuracy. All simulation studies published so far have assumed that different genes share the same history, and have been based on combined analysis of genes (either implicitly or explicitly). The impact of missing data on methods that estimate species trees without concatenation (e.g., Edwards et al. 2007) also requires study.

The impact of missing data on support values would also benefit from additional study. For examples, simulations are needed to address whether the standard interpretation of support values (e.g., likelihood bootstrap support, Bayesian Pp) remains valid for taxa with extensive missing data.

CONCLUSIONS

LEA state (p. 130) that the results of their study "have major implications for all analyses that rely on accurate estimates of topology or branch lengths, including divergence time estimation, ancestral state reconstruction, tree-dependent comparative methods, rate variation analysis, phylogenetic hypothesis testing, and phylogeographic analysis." However, examination of their results shows that their evidence for the negative impacts of missing data hinge largely on methodological choices that would presumably not be made by most empirical systematists (e.g., adding data sets consisting of invariant or "saturated" characters, failing to partition data sets evolving at dramatically different rates). Unless those choices are made, their sweeping generalizations are not supported by their own results. These generalizations are also contradicted by many previous simulation and empirical studies, and also by new results from simulations that incorporate larger numbers of taxa and data partitioning (Fig. 1), from reanalysis of their plethodontid salamander example (Fig. 2), and from eight empirical data sets analyzed here (Fig. 3). In contrast to the idea of discordance among studies promoted by LEA, we argue that most results on missing data can be explained in a common theoretical framework (Wiens 2003b), and that most studies suggest that it should generally be possible to accurately place incomplete taxa in phylogenies, if enough informative characters are sampled. We think that there is a need for continued investigation of the impact of missing data on phylogenetics, and we point out specific topics in particular need of focused research. However, future studies should strive to reconcile their new results with those from previous studies in order to make real progress in this area.

FUNDING

This work was supported by the U.S. National Science Foundation (EF 0334923 to J.J.W).

ACKNOWLEDGMENTS

For initial comments on the manuscript, we thank N. Butler, M. C. Fisher-Reid, X. Hua, D. Moen, A. Pyron.

For comments on the submitted version, we thank the reviewers (H. Philippe, A. Lemmon, and anonymous) and associate editor (K. Kjer).

REFERENCES

Alfaro M.E., Zoller S., Lutzoni F. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. 20:255–266.

Anderson J.S. 2001. The phylogenetic trunk: maximal inclusion of taxa with missing data in an analysis of the Lepospondyli (Vertebrata, Tetrapoda). Syst. Biol. 50:170–193.

Bossuyt F., Brown R.M., Hillis D.M., Cannatella D.C., Milinkovitch M.C. 2006. Phylogeny and biogeography of a cosmopolitan frog radiation: Late Cretaceous diversification resulted in continent-scale endemism in the family Ranidae. Syst. Biol. 55:579–594.

Cobbett A., Wilkinson M., Wills M.A. 2007. Fossils impact as hard as living taxa in parsimony analyses of morphology. Syst. Biol. 56:753–766.

Donoghue M.J., Doyle J.A., Gauthier J., Kluge A.G., Rowe T. 1989. The importance of fossils in phylogeny reconstruction. Annu. Rev. Ecol. Syst. 20:431–460.

Dragoo J.W., Honeycutt R.L. 1997. Systematics of mustelid-like carnivores. J. Mammal. 78:426-443.

Driskell A.C., Ané C., Burleigh J.G., McMahon M.M., O'Meara B.C., Sanderson M.J. 2004. Prospects for building the Tree of Life from large sequence databases. Science. 306:1172–1174.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Dunn K.A., McEachran J.D., Honeycutt R.L. 2003. Molecular phylogenetics of myliobatiform fishes (Chondrichthyes: Myliobatiformes), with comments on the effects of missing data on parsimony and likelihood. Mol. Phylogenet. Evol. 27:259–270.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

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

Frost D.R., Grant T., Faivovich J., Bain R.H., Haas A., Haddad C.F.B., de Sá R.O., Channing A., Wilkinson M., Donnellan S.C., Raxworthy C.J., Campbell J.A., Blotto B.L., Moler P., Drewes R.C., Nussbaum R.A., Lynch J.D., Green D.M., Wheeler W.C. 2006. The amphibian tree of life. Bull. Am. Mus. Nat. Hist. 297:1–370.

Gouveia-Oliveira R., Sackett P.W., Pedersen A.G. 2007. MaxAlign: maximizing usable data in an alignment. BMC Bioinformatics. 8:312.

Hartmann S., Vision T.J. 2008. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evol. Biol. 8:95.

Hua X., Fu C., Li J., Nieto-Montes de Oca A., Wiens J.J. 2009. A revised phylogeny of Holarctic treefrogs (genus *Hyla*) based on nuclear and mitochondrial DNA sequences. Herpetologica. 65:246–259.

Huelsenbeck J.P. 1991. When are fossils better than extant taxa in phylogenetic analysis? Syst. Zool. 40:458–469.

Huelsenbeck J.P. 1995. The performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.

Huelsenbeck J.P., Ronquist F. 2001. MrBayes: Bayesian inference of phylogeny. Bioinformatics. 17:754–755.

Huelsenbeck J.P., Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities. Syst. Biol. 53:904–913.

Kearney M. 2002. Fragmentary taxa, missing data, and ambiguity: Mistaken assumptions and conclusions. Syst. Biol. 51:369–381.

Kozak K.H., Mendyk R.W., Wiens J.J. 2009. Can parallel diversification occur in sympatry? Repeated patterns of body-size evolution in co-existing clades of North American salamanders. Evolution. 63:1769–1784.

Lemmon A.R., Brown J.M., Stanger-Hall K., Moriarty-Lemmon E. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58:130–145.

Lewis P.O., Holder M.T., Holsinger K.E. 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol. 54:241–253

Lynch V.J., Wagner G.P. 2010. Did egg-laying boas break Dollo Law? Phylogenetic evidence for reversal to oviparity in sand boas (*Eryx*: Boidae). Evolution. 64:207–216.

Manos P.S., Soltis P.S., Soltis D.E., Manchester S.R., Oh S.-H., Bell C.D., Dilcher D.L., Stone D.E. 2007. Phylogeny of extant and extinct Juglandaceae inferred from the integration of molecular and morphological data sets. Syst. Biol. 56:412–430.

Mueller R.L., Macey J.R., Jaekel M., Wake D.B., Boore J.L. 2004. Morphological homoplasy, life history evolution, and historical biogeography of plethodontid salamanders inferred from complete mitochondrial genomes. Proc. Natl. Acad. Sci. U.S.A. 101:13820–13825.

Novacek M.J. 1992. Fossils, topologies, missing data, and the higher level phylogeny of eutherian mammals. Syst. Biol. 41:58–73.

Parra-Olea G., Wake D.B. 2001. Extreme morphological and ecological homoplasy in tropical salamanders. Proc. Natl. Acad. Sci. U.S.A. 98:7888-7891.

Philippe H., Snell E.A., Bapteste E., Lopez P., Holland P.W.H., Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21:1740–1752.

Platnick N.I., Griswold C.E., Coddington J.A. 1991. On missing entries in cladistic analysis. Cladistics. 7:337–343.

Poe S. 2003. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. Syst. Biol. 52:423–428.

Pyron R.A., Burbrink F.T., Colli G.R., Nieto Montes de Oca A., Vitt L.J., Kuczynski C.A., Wiens J.J. 2011. The phylogeny of advanced snakes (Colubroidea), with discovery of a new subfamily and comparison of support methods for likelihood trees. Mol. Phylogenet. Evol. 58:329–342.

Rannala B., Huelsenbeck J.P., Yang Z., Nielsen R. 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. 47:702–710.

Sanderson M.J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19:101–109.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–2690.

Stanger-Hall K.F., Lloyd J.E., Hillis D.M. 2007. Phylogeny of North American fireflies (Coleoptera: Lampyridae): implications for the evolution of light signals. Mol. Phylogenet. Evol. 45:33–49.

Thomson R.C., Shaffer H.B. 2010. Sparse supermatrices for phylogenetic inference: Taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. Syst. Biol. 59:42–58

Wiens J.J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst. Biol. 47:625–640.

Wiens J.J. 2003a. Incomplete taxa, incomplete characters, and phylogenetic accuracy: what is the missing data problem? J. Vertebr. Paleontol. 23:297–310.

Wiens J.J. 2003b. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52:528–538.

Wiens J.J. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst. Biol. 54:731–742.

Wiens, J.J. 2006. Missing data and the design of phylogenetic analyses. J. Biomed. Inform. 39:34–42.

Wiens J.J. 2007. Global patterns of species richness and diversification in amphibians. Am. Nat. 170:S86–S106.

Wiens J.J. 2009. Paleontology, genomics, and combined-data phylogenetics: can molecular data improve phylogeny estimation for fossil taxa? Syst. Biol. 58:87–99.

Wiens J.J., Fetzner J.W., Parkinson C.L., Reeder T.W. 2005. Hylid frog phylogeny and sampling strategies for speciose clades. Syst. Biol. 54:719-748.

Wiens J.J., Kuczynski C.A., Arif S., Reeder T.W. 2010. Phylogenetic relationships of phrynosomatid lizards based on nuclear and mitochondrial data, and a revised phylogeny for *Sceloporus*. Mol. Phylogenet. Evol. 54:150–161.

Wiens J.J., Kuczynski C., Duellman W.E., Reeder T.W. 2007. Loss and re-evolution of complex life cycles in marsupial frogs: can ancestral trait reconstruction mislead? Evolution 61:1886–1899.

Wiens J.J., Kuczynski C.A., Smith S.A., Mulcahy D.G., Sites J.W., Townsend T.M., Reeder T.W. 2008. Branch lengths, support, and congruence: testing the phylogenomic approach with 20 nuclear loci in snakes. Syst. Biol. 57:420–431.

Wiens J.J., Kuczynski C.A., Stephens P.R. 2010. Discordant mitochondrial and nuclear gene phylogenies in emydid turtles: implications for speciation and conservation. Biol. J. Linn. Soc. 99: 445–461.

Wiens J.J., Kuczynski C.A., Townsend T., Reeder T.W., Mulcahy D.G., Sites J.W. Jr. 2010. Combining phylogenomics and fossils in higher level squamate reptile phylogeny: molecular data change the placement of fossil taxa. Syst. Biol. 59:674–688.

Wiens J.J., Moen D.S. 2008. Missing data and the accuracy of Bayesian phylogenetics. J. Syst. Evol. 46:307–314.

Wiens J.J., Parra-Olea G., Garcia-Paris M., Wake D.B. 2007. Phylogenetic history underlies elevational patterns of biodiversity in tropical salamanders. Proc. R. Soc. Lond. B 274:919–928.

Wiens J.J., Reeder T.W. 1995. Combining data sets with different numbers of taxa for phylogenetic analysis. Syst. Biol. 44:548–558.

Wiens J.J., Sukumaran J., Pyron R.A., Brown R.M. 2009. Evolutionary and biogeographic origins of high tropical diversity in Old World frogs (Ranidae). Evolution 63:1217–1231.

Wilcox T.P., Zwickl D.J., Heath T.A., Hillis D.M. 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Mol. Phylogenet. Evol. 25:361–371.

Wilkinson M. 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. Syst. Biol. 44:501–514.