Journal of Systematics and Evolution 46 (3): 307–314 (2008)
(formerly Acta Phytotaxonomica Sinica)
doi: 10.3724/SP.J.1002.2008.08040
http://www.plantsystematics.com

Missing data and the accuracy of Bayesian phylogenetics

John J. WIENS* Daniel S. MOEN
(*Department of Ecology and Evolution, Stony Brook University,* Stony Brook, NY 11794-5245, USA)

Abstract The effect of missing data on phylogenetic methods is a potentially important issue in our attempts to reconstruct the Tree of Life. If missing data are truly problematic, then it may be unwise to include species in an analysis that lack data for some characters (incomplete taxa) or to include characters that lack data for some species. Given the difficulty of obtaining data from all characters for all taxa (e.g., fossils), missing data might seriously impede efforts to reconstruct a comprehensive phylogeny that includes all species. Fortunately, recent simulations and empirical analyses suggest that missing data cells are not themselves problematic, and that incomplete taxa can be accurately placed as long as the overall number of characters in the analysis is large. However, these studies have so far only been conducted on parsimony, likelihood, and neighbor-joining methods. Although Bayesian phylogenetic methods have become widely used in recent years, the effects of missing data on Bayesian analysis have not been adequately studied. Here, we conduct simulations to test whether Bayesian analyses can accurately place incomplete taxa despite extensive missing data. In agreement with previous studies of other methods, we find that Bayesian analyses can accurately reconstruct the position of highly incomplete taxa (i.e., 95% missing data), as long as the overall number of characters in the analysis is large. These results suggest that highly incomplete taxa can be safely included in many Bayesian phylogenetic analyses.
Key words accuracy, Bayesian analysis, missing data, phylogenetic analysis.

The impact of missing data is a potentially important issue in phylogenetic analyses, particularly if the goal is to reconstruct a comprehensive Tree of Life that includes both fossil and living taxa. Missing data are often encountered when combining data from two or more different genes, when some of the taxa have sequence data available for one gene but not the other. If the taxa lacking data for a gene are included in the combined analysis, then the characters associated with this gene are typically coded as missing or unknown (often denoted with a "?"). Similarly, missing data are often encountered in analyses that include fossil taxa, when certain taxa must be scored as unknown for certain characters because the relevant features have not been adequately preserved.

Concerns about missing data may often determine what characters and taxa will be included in an analysis (Wiens, 2006), even if this is not always stated explicitly by researchers. For example, if missing data are considered to be problematic, then one should only include species that have complete data for all characters or else only include characters that have complete data for all species. Thus, one may have to reduce the number of taxa or characters in an analysis in order to avoid including missing data cells. Furthermore, it would be difficult (if not impossible) to combine molecular and morphological character data from fossil and living taxa in the same analysis, because the fossil taxa will almost always lack molecular data.

But are missing data truly problematic? Several authors have suggested that including taxa with a high proportion of missing data cells is potentially problematic for phylogeny reconstruction, based on both empirical data (e.g., Novacek, 1992; Wiens & Reeder, 1995; Wilkinson, 1995; Kearney, 2002) and computer simulations (Huelsenbeck, 1991; Hartmann & Vision, 2008). These authors suggested that including highly incomplete taxa can potentially lead to uncertain relationships (e.g., Novacek, 1992) and an overall decrease in the accuracy of the reconstructed trees (Huelsenbeck, 1991; Hartmann & Vision, 2008). By accuracy, we mean the frequency with which the true phylogeny is reconstructed correctly.

Computer simulations, such as those of Huelsenbeck (1991), can offer important insights into whether a given phylogenetic method is able to accurately reconstruct the true phylogeny under a broad range of conditions (Hillis, 1995; Huelsenbeck, 1995). Simulations are important because in most empirical studies the true phylogeny of the organisms is unknown. In contrast, simulations provide a context where the true phylogeny is known and the conditions that affect the

Received: 1 April 2008 Accepted: 2 May 2008
* Author for correspondence. E-mail: wiensj@life.bio.sunysb.edu; Tel.: 631-632-1101.

phylogenetic accuracy of a method can be varied in a controlled, experimental fashion. However, simulations always require making many simplifying assumptions, and the results may depend entirely on the set of simulated conditions that were examined (Hillis, 1995; Huelsenbeck, 1995). For example, Huelsenbeck (1991) examined the effects of missing data on parsimony analysis under a broad range of conditions, but did not vary the number of characters (only 100 characters were included in the simulations).

Wiens (2003) conducted simulations (in which the number of characters was varied extensively) which indicate that missing data are not themselves problematic. Instead, problems with highly incomplete taxa arise because there are too few characters in these taxa to accurately place them on the tree. If the overall number of characters in the analysis is small, then overall accuracy may be low when many of the taxa are incomplete (i.e., if the overall number of characters in the analysis is only 100, a taxon with 95% missing data will have data for only 5 characters). Conversely, if the overall number of characters is large, then even highly incomplete taxa should have enough characters to allow them to be accurately placed in the tree (i.e., if there are 2,000 characters, and a taxon has 95% missing data, there are still 100 characters that can allow it to be placed on the tree). Based on these results, the missing data themselves are irrelevant, and the more important parameter is the quantity of the characters that are present. However, these results are dependent not only on the simulated conditions, but also on how each phylogenetic method and software package deals with missing data. Wiens (2003) examined the performance of parsimony, likelihood and neighbor-joining as implemented by the widely used software package PAUP* (Swofford, 2002). Subsequently, an empirical molecular study of plant phylogeny by Driskell et al. (2004) and a combined empirical and simulation study of higher-level eukaryote molecular phylogeny by Philippe et al. (2004) also found results suggesting that incomplete taxa can be accurately placed in phylogenetic analyses. The results of Driskell et al. (2004) were based on parsimony (using PAUP*), and those of Philippe et al. (2004) were based on parsimony, likelihood, and neighbor-joining (using various programs, but especially PHYML; Guindon & Gascuel, 2003). A recent simulation study analyzed the effect of missing data in datasets resembling those from ESTs (Hartmann & Vision, 2008) and found that replacing complete data with missing data cells decreased accuracy, especially when the missing data are randomly distributed among characters and taxa (but note that if data are simply replaced with missing data cells, one expects accuracy to decrease simply because you are decreasing the overall amount of data).

The papers by Wiens (2003), Driskell et al. (2004), Philippe et al. (2004), and Hartmann & Vision (2008) all shared a serious omission, however. Starting around 2002, Bayesian methods have become widely used for reconstructing phylogenies. Although the performance of Bayesian phylogenetic methods has become relatively well understood through a large number of simulation studies (e.g., Suzuki et al., 2002; Wilcox et al., 2002; Alfaro et al., 2003; Cummings et al., 2003; Douady et al., 2003; Erixon et al., 2003; Huelsenbeck & Rannala, 2004; Lewis et al., 2005), how Bayesian methods perform when faced with missing data is largely unknown.

Two previous studies, one empirical and one based on computer simulations, suggest that missing data may not be problematic for Bayesian analyses. First, Wiens et al. (2005) examined relationships among hylid frog species using Bayesian analysis of molecular and morphological data. They had relatively complete data for 81 species and incomplete data (typically with data for only one gene) for an additional 117. They found that the taxa with only one gene were placed in the expected clades, despite their missing data. For example, the eight species in the analysis that had >90% missing data were each placed in the clades expected based on their current taxonomy (e.g., species of the genus *Scinax* were placed with other species of *Scinax*), and the Bayesian support for the monophyly of these clades was very high (posterior probability of all clades=1.00). Furthermore, these authors found no relationship between levels of completeness (100 - % missing data) in each species and the level of support (Bayesian posterior probability) for the placement of that species on the terminal branches of the tree. Instead, there was a significant relationship between the level of support in the combined analysis (including missing data) and the level of support in the analyses of the gene that was sequenced in almost all taxa (mitochondrial ribosomal 12S). In other words, the level of support seemed to depend on the data that were present, not the amount of data that was absent. However, it is important to note that the actual phylogenetic relationships of the species were unknown, and so this study did not directly assess the impact of missing data on phylogenetic accuracy of Bayesian analysis.

In the second study, Wiens (2005) used simulations to test the ability of added taxa to improve

phylogenetic accuracy (for the complete taxa) for various phylogenetic methods when the added taxa had a high proportion of missing data. These analyses included Bayesian analyses in addition to parsimony, likelihood, and neighbor-joining. Taxa were added under simulated conditions where there was long-branch attraction (which greatly reduces phylogenetic accuracy for most methods; Felsenstein, 1978; Huelsenbeck, 1995) and the added taxa could subdivide or "break up" these long branches (e.g., Poe, 2003). These analyses showed that incomplete taxa can successfully subdivide long branches and thereby increase phylogenetic accuracy, in many cases, as well as complete taxa can. Although these results are encouraging about the ability of Bayesian analyses to cope with missing data, these analyses did not address the overall accuracy of the trees, only the accuracy of the relationships among the complete taxa. Thus, it is theoretically possible that the incomplete taxa were placed inaccurately. In summary, despite some encouraging results from two previous studies, the impact of missing data on the accuracy of Bayesian analysis is in need of further study.

In this paper, we use computer simulations to explicitly examine the accuracy of Bayesian phylogenetic analysis when many of the taxa are incomplete. These analyses follow closely the protocols used by Wiens (2003), in order to make the results easily comparable to those based on other phylogenetic methods. Indeed, we find here that the results from Bayesian analyses mirror those from likelihood, parsimony, and neighbor-joining. In Bayesian analyses, highly incomplete taxa can be accurately placed if the overall number of characters is large. As in previous studies, we find that the missing data cells themselves do not appear to be problematic for phylogeny reconstruction.

1 Material and Methods

The general methodology for simulating data and analyzing these data followed Wiens (2003), and only a brief explanation is provided here. The overall design was to test the accuracy of Bayesian analysis when many of the taxa have different proportions of missing data, and to test this across different numbers of characters. Based on previous studies, we anticipated that analyses including taxa with a high proportion of missing data would have low accuracy when the number of characters in the analysis was small, but relatively high accuracy when the number of characters was large.

We initially simulated a 16-taxon phylogeny that was fully asymmetric and had equal branch lengths. Asymmetric trees are expected to be more common when all topologies are considered to be equally likely (e.g., Huelsenbeck & Kirkpatrick, 1996), and previous studies suggest that tree shape will have little impact on the results (Wiens, 2003). Characters were simulated along this phylogeny to create a complete set of character data for each taxon. We used DNA sequence data evolving according to the simple Jukes-Cantor model (Jukes & Cantor, 1969), with equal rates of change between all substitution types, equal base frequencies, and equal rates of change between characters. We focused primarily on how Bayesian analyses are affected by missing data, and not other issues (e.g., how they deal with more complex models of evolution). Different branch lengths were also analyzed ranging from 0.05 (i.e., 1 in 20 characters expected to change from the beginning to the end of the branch), 0.10, 0.20, to 0.30. These branch lengths span a broad range of levels of variability and homoplasy, ranging from conditions where phylogeny reconstruction is relatively easy to those where it is relatively difficult. Different numbers of characters were also analyzed including 100, 500, 1,000, and 2,000. These total numbers of characters included both sites that were variable and invariant among taxa.

For a given simulation replicate, 8 taxa were then randomly chosen to be incomplete. These taxa had a certain proportion of their characters replaced with missing data cells ("?"). The same characters were made incomplete in each incomplete taxon (but see below). We systematically varied the level of completeness from 5% (95% of data cells missing), to 10%, 25%, 50%, 75%, and 100% (no missing data). 100 replicates were examined for each combination of level of completeness, branch length, and number of characters.

Each simulated data set was analyzed using MrBayes version 3.0b4 (Huelsenbeck & Ronquist, 2001) and the resulting tree was compared to the true, known phylogeny that was used to generate the data. Accuracy for a given replicate was measured as the proportion of nodes from the estimated Bayesian tree that matched nodes in the true species phylogeny. The accuracy for a given set of simulated conditions was the average accuracy across all 100 replicates.

In general, we used default options for MrBayes in the Bayesian analyses. These default options included use of the Jukes-Cantor model with no among-site rate variation or invariant sites (thus, there was a close correspondence between the model used to simulate the data and reconstruct the tree). The

number of generations was set to 50,000. The first 5,000 generations were discarded as burn-in, and the phylogeny was estimated as the majority-rule consensus tree of the post burn-in trees, following standard practice. Some readers may be surprised by the low number of generations used. After all, empirical analyses typically use several million. However, the overall number of taxa analyzed in each replicate is relatively low (16; making thorough searching of tree space much easier). Furthermore, prior to selecting this number of generations, we analyzed a subset of the results using twice as many generations (100,000) and found no detectable difference in the results. We also examined the results using 100,000 generations to determine when stationarity was achieved, and found that it was consistently reached in less than 5,000 generations (based on a plateau in a plot of likelihood values against number of generations). Finally, we found that even when using only 50,000 generations, Bayesian analyses had an accuracy of 100% (all nodes correctly reconstructed) under many different conditions. This result demonstrates that, at least under these conditions, there are no random errors in tree reconstruction associated with inadequate tree searching.

Programs for simulating the data and compiling the results were written in C by J.J.W. Analyses were conducted using PAUP* version 4.0b10 (Swofford, 2002) to make consensus trees from the Bayesian analyses and to compare the estimated Bayesian trees to the true trees.

In addition to these basic simulations, we also performed a smaller set of simulations to test how robust the results were to changes in different parameters. First, we examined more complex models of DNA sequence evolution. We performed analyses that incorporated unequal base frequencies, a different ratio of transitions and transversion, and different rates of change among sites. We assumed a 3:1 transition: transversion ratio and base frequencies of A=37%, G=12%, C=24%, and T=27% (parameter values based on mammalian sequences as reported by Zwickl & Hillis, 2002). We simulated among-site rate variation by modeling the data to resemble protein-coding sequences. Thus, the first two characters of every three had branch lengths of 0.02 and 0.02, whereas the third had a branch length of 0.20 (the ten-fold difference in rates was initially chosen based on protein-coding genes in salamanders; Wiens, unpubl.). Overall, the simulated data corresponded to the HKY model (Hasegawa et al., 1985). We performed two sets of analyses under these conditions. In the first, we analyzed the data using the simple Jukes-Cantor model, to evaluate the combined effects of missing data and an oversimplified model. In the second, we analyzed the data using a more appropriate model (HKY, with the gamma parameter added to account for among-site rate variation; Yang, 1993).

We also examined the effects of changing tree shape. We analyzed a fully symmetric 16-taxon tree for a limited set of conditions (branch length=0.05, Jukes-Cantor model of sequence evolution).

Finally, we changed the way in which missing data were distributed among characters in the incomplete taxa. Instead of having the same set of characters lacking data in all of the incomplete taxa, we randomly distributed missing data cells among characters. Again, we analyzed a limited set of conditions to address the effects of changing this parameter (asymmetric tree with branch lengths=0.05, Jukes-Cantor model).

2 Results

The main results of the study are summarized in Fig. 1. These results show that highly incomplete taxa can be accurately placed in Bayesian analyses as long as the overall number of characters in the analysis is large. When the number of characters is low, Bayesian analyses that include highly incomplete taxa may have relatively low accuracy. But it is clear that the low accuracy in analyses with 100 characters and 75%–90% missing data is not directly caused by a large number or proportion of missing data cells, because analyses with 2,000 characters and 95% missing data have relatively high accuracy but a larger number and higher proportion of missing data cells. Instead, the low accuracy in analyses with 100 characters is presumably associated with the limited number of characters that are present and that can place these highly incomplete taxa on the tree. These general results are robust across various branch lengths (Fig. 1), and closely parallel those for parsimony, likelihood, and neighbor-joining under comparable simulation conditions (Wiens, 2003). These results also appear to robust when analyzing data evolved under more complex models of sequence evolution (regardless of whether that complexity is included in the analysis; Fig. 2: A, B), under different tree shapes (Fig. 2: C), and different ways of distributing missing data among characters (Fig. 2: D).

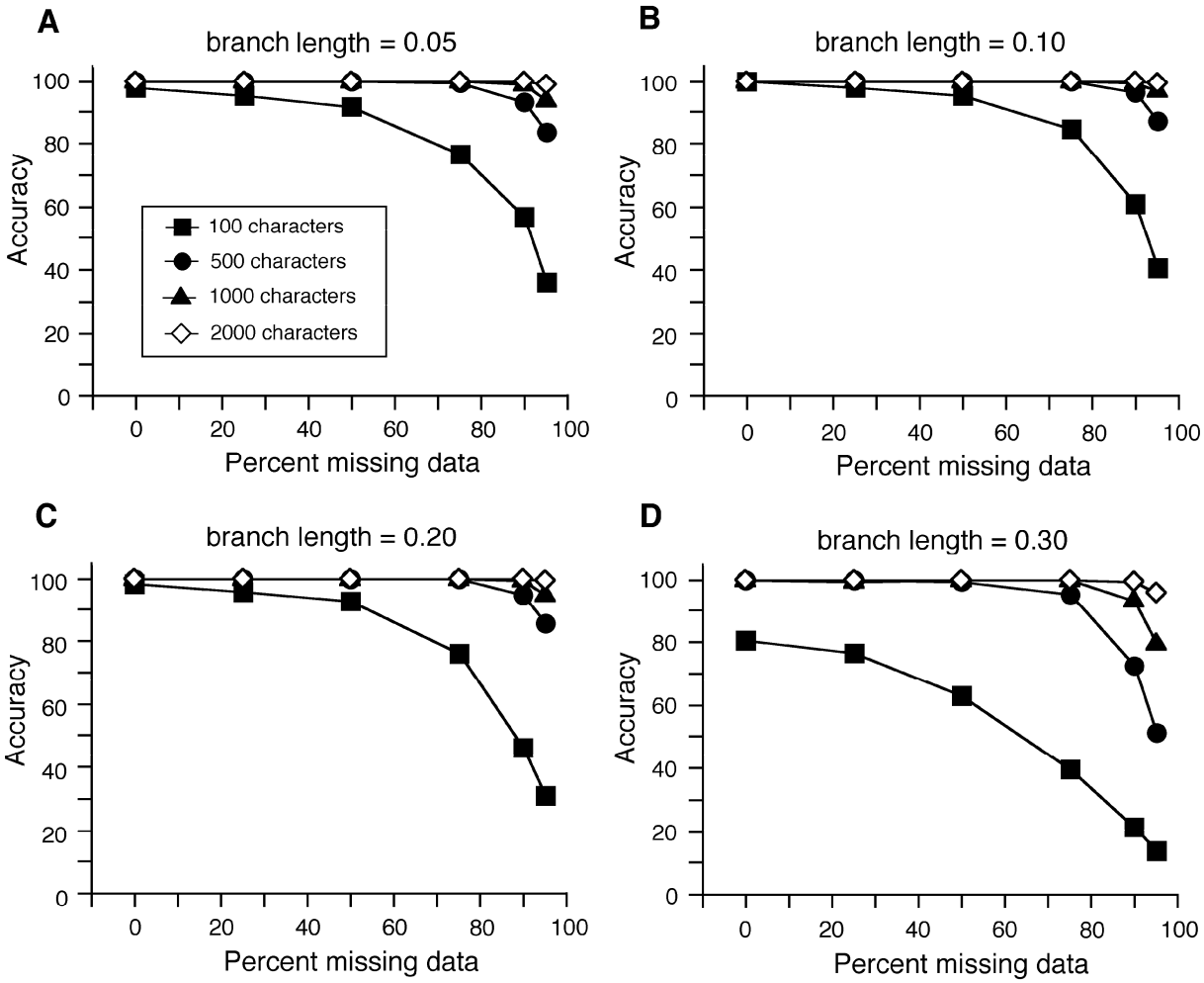

Fig. 1. Results of simulations showing the effects of missing data on the accuracy of Bayesian phylogenetic analysis. The results show that highly incomplete taxa are only problematic when the number of characters is very low. When the number of characters in the analysis is large, even taxa with 95% missing data can be accurately placed. Each data point is the mean of 100 replicates. **A–D** refers to different branch lengths on the simulated phylogeny. Accuracy refers to the percentage of nodes that are correctly reconstructed in the Bayesian majority-rule consensus tree of each replicate, averaged across the 100 replicates. The percentage of missing data refers to the proportion of missing data cells in each of the 8 taxa that are randomly selected to be incomplete in each replicate (out of 16 taxa total).

3 Discussion

Our results suggest that highly incomplete taxa can be accurately placed in Bayesian phylogenetic analyses, as long as the number of characters in the analyses is not unusually low. These results further support those of empirical analyses (Wiens et al., 2005), simulation analyses with a different design (Wiens, 2005), and empirical and simulation analyses based on other methods, such as parsimony and likelihood (e.g., Wiens, 2003; Phillipe et al., 2004). Taken together, these results indicate that taxa should not be excluded from Bayesian analyses merely because they have many missing data cells. Furthermore, these results suggest considerable promise for constructing a comprehensive Tree of Life using Bayesian methods, even though some taxa may be missing data for many characters. However, a number of caveats should be noted.

3.1 Simulations versus the real world

Simulations involve many simplifying assumptions, and considerable caution must be taken when using the results of simulation studies to inform empirical analyses (Hillis, 1995; Huelsenbeck, 1995). For example, the combination of tree shapes (fully asymmetric or fully symmetric), branch lengths (all

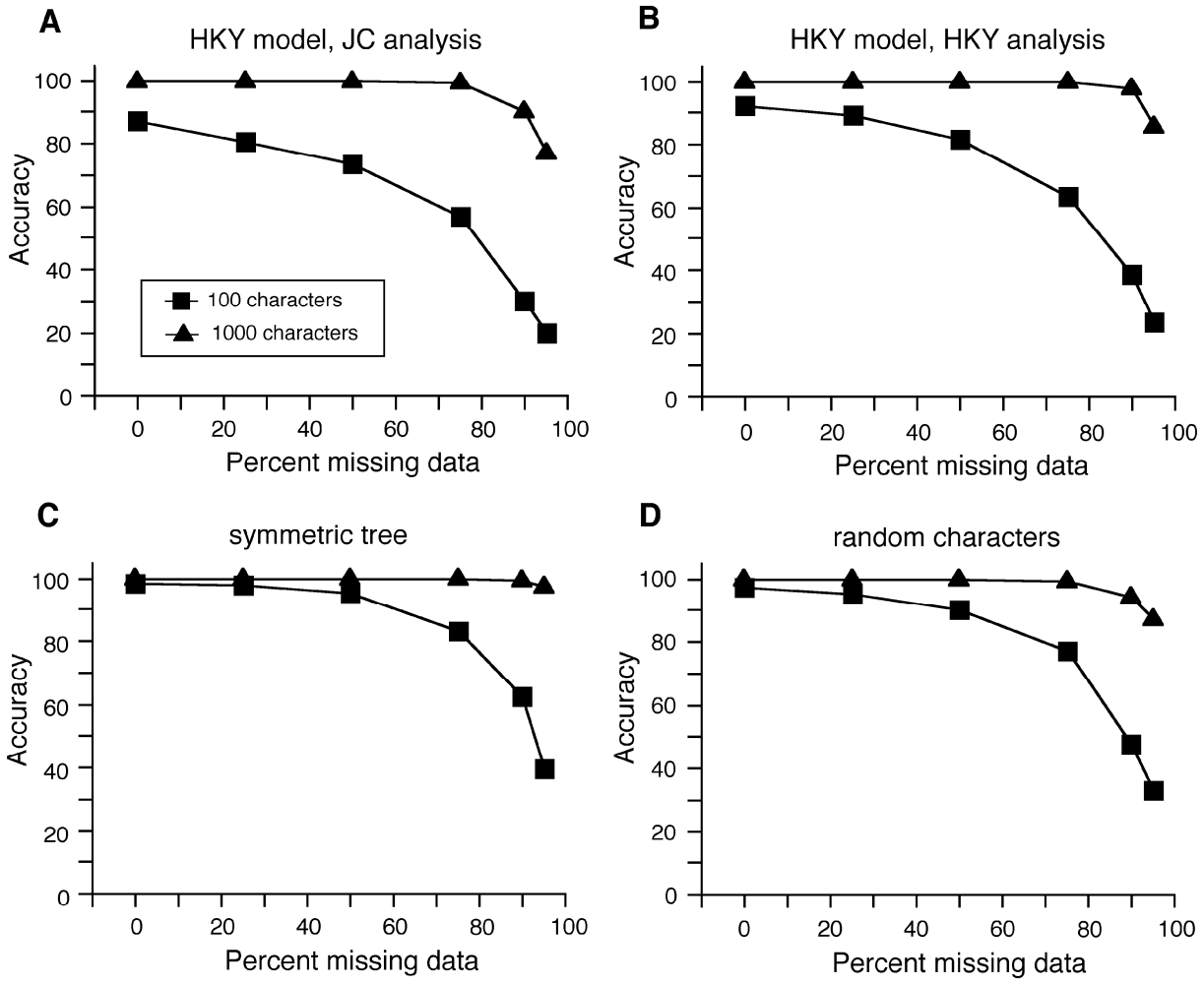

Fig. 2. Results of simulations showing the effects of missing data on the accuracy of Bayesian phylogenetic analysis, as in Fig. 1. **A,** Data simulated under the HKY model with unequal rates of change among characters, but analyzed under the JC model. **B,** Data simulated under the HKY model with unequal rates of change among characters, analyzed under the HKY + Γ model. **C,** Data simulated on a fully symmetric tree, using the JC model with branch lengths of 0.05 (analyzed using JC model). **D,** Data simulated using the JC model with branch lengths of 0.05 (analyzed using JC model), with missing data randomly distributed among characters in the 8 incomplete taxa.

equal), and simple models of sequence evolution that were simulated here will not be encountered in every (or perhaps any) empirical data set. Nevertheless, our analyses here suggest that tree shape *per se* has little impact on the results, given that the two most extreme tree shapes possible gave similar results (Figs. 1, 2). Our basic results are also supported under a broad range of equal branch lengths (Fig. 1). In theory, there might be additional negative consequences of missing data for Bayesian analysis when analyzing certain combinations of unequal branch lengths (e.g., Huelsenbeck, 1995), but these were not apparent in a previous simulation study (Wiens, 2005). The results were also robust to increasing the complexity of the model of sequence evolution, even when the analysis failed to account for that complexity (i.e., analyzing data evolved under the HKY model with unequal rates among sites, but using only the simple JC model).

Although our results were robust under many different simulated conditions, we acknowledge that it is theoretically possible that there might be negative impacts of missing data when combined with other problems. For example, we assumed that the sets of missing and non-missing characters were basically equivalent. We expect that cases of extensive missing data could be far more problematic if the only characters that were non-missing were themselves problematic for some reason (e.g., evolving too slowly to be

informative or too quickly to be accurate). In such a situation, we would not expect highly incomplete taxa to be accurately placed by Bayesian analysis, or by any other method.

Finally, there are some particular distributions of missing data that most phylogenetic methods will not be readily able to deal with. For example, imagine that there are five species (A–E) having data for four genes (1–4) and another five species (F–J) having data for a different set of four genes (5–8). If these data sets are combined, there will be considerable missing data and the analysis will not be able to simultaneously resolve the relationships of all 10 species. However, the problem is that there is no overlap between the two data sets, not the amount of missing data *per se* (i.e., the matrix here would have 50% missing data cells, an amount which is unproblematic under most conditions we analyzed). Our simulations have focused primarily on the issue of including some incomplete taxa in analyses that include some complete taxa, and not that of combining poorly overlapping datasets with few taxa or characters in common (e.g., Sanderson et al., 2003).

3.2 Levels of support

Another issue that we did not examine here is the relationship between missing data and levels of support. It is theoretically possible that, under conditions when accuracy is low, missing data might cause Bayesian analyses to yield incorrect results with strong statistical support. Other studies have found that Bayesian analysis may overestimate support when relationships are highly uncertain (e.g., Suzuki et al., 2002), such as when three taxa split almost simultaneously (e.g., Lewis et al., 2005). This topic might be worthy of additional investigation. However, our results do suggest that the circumstances under which extensive missing data would be associated with low phylogenetic accuracy are quite limited.

3.3 Missing data and analysis time

A final issue that we did not address is the effect of missing data on the duration of analyses and the time to reach stationarity. One can imagine that more missing data cells might slow down a Bayesian analysis. Similarly, one could also imagine that the number of generations required to achieve stationarity might be increased by extensive missing data. We briefly tested these assumptions with our simulations. Under the conditions where Bayesian analysis has relatively low accuracy (100 characters, branch length = 0.05), the duration of the analysis is actually shorter with 95% missing data (in the 8 taxa) than with no missing data (mean with 95% missing data=73.4 seconds, range= 62–85; vs. mean with no missing data=102.1, range= 89–113; t-test, $P<0.0001$). Further, we were unable to detect any difference in the time to reach stationarity (i.e., before 5,000 generations in all cases). Although more analysis of these issues would be welcome, we have found no evidence so far to suggest that having extensive missing data leads to slower analyses or longer times to achieve stationarity.

Acknowledgements We would like to thank Yin-Long Qiu for helping to organize the Beijing Tree of Life symposium in 2007 and for inviting Wiens to participate and contribute this paper. We thank the U.S. National Science Foundation for financial support (EF 0334923 to J.J.W. and a Graduate Research Fellowship to D.S.M.). We are grateful to two anonymous reviewers for helpful comments on the manuscript.

References

Alfaro ME, 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. Molecular Biology and Evolution 20: 255–266.

Cummings MP, Handley SA, Myers DS, Reed DL, Rokas A, Winka K. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Systematic Biology 52: 477–487.

Douady CJ, Delsuc F, Boucher Y, Doolittle WF, Douzery EJP. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Molecular Biology and Evolution 20: 248–254.

Driskell AC, Ané C, Burleigh JG, McMahon MM, O'Meara BC, Sanderson MJ. 2004. Prospects for building the Tree of Life from large sequence databases. Science 306: 1172–1174.

Erixon P, Svennblad B, Britton T, Oxelman B. 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Systematic Biology 52: 665–673.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology 27: 401–410.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52: 696–704.

Hartmann S, Vision TJ. 2008. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evolutionary Biology 8: 95.

Hasegawa M, Kishino H, Yano T. 1985. Dating the human-ape spit by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22: 160–174.

Hillis DM. 1995. Approaches for assessing phylogenetic accuracy. Systematic Biology 44: 3–16.

Huelsenbeck JP. 1991. When are fossils better than extant taxa

in phylogenetic analysis? Systematic Zoology 40: 458–469.
Huelsenbeck JP. 1995. The performance of phylogenetic methods in simulation. Systematic Biology 44: 17–48.
Huelsenbeck JP, Kirkpatrick M. 1996. Do phylogenetic methods produce trees with biased shapes? Evolution 50: 1418–1424.
Huelsenbeck JP, Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities. Systematic Biology 53: 904–913.
Huelsenbeck JP, Ronquist F. 2001. MrBayes: Bayesian inference of phylogeny. Bioinformatics 17: 754–755.
Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro H ed. Mammalian protein metabolism. New York: Academic Press. 21–132.
Kearney M. 2002. Fragmentary taxa, missing data, and ambiguity: Mistaken assumptions and conclusions. Systematic Biology 51: 369–381.
Lewis PO, Holder MT, Holsinger KE. 2005. Polytomies and Bayesian phylogenetic inference. Systematic Biology 54: 241–253.
Novacek MJ. 1992. Fossils, topologies, missing data, and the higher level phylogeny of eutherian mammals. Systematic Biology 41: 58–73.
Philippe HE, Snell EA, Bapteste E, Lopez P, Holland PWH, Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Molecular Biology and Evolution 21: 1740–1752.
Poe S. 2003. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. Systematic Biology 52: 423–428.
Sanderson MJ, Driskell AC, Ree RH, Eulenstein O, Langley S. 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Molecular Biology and Evolution 20: 1036–1042.
Suzuki Y, Glazko GV, Nei M. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proceedings of the National Academy of Sciences USA 99: 16138–16143.
Swofford DL. 2002. PAUP*: Phylogenetic analysis using parsimony*, version 4.0b10. Sunderland, MA: Sinauer.
Wiens JJ. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Systematic Biology 52: 528–538.
Wiens JJ. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Systematic Biology 54: 731–742.
Wiens JJ. 2006. Missing data and the design of phylogenetic analyses. Journal of Biomedical Informatics 39: 34–42.
Wiens JJ, Fetzner JW, Parkinson CL, Reeder TW. 2005. Hylid frog phylogeny and sampling strategies for speciose clades. Systematic Biology 54: 719–748.
Wiens JJ, Reeder TW. 1995. Combining data sets with different numbers of taxa for phylogenetic analysis. Systematic Biology 44: 548–558.
Wilcox TP, Zwickl DJ, Heath TA, Hillis DM. 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Molecular Phylogenetics and Evolution 25: 361–371.
Wilkinson M. 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. Systematic Biology 44: 501–514.
Yang Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular Biology and Evolution 10: 1396–1401.
Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. Systematic Biology 51: 588–598.