Molecular Phylogenetics and Evolution 80 (2014) 308–318

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Should genes with missing data be excluded from phylogenetic analyses?

Wei Jiang [a,b,c], Si-Yun Chen [b], Hong Wang [a,b,c], De-Zhu Li [a,b,c,*], John J. Wiens [d]

[a] Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650201, China
[b] Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650201, China
[c] Kunming College of Life Sciences, University of Chinese Academy of Sciences, Kunming, Yunnan 650201, China
[d] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721-088, USA

ARTICLE INFO

Article history:
Received 25 January 2014
Revised 15 July 2014
Accepted 3 August 2014
Available online 11 August 2014

Keywords:
Accuracy
Maximum likelihood
Missing data
Phylogeny

ABSTRACT

Phylogeneticists often design their studies to maximize the number of genes included but minimize the overall amount of missing data. However, few studies have addressed the costs and benefits of adding characters with missing data, especially for likelihood analyses of multiple loci. In this paper, we address this topic using two empirical data sets (in yeast and plants) with well-resolved phylogenies. We introduce varying amounts of missing data into varying numbers of genes and test whether the benefits of excluding genes with missing data outweigh the costs of excluding the non-missing data that are associated with them. We also test if there is a proportion of missing data in the incomplete genes at which they cease to be beneficial or harmful, and whether missing data consistently bias branch length estimates. Our results indicate that adding incomplete genes generally increases the accuracy of phylogenetic analyses relative to excluding them, especially when there is a high proportion of incomplete genes in the overall dataset (and thus few complete genes). Detailed analyses suggest that adding incomplete genes is especially helpful for resolving poorly supported nodes. Given that we find that excluding genes with missing data often decreases accuracy relative to including these genes (and that decreases are generally of greater magnitude than increases), there is little basis for assuming that excluding these genes is necessarily the safer or more conservative approach. We also find no evidence that missing data consistently bias branch length estimates.

1. Introduction

The problem of missing data in phylogenetic analysis is an important issue because missing data are common in many data matrices (e.g., Philippe et al., 2004; Fulton and Strobeck, 2006; Burleigh et al., 2009), and are only absent in many others because taxa and genes are deliberately excluded in order to avoid them. For example, the issue of missing data may arise because of gaps in alignments, because data are unavailable for some species for some genes, or because molecular data are lacking entirely (e.g., fossils). There has been extensive debate about whether missing data should be included in phylogenetic analyses or not, and the possible consequences of both approaches (e.g., Huelsenbeck, 1991; Wiens and Reeder, 1995; Wiens, 1998, 2003a,b, 2005; Driskell et al., 2004; Philippe et al., 2004; Wiens et al., 2005, 2010; Wiens and Moen, 2008; Burleigh et al., 2009; Lemmon et al., 2009; Sanderson et al., 2010, 2011; Wiens and Morrill, 2011; Wiens and Tiu, 2012; Roure et al., 2013). In this debate, it is important to remember that missing data cells are only included because excluding missing data also requires excluding some taxa and/or characters from the analysis, which have non-missing data (Wiens, 1998; Cho et al., 2011; Schaefer and Renner, 2011; Zwick et al., 2011). The fundamental question is: when do the benefits of excluding the missing data outweigh the costs of excluding the non-missing data that are associated with them?

Missing data can be added to an analysis by two primary mechanisms: by adding incomplete taxa or by adding incomplete characters (Wiens, 2003a). Many studies have shown that incomplete taxa can often be included with relatively limited negative impacts, especially when the number of characters is large (e.g., Wiens, 2003b; Driskell et al., 2004; Philippe et al., 2004; Wiens and Moen, 2008; Cho et al., 2011; Wiens and Morrill, 2011; Wiens and Tiu, 2012; Roure et al., 2013). Specifically, these studies show

* Corresponding author at: Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650201, China.
E-mail addresses: jiangwei@mail.kib.ac.cn (W. Jiang), chensiyun@mail.kib.ac.cn (S.-Y. Chen), wanghong@mail.kib.ac.cn (H. Wang), dzl@mail.kib.ac.cn (D.-Z. Li), wiensj@email.arizona.edu (J.J. Wiens).

http://dx.doi.org/10.1016/j.ympev.2014.08.006
1055-7903/

that incomplete taxa can be placed correctly in phylogenies (based on simulations with a known true topology or based on concordance with other empirical studies), when sufficient characters have been sampled overall (review in Wiens and Morrill, 2011). Some studies have also shown that adding incomplete taxa can improve the accuracy of estimated relationships among the complete taxa (by breaking up long branches), using both simulated data (Wiens, 2005) and empirical data (Wiens and Tiu, 2012; Roure et al., 2013). In other words, adding incomplete taxa can potentially have similar benefits to adding complete taxa in these cases.

Far fewer studies have addressed the costs and benefits of adding characters with missing data. In a simulation study, Wiens (1998) found that for parsimony analyses adding incomplete characters was often beneficial, but became less beneficial with a greater proportion of missing data. Although this study found little evidence that adding characters with missing data generally decreased accuracy, it also showed that some patterns of missing data could create a problem of long-branch attraction among the species with non-missing data. Lemmon et al. (2009) analyzed simulations of the 4-taxon case and suggested that missing data could cause misleading results in Bayesian analyses with missing data in 2 of 4 taxa, especially when combining data from genes with very low rates of change (approaching invariant data) and very high rates (effectively randomized data). Wiens and Morrill (2011) found that in simulations utilizing rates and numbers of taxa more typical of empirical phylogenetic studies, adding characters with missing data tended to either increase or have little effect on mean accuracy for Bayesian phylogenetics. However, all three of these simulation studies were relatively simplistic. For example, none explored more realistic situations with multiple genes where gene topologies could potentially disagree. Nevertheless, discordance among gene trees is pervasive in empirical multi-locus datasets (Rokas et al., 2003; Cranston et al., 2009), especially when the underlying species topology includes one or more relatively short branches (e.g., Wiens et al., 2008).

Thus, a critical but unresolved question for empirical systematists is whether it is better to include or exclude genes that have some missing data. Specifically, do the benefits of increasing the number of genes outweigh the potential consequences of increasing the overall amount of missing data? This question is particularly relevant for short nodes that are difficult to resolve, nodes which may require the addition of many genes to resolve (e.g., Rokas et al., 2003) but for which gene topologies are especially likely to disagree (e.g., Wiens et al., 2008). Some empirical results on this issue were obtained by Wiens et al. (2005) and Cho et al. (2011), who both found that adding incomplete genes seemed to give more well-supported results that were more consistent with previous taxonomy and phylogenetic estimates (whereas excluding incomplete genes gave weaker support and/or relationships inconsistent with previous taxonomy and phylogenetic hypotheses). However, these authors did not perform detailed experiments examining the impact of missing data in the added genes.

In this study, we use analyses of real datasets to explore the consequences of including versus excluding genes with missing data on the accuracy of concatenated likelihood analyses. We use the similarity of the estimated trees to the phylogeny based on the complete data as a proxy for accuracy (which we define as the similarity of the estimated tree to the true phylogeny). We analyze data from yeast to represent datasets with many genes and extensive genetic divergence among taxa (despite most taxa being congeners in this case) and a dataset from plants representing those with fewer genes and more limited genetic divergence among taxa (despite many species being in different families). We analyze these datasets to address the following questions: (1) is accuracy of concatenated likelihood analyses increased or decreased by adding genes with missing data? (2) If adding genes with missing data is beneficial, is there a proportion of missing data at which adding these incomplete genes ceases to be useful? (3) If adding genes with missing data is detrimental, at what proportion of missing data does this occur? (4) How do the advantages and disadvantages of adding incomplete genes change with the overall number of genes in the analysis?

We also test for potential biases in branch length estimation caused by including versus excluding genes with missing data. Accurate branch-length estimates may be critically important for phylogenetic comparative analyses and for divergence-date estimation. Some authors have suggested that missing data can lead to strongly biased and inaccurate estimates of branch lengths (i.e., Lemmon et al., 2009), whereas other authors have suggested that those results may have been artifacts of the methods used by those authors (e.g., Wiens and Morrill, 2011; Roure et al., 2013). At least two recent studies have tested for biases in branch-length estimation caused by missing data in empirical datasets, and found no evidence for such biases (Pyron et al., 2011; Wiens and Tiu, 2012). Here, we explicitly contrast the impacts of including versus excluding genes with missing data on branch-length estimation, comparing these estimated branch lengths to those from the complete datasets with all sampled genes.

2. Materials and methods

2.1. Yeast data

2.1.1. Basic information on the yeast dataset

We selected an empirical dataset consisting of 8 yeast species (Rokas et al., 2003) and 106 orthologous genes. The dataset includes seven species of *Saccharomyces*, with a more distant relative (*Candida albicans*) included as an outgroup. There are very few missing data in the original data set (only 0.0063%). Separate analyses of each gene revealed considerable discordance among the estimated gene trees (Rokas et al., 2003). However, combining all genes yielded a single tree with 100% likelihood bootstrap values at every branch (Fig. 1; Rokas et al., 2003). The same topology was also found using a coalescent-based species-tree approach (BEST; Edwards et al., 2007). Therefore, we assumed that this tree reflects the true phylogenetic relationships among these eight species.

2.1.2. Design of missing data experiments

The overall design of the yeast experiments was as follows. First, we estimated a phylogeny for the complete data (106 genes). We then created smaller datasets by randomly sampling smaller numbers of genes (5, 10, 20, and 50), creating 100 new data matri-

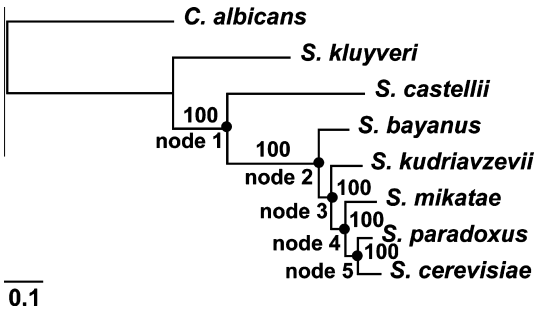

Fig. 1. Maximum likelihood estimate of phylogeny for 8 yeast species (seven species of *Saccharomyces* and an outgroup, *Candida albicans*) based on concatenated analysis of 106 genes (originally from Rokas et al., 2003), showing numbered nodes (for assessing accuracy) and bootstrap support.

ces (replicates) for each level of sampling. For each of these replicates, we then created new data matrices with different amounts of missing data, by replacing sequence data with missing data cells for selected genes (incomplete genes hereafter). These genes were made incomplete by randomly selecting certain taxa to have all their data for that gene replaced by missing data cells. We then created another set of matrices in which all the incomplete genes were excluded. For a given set of conditions, we then analyzed these matrices with maximum likelihood and compared the trees to the tree based on all 106 genes with no missing data. Specifically, we evaluated whether trees were more accurate (more similar to the tree from 106 genes) when they included some genes with missing data or had fewer genes but with no missing data. For example, one set of conditions would be based on a sample of 5 genes, of which 3 were selected to be incomplete, and these 3 incomplete genes would be made incomplete by replacing the sequencing data with missing data for 4 of the 8 species. We would then analyze the data matrices based on all 5 genes (including 3 with missing data) and also the data matrices with these three incomplete genes excluded (i.e. data matrices containing only 2 genes, with no missing data). The results for each set of conditions were based on the average results from 100 replicate matrices.

More specifically, we subsampled datasets with different number of genes (5, 10, 20, and 50), with genes randomly sampled without replacement from the total pool of 106 genes to yield each 8-taxon concatenated data matrix. Then, we replaced the data for selected genes in selected taxa to make different proportions of genes with missing data (20%, 40%, 60% and 80%) in the concatenated data matrices. Thus, given a total sample of 5 genes, either 1, 2, 3, or 4 genes would have missing data, given 10 genes, 2, 4, 6, or 8 genes would be incomplete, given 20 genes, 4, 8, 12, and 16 genes would be incomplete, and given 50 genes, 10, 20, 30, or 40 genes would be incomplete. These incomplete genes were randomly chosen from among those in the subsampled datasets. Then, for each gene that was selected to be incomplete, the gene was made incomplete by selecting different numbers of species to have missing data (2, 4, 6, and 7 species out of a total of 8). Thus, for each incomplete gene, ~25%, ~50%, ~75%, or ~87.5% missing data were introduced. The total amount of missing data in each matrix is therefore based on both the number of incomplete genes and the number of species with missing data in each gene. Species were randomly selected to have missing data, and sequences of these randomly selected species were entirely replaced by "?". During this process, we ensured that no species and no genes were completely replaced by missing data in the subsampled datasets; a gene with all missing data should be uninformative, and a species with all missing data cannot be placed in a phylogenetic analysis.

In the analyses described above, each incomplete gene had missing data in different, randomly selected taxa (as might occur when data are missing due to problems in obtaining sequence data). We also performed a limited set of analyses in which the same set of species had missing data across all incomplete genes (as might occur when assembling a supermatrix with data from previous studies with different sampling strategies), in order to evaluate whether these different distributions impact the results. The set of species lacking missing data across genes was randomly selected in each replicate. For these analyses, we analyzed conditions in which 50% of the species were incomplete (4 species with missing data), with few genes and many genes (10, 50), and in which 20% and 80% of the genes were incomplete. These analyses were not comprehensive, but were intended to represent a broad range of the overall conditions examined.

Finally, for the main analyses, we also evaluated how including versus excluding genes with missing data influenced estimates of branch lengths. Three different maximum likelihood trees were estimated for each set of conditions and each replicate (the complete data for a sample of genes, the analysis including genes with missing data, and the analysis excluding those genes with missing data). For each tree, we averaged the likelihood-estimated lengths of all branches to estimate the mean branch length for that tree. We then compared the mean branch lengths among the three classes of trees across the 100 replicates for each set of conditions. We tested for significant differences among these classes using ANOVA in SPSS 16.0. We note that this analysis specifically tests whether missing data consistently bias the estimates of branch lengths (e.g., consistently shorter or longer branch lengths given more missing data).

2.1.3. Assessment of accuracy

We estimated likelihood trees for: (1) the complete, subsampled datasets with 5, 10, 20, and 50 genes, (2) the subsampled datasets including genes with missing data, and (3) the subsampled datasets after excluding genes with missing data (see below for details of these analyses). We then compared these estimated trees to the "true" tree based on all 106 genes. This design allowed us to evaluate whether accuracy was increased by including or excluding genes with missing data.

In order to evaluate the effects of including or excluding genes with missing data, we used the ratio of the accuracy (proportion of correctly resolved nodes relative to the total number of nodes) of trees estimated from data matrices including incomplete genes (I_A) to the accuracy of trees estimated from data matrices excluding those incomplete genes (E_A). When this ratio is >1 the accuracy of trees from data matrices including incomplete genes is higher than that excluding incomplete genes. When the ratio equals 1, the trees from data matrices with and without missing data have equal accuracy. When the ratio is <1 the accuracy of trees from data matrices with missing data is lower than data matrices excluding incomplete genes. Estimated trees were fully resolved, and so each node was either correct or incorrect (in comparison to the "true" tree based on all the data).

Each treatment was repeated 100 times (i.e., 100 random subsamples of genes, each with a different random selection of genes and species that were incomplete). For a given set of conditions, accuracy was measured as the proportion of nodes in common between estimated trees and the concatenated tree of 106 genes, averaged across the 100 replicates for a given set of conditions. We first investigated accuracy across all five nodes and then considered accuracy for two difficult nodes (nodes 1 and 3 in Fig. 1; for evidence of their difficulty see Rokas et al., 2003). As an alternative measure of accuracy, we also estimated the percentage of replicates in which each method (including vs. excluding incomplete genes) estimated a topology that fully matched the correct topology based on all the data.

2.1.4. Maximum likelihood tree estimation

Choosing an appropriate partitioning scheme is an important issue in phylogenetic analysis (e.g., Lanfear et al., 2012). Furthermore, some evidence suggests that inaccurate results may be obtained from the combination of missing data and failure to partition (e.g., Lemmon et al., 2009). Therefore, prior to conducting the main analyses of our study (see above), we conducted a separate set of analyses designed to compare 4 different partitioning schemes: (1) no partitions, (2) partitioned by gene, (3) partitioned by codon position and (4) partitioned by gene and codon position. We used these analyses to both test for potential interactions between partitioning and missing data on accuracy and to evaluate partitioning strategies for subsequent analyses. We compared these 4 partitioning strategies in the following treatments: (1) datasets including the minimum proportion of missing data (20% of genes incomplete, with each incomplete gene containing 25% missing data), (2) datasets including the maximum proportion of

missing data (80% of genes incomplete, each incomplete gene with 87.5% of missing data), and (3) datasets excluding these incomplete genes. Each treatment was repeated 100 times for each partitioning scheme.

We first compared only the topologies estimated from each treatment. We found that the estimated topology varied among partitioning schemes in some replicates, depending on the treatments (Table S1). Topologies differed among partitioning schemes more often given fewer genes and more missing data, but patterns were similar regardless of whether the incomplete genes were included or excluded (Table S1). We also found that failing to partition did not make the incomplete genes misleading, because the results on the impact of including versus excluding incomplete genes were similar under different partitioning schemes (Table S2).

To find the overall best partitioning scheme, we then held the topology constant and compared the fit of the data under the different partitioning schemes. We first estimated a tree for the 106-gene concatenated matrix using each of the four partition schemes. All four yielded the same topology (Fig. 1). We then calculated the log likelihood of the data under each partitioning scheme using RAxML 7.2.6 (and the fixed topology). The best partitioning scheme was selected by calculating and comparing the Akaike information criterion (AIC; Akaike, 1974; Posada and Buckley, 2004). Partitioning by gene and codon gave the lowest, optimal AIC value (Table S3). But, as no partitioning was used in the original likelihood analysis of Rokas et al. (2003), and because we found that partitioning had little impact on our results regarding missing data (Table S2), we used the unpartitioned scheme in our main analyses. Furthermore, the unpartitioned analysis is (in some ways) the most conservative for our analyses, since previous studies (e.g., Lemmon et al., 2009) suggest that failing to partition may increase the negative impacts of missing data.

Each data matrix was analyzed with maximum likelihood in RAxML version 7.2.6 (Stamatakis, 2006; Stamatakis et al., 2008). All RAxML analyses employed the GTR + Γ model (general time reversible model with the gamma distribution of rates among sites) and used the "f-a" option to conduct a rapid bootstrap analysis with 100 replicates combined with 20 searches for the optimal tree. Note that the GTR model is the only substitution model used in RAxML, and is the most general model of sequence evolution (all other models are special cases of this model; Felsenstein, 2004). We used the gamma model (and not the gamma + invariant sites model) given that the large number of rate categories (25) used during tree searches should effectively encompass the invariant sites model.

2.2. Rosales data

2.2.1. Basic information on the Rosales dataset

We selected a multi-locus dataset from plants that included few missing data and provided a well-resolved phylogeny. Zhang et al. (2011) analyzed the phylogeny of Rosales with a total of 12 nuclear and plastid genes. They sampled 25 ingroup species to represent all nine families of Rosales and 13 outgroup species. Most family-level relationships in Rosales estimated by Zhang et al. (2011) are consistent with estimates from previous studies (e.g., Sytsma et al., 2002; Wang et al., 2009). Specifically, Rosaceae is sister to all other families in this order, and the other families are divided into two clades: (Ulmaceae, (Cannabaceae, (Moraceae, Urticaceae))), and (Rhamnaceae, (Elaeagnaceae, (Barbeyaceae, Dirachmaceae))). Our goal was to subsample genes from this "complete" dataset and introduce missing data experimentally. Therefore, to decrease missing data in the complete dataset, we excluded 2 genes which had missing data in more than 2 taxa (*rps*4 gene and 26S rDNA). We also excluded 4 taxa because these taxa had missing data in at least one gene (3 outgroup species: *Anisophyllea fallax*, *Polygala cruciata* and *Quillaja saponaria*; and one ingroup species *Dirachma socotrana*). After these deletions, only 1.05% missing data cells were present (including gaps) in the full data matrix.

The new, complete data matrix (10 genes × 34 taxa, Fig. 2) was then analyzed using maximum likelihood. In the complete, combined-data tree (Fig. 2), the only species of Dirachmaceae was excluded (see above), but the complete-data tree is otherwise identical to that estimated by Zhang et al. (2011). Furthermore, most of the relevant ingroup branches (nodes 1–7 and monophyly of families; see below) have relatively high bootstrap support (Fig. 2), with values of 100% for all nodes except node 7 (bootstrap = 76%). Therefore, we used this phylogenetic tree as the reference tree for assessing accuracy.

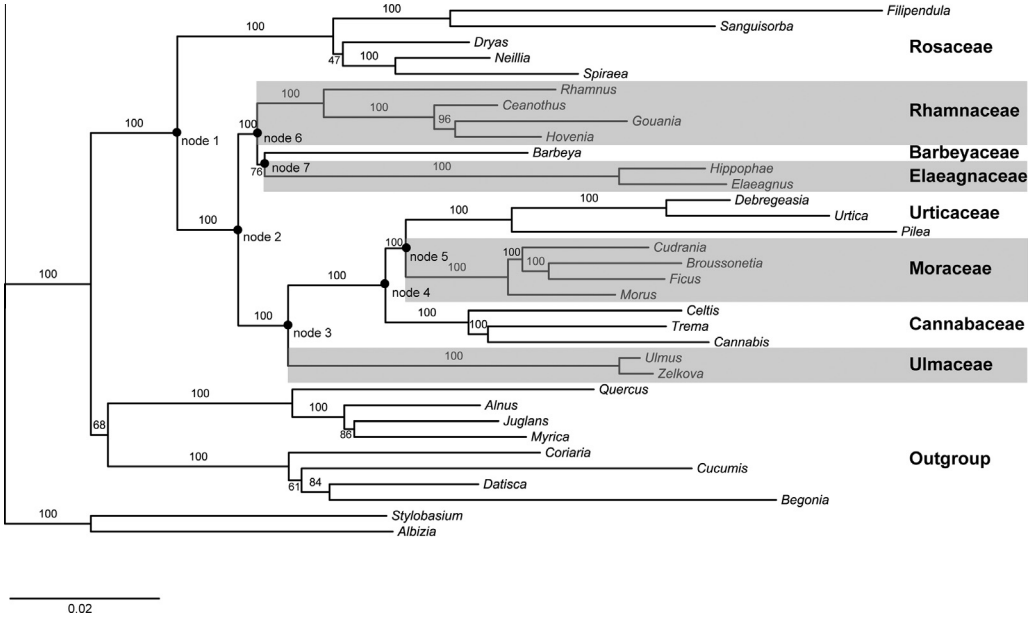

Fig. 2. Maximum likelihood estimate of phylogeny for Rosales from analysis of a reduced data matrix containing 10 genes and 34 taxa (full dataset from Zhang et al., 2011). Bootstrap values are shown above branches. Focal nodes (for measuring accuracy) are indicated below the branches.

We also analyzed each gene separately using maximum likelihood. However, *psb*BTNH consisted of four continuous genes and was treated as single locus (Soltis et al., 2011). The results showed discordance among the estimated gene trees (as in the yeast), especially over relationships between Moraceae, Urticaceae, and Cannabaceae and relationships between Elaeagnaceae and Barbeyaceae (Fig. S1).

2.2.2. Design of missing data experiments

The experimental design for the Rosales data was similar to that for the yeast. From the complete data matrix of 10 genes, we randomly selected 2 genes, 4 genes, and 6 genes to be incomplete and then separately introduced ~15%, ~30%, ~60%, ~90% missing data into these genes. To generate these different amounts of missing data, different numbers of taxa (5, 10, 20 and 30) were selected to have missing data for a given gene. For these randomly selected species, nucleotide sequence data were completely replaced by "?" for a given gene. During this process, we again ensured that no gene or species had its data completely replaced with missing cells.

As for the yeast data, we also performed a limited set of analyses in which the missing data cells were placed in the same taxa for all incomplete genes, to compare to the main results (in which incomplete species were randomly selected for each gene). Again, one set of incomplete species was randomly selected for all incomplete genes in each replicate. We selected an intermediate set of conditions to test (extensive missing data but not the maximum missing data), focusing on the case in which 4 of 10 genes were incomplete, with missing data in 20 of 34 taxa (~60%).

Finally, we tested whether missing data consistently bias estimates of branch lengths in the Rosales data. We compared the mean branch lengths for the tree with complete sampling of taxa and characters to trees estimated with genes with missing data included and excluded. Results were averaged across each set of 100 replicates and tested statistically using ANOVA.

2.2.3. Assessment of accuracy

All treatments were repeated 100 times. To test the impact of including versus excluding incomplete genes, the same index was used as above (ratio of mean accuracy of trees including and excluding incomplete genes). We compared estimated trees from subsampled data matrices with and without incomplete genes to the concatenated tree from the full data matrix of 10 genes with minimal missing data. Fourteen nodes of the concatenated tree were investigated (Fig. 2): 7 nodes support the monophyly of families, and the other 7 resolve relationships among them. Among the latter 7 nodes, the nodes relevant to relationships among Moraceae, Urticaceae, and Cannabaceae (node 5) and relationships among Rhamnaceae, Elaeagnaceae, and Barbeyaceae (node 7) were deemed as difficult nodes because of short branch lengths and discordance among genes trees (reviews in Sytsma et al., 2002; Zhang et al., 2011). Accuracy was estimated as the proportion of these 14 nodes shared between the estimated trees for the subsampled and completed data matrices. Accuracy for the two difficult nodes (5 and 7) was also estimated and summarized separately. As with the yeast data, an alternative approach for assessing accuracy was also used: we estimated the percentage of replicates in which each method (including vs. excluding incomplete genes) estimated a topology that fully matched the correct topology based on all the data.

2.2.4. Maximum likelihood tree estimation

The methods for likelihood analyses followed those described above. Specifically, we used RAxML version 7.2.6 (Stamatakis, 2006; Stamatakis et al., 2008), with the GTR + Γ model and using the "f-a" option to conduct a rapid bootstrap analysis with 100 replicates combined with 20 searches for the optimal tree for each data matrix. Prior to the main analyses, we also analyzed the complete Rosales data matrix using the four partitioning schemes described above. The tree topologies estimated from these partition schemes were almost the same, except for the position of *Dryas* within Rosaceae. The likelihoods and AIC for the four partitioning schemes were then estimated and compared using the tree topology from the unpartitioned analysis. The lowest AIC value was obtained from partitioning by both gene and codon (Table S3). However, we followed the original authors (Zhang et al., 2011) and performed our main analyses without partitions. Again, use of the unpartitioned analyses should make our results more conservative with respect to the negative impacts of missing data.

3. Results

3.1. Yeast results

The impact of including versus excluding genes with missing data depended on the amount of missing data (Fig. 3). Given ~25% missing data in the incomplete genes (missing data in 2 of 8 species), accuracy when including these genes was higher or equal to the accuracy excluding these incomplete genes (Fig. 3a), especially when 60% or 80% of the genes contain missing data and when less than 50 genes are sampled overall. Given ~50% missing data in the incomplete genes (Fig. 3b), accuracy also often increased when including the incomplete genes, especially when few genes were sampled overall and the proportion of genes with missing data was high. In contrast, including genes with ~75% missing data (missing data in 6 of 8 taxa) typically decreased accuracy compared to data matrices excluding these genes (Fig. 3c), although these decreases were consistently low (less than 2%). When there were 87.5% missing data in the incomplete genes (missing data in 7 of 8 taxa), adding these genes had little or no impact on accuracy (Fig. 3d). The overall results were generally similar when missing data were present in the same set of taxa across all incomplete genes, relative to the main results in which the incomplete taxa are randomly and independently selected for each gene (Table 1).

Accuracy was also analyzed separately for two difficult nodes (Fig. 4). The effects of including and excluding incomplete genes were consistent with the results for all nodes. However, adding incomplete genes caused much larger increases in accuracy for these less certain nodes (almost 30% in some cases; Fig. 4a) relative to the positive impacts on the overall dataset (less than 15%; Fig. 3), whereas the worst decreases associated with missing data were still low (less than 5%; Fig. 4c).

These conclusions were based on comparisons of average accuracy across replicates (proportion of correctly resolved nodes). As an alternative measure of accuracy, we also compared the percentage of replicates in which the estimated topologies matched the "true" topology (Table 2). Given ~25% and ~50% missing data in the incomplete genes, including incomplete genes yielded the correct topology more frequently than excluding them did. In contrast, given ~75% missing data, excluding the incomplete genes yielded the correct topology more frequently.

Including genes with missing data generally had little or no impact on mean branch lengths, both relative to analyses excluding these genes and relative to the complete datasets (Table 3). With a small number of loci (5 or 10 genes) there were some cases in which branch lengths were significantly longer when 80% of the genes had missing data and were included. However, these differences were still relatively small. Moreover, these were also cases in which excluding these genes also caused a similar, statistically significant increase in branch lengths. Thus, excluding incomplete genes did not appear to improve branch-length estimation relative to including them.

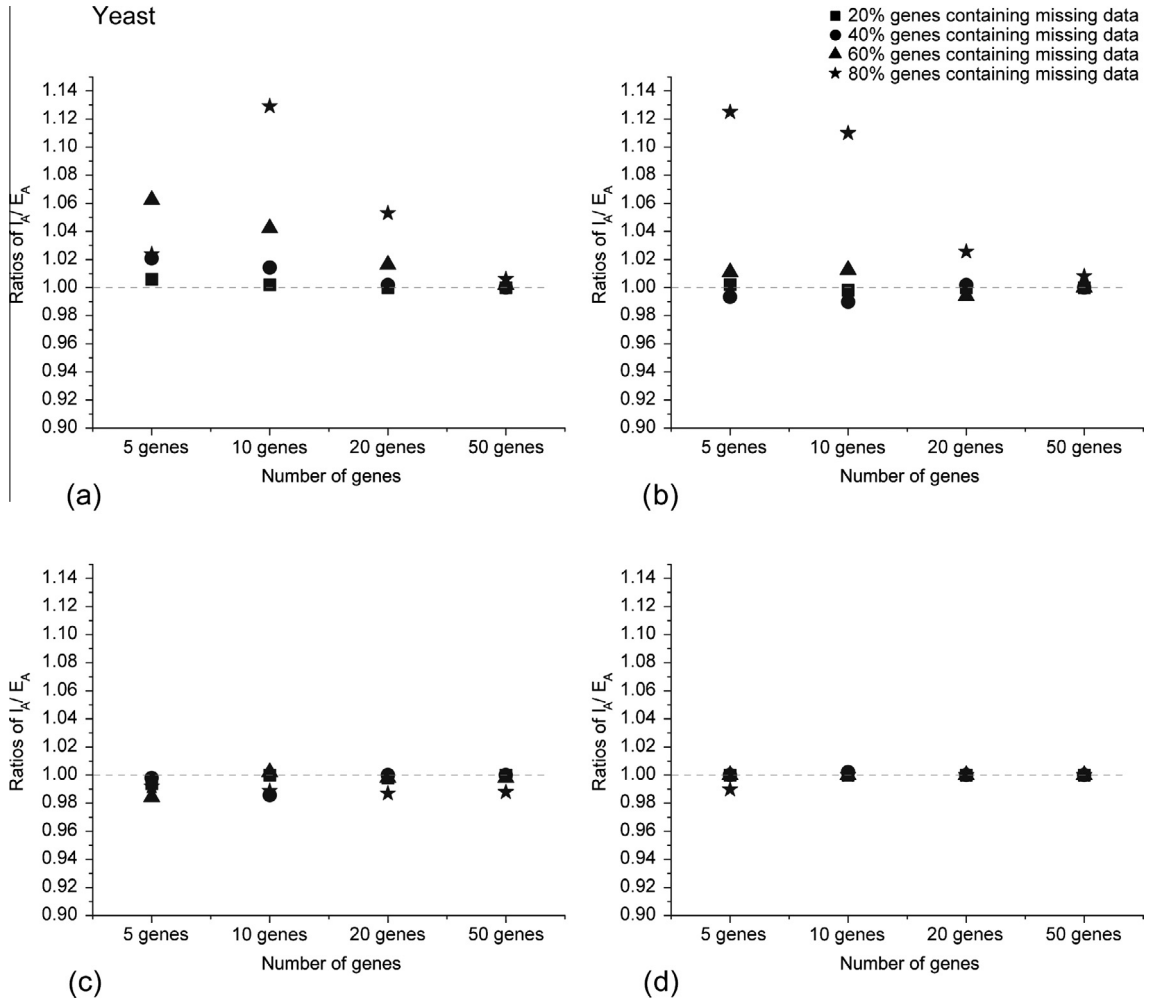

Fig. 3. Impacts of including versus excluding incomplete genes on phylogenetic accuracy for concatenated maximum likelihood analysis of multi-locus data for 8 species of yeast. Accuracy is based on all nodes. Ratios >1 indicate that including incomplete genes (I_A) increases accuracy relative to excluding these genes (E_A). Results are shown separately for (a) ~25%, (b) ~50%, (c) ~75%, and (d) 87.5% missing data in the incomplete genes in data matrices of four different sizes (total of 5, 10, 20, and 50 genes, when all genes are included). The four symbols indicate results when 20%, 40%, 60%, and 80% of the genes contain missing data.

Table 1
Comparison of two different ways of distributing missing data on the accuracy of maximum likelihood analyses including and excluding genes with missing data, for both the yeast and Rosales data. Missing data are either distributed in the same taxa across all incomplete genes, or else in a different set of randomly selected taxa for each gene. Most of the results of the study are based on the latter approach. The results here show that patterns of accuracy are generally similar using both approaches.

Total number of complete genes	Number of incomplete genes	Distribution of missing data	I_A: Accuracy including incomplete genes (%)	E_A: Accuracy excluding incomplete genes (%)	Ratio of I_A/E_A
Yeast					
10 Genes	2 Incomplete genes	4 Randomly selected taxa incomplete for all incomplete genes	99.0	99.4	0.9960
		4 Incomplete taxa randomly selected for each incomplete gene	98.0	98.2	0.9980
	8 Incomplete genes	4 Randomly selected taxa incomplete for all incomplete genes	87.8	84.3	1.0415
		4 Incomplete taxa randomly selected for each incomplete gene	90.6	81.2	1.1158
50 Genes	10 Incomplete genes	4 Randomly selected taxa incomplete for all incomplete genes	100.0	100.0	1.0000
		4 Incomplete taxa randomly selected for each incomplete gene	100.0	100.0	1.0000
	40 Incomplete genes	4 Randomly selected taxa incomplete for all incomplete genes	98.0	99.2	0.9879
		4 Incomplete taxa randomly selected for each incomplete gene	99.8	99.0	1.0081
Rosales					
10 Genes	4 Incomplete genes	20 Randomly selected taxa incomplete for all incomplete genes	97.5	96.0	1.0156
		20 Incomplete taxa randomly selected for each incomplete gene	98.5	96.1	1.0250

3.2. Rosales results

For the Rosales data, including incomplete genes generally increased accuracy relative to excluding them, both for all nodes and for the two difficult nodes (Fig. 5). Overall, including genes with missing data became much more beneficial than excluding them when a higher proportion of genes were incomplete (Fig. 5b and c). Conversely, there was less benefit to adding incom-

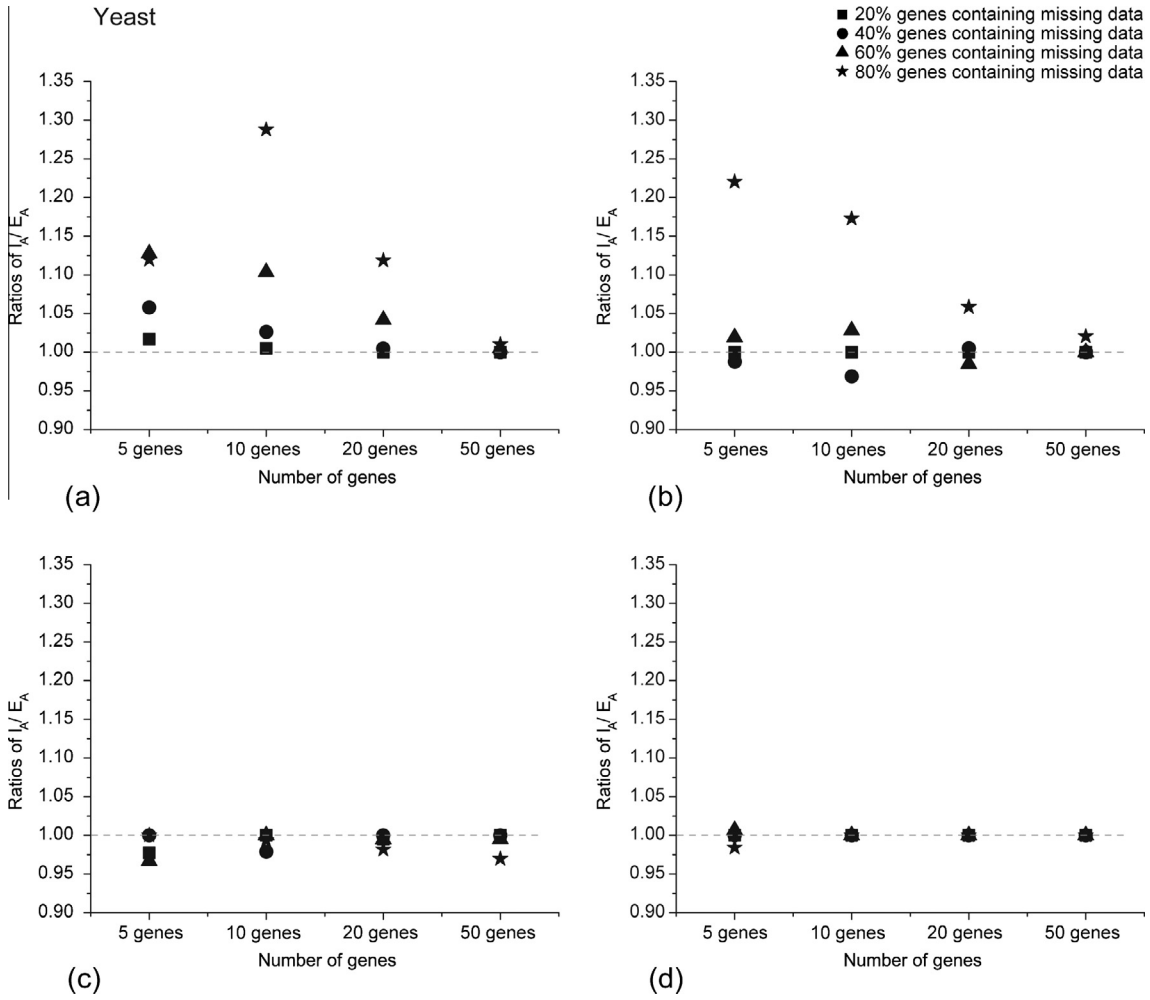

Fig. 4. Impacts of including versus excluding incomplete genes on phylogenetic accuracy for concatenated maximum likelihood analysis of multi-locus data for 8 species of yeast. Accuracy is based on nodes 1 and 3 only (Fig. 1). Ratios >1 indicate that including incomplete genes (I_A) increases accuracy relative to excluding these genes (E_A). Results are shown separately for (a) ~25%, (b) ~50%, (c) ~75% and (d) 87.5% missing data in the incomplete genes in data matrices of four different sizes (total of 5, 10, 20, and 50 genes, when all genes are included). The four symbols indicate results with 20%, 40%, 60%, and 80% of the genes containing missing data.

plete genes when relatively few of these genes were added (Fig. 5a), and when the genes that were added contained extensive missing data. The positive impacts of including genes with missing data were much stronger when considering only the two more difficult nodes (Fig. 5b and c). Different ways of distributing missing data cells appeared to have little impact on the results: results were similar when missing data were present in the same set of taxa across all incomplete genes, relative to the main analyses in which incomplete taxa were independently selected for each gene (Table 1). We also compared the number of replicates in which correct trees were estimated in the two treatments (with incomplete genes and without incomplete genes). We found that including incomplete genes yielded the correct topology more frequently than excluding them did under all conditions, although the numbers become more similar given fewer incomplete genes or more missing data per incomplete gene (Table 4). Comparison of mean branch lengths showed no significant differences between branch lengths based on the complete data and branch lengths estimated from datasets including incomplete genes (Table 5). However, there is a trend for mean branch lengths to be somewhat shorter with extensive missing data in these datasets, relative to the complete dataset and relative to analyses excluding these genes.

4. Discussion

Our study addresses a common question in phylogenetic analyses: should more genes be added to an analysis, even when these genes are missing data for some taxa? This question is likely to be encountered in almost every empirical phylogenetic study, and especially those using supermatrices or RAD sequence data (e.g., Rubin et al., 2012). However, relatively few studies have tested the potential impacts of adding genes with missing data on phylogenetic accuracy (e.g., Wiens, 1998; Wiens et al., 2005; Lemmon et al., 2009; Cho et al., 2011). Here, we address this question with experimental empirical analyses in two clades (yeast, Rosales). Overall, our results suggest that adding genes with some missing data seems to increase accuracy of concatenated likelihood analyses under many conditions, and only causes slight decreases under others. Interestingly, the distribution of missing data is critically important. Adding many genes with small amounts of missing data can dramatically increase accuracy relative to excluding these genes (despite the fact that these matrices have many missing data cells overall; Figs. 3a and b and 4a and b). On the other hand, when many taxa are missing data in the incomplete genes (i.e., 6 of 8 in yeast), adding many incomplete genes or few incomplete genes can both cause slight decreases in accuracy (Figs. 3c

Table 2
Percentage of replicates (out of 100) in which the estimated maximum likelihood topology fully matches the "true" topology, comparing results both including incomplete genes (I) and excluding incomplete genes (E), using the data from yeast.

Proportion of incomplete genes	5 Genes			10 Genes			20 Genes			50 Genes		
	Missing data (%)	Percent correct replicates (I)	Percent correct replicates (E)	Missing data (%)	Percent correct replicates (I)	Percent correct replicates (E)	Missing data (%)	Percent correct replicates (I)	Percent correct replicates (E)	Missing data (%)	Percent correct replicates (I)	Percent correct replicates (E)
20% Incomplete genes	25.00	84	81	24.93	99	98	25.23	100	100	24.86	100	100
	50.00	68	66	49.64	92	92	50.12	100	100	49.58	100	100
	75.00	76	80	75.18	97	97	75.08	99	100	75.04	100	100
	87.50	83	83	87.50	93	93	87.50	100	100	87.50	100	100
40% Incomplete genes	24.63	83	74	24.88	94	89	24.93	100	99	25.07	100	100
	50.23	66	69	50.43	87	93	50.07	99	98	50.20	100	100
	75.26	79	79	75.04	87	91	74.67	100	100	74.46	100	100
	87.50	69	69	87.50	88	88	87.50	98	98	87.50	100	100
60% Incomplete genes	24.90	69	60	25.09	92	76	24.83	99	91	25.03	100	98
	50.04	64	61	50.05	83	80	50.22	96	99	50.29	100	100
	75.09	55	61	74.75	75	74	75.25	92	93	74.97	98	99
	87.50	56	55	87.50	79	79	87.50	95	95	87.50	100	100
80% Incomplete genes	25.29	45	39	24.86	81	52	25.08	98	79	24.93	100	98
	49.61	49	35	49.88	60	46	49.92	82	74	50.15	99	95
	74.76	32	32	75.13	61	63	74.80	64	67	75.07	92	98
	87.50	41	41	87.50	60	60	87.50	74	74	87.50	96	96

and 4c). These results are broadly concordant with simulation results focusing on a smaller number of characters and using parsimony (Wiens, 1998). They also support previous empirical studies using model-based analyses of molecular data, suggesting that adding incomplete genes can improve support and congruence, despite their missing data (e.g., Wiens et al., 2005; Cho et al., 2011).

In yeast, we found some conditions under which there was a decrease in accuracy. However, these decreases were typically small. Most importantly, these involved cases where only 2 species had non-missing data (Figs. 3c and 4c). Under these conditions, there is no a priori reason to expect the added genes to contribute positively to the analyses. It appears that the particular percentage of missing data is not actually problematic. For example, in Rosales, genes with >90% missing data still cause an average increase in accuracy, even though the increase is slight (Fig. 5). Nevertheless, an important question is why the missing data should cause decreases at all. It makes intuitive sense that there may be issues of long-branch attraction when few taxa have non-missing data (e.g., Wiens, 1998), and this may explain some cases of slight decreases when 4 of 8 taxa have non-missing data in incomplete genes in yeast (Figs. 3b and 4b). However, this may not explain the pattern when only 2 of 8 taxa have non-missing data. In this case, it appears that these genes should have no impact on accuracy, as we find when 7 of 8 taxa have missing data (Figs. 3d and 4d). Nevertheless we find slight decreases in mean accuracy, even when many genes are sampled overall (i.e., 50; Fig. 3c). These results may be consistent with those of Lemmon et al. (2009), who also found that adding seemingly uninformative genes with extensive missing data might negatively impact accuracy. The underlying cause is not entirely clear, but may be related to distorted branch lengths or distorted parameter estimates (Lemmon et al., 2009; Roure et al., 2013). However, this still begs the question of why one would add such genes to a phylogenetic analysis in the first place, given that there is no reason to expect them to contribute positively. Clearly, the easiest solution is to simply not add such characters. This solution makes the question of why exactly this happens of limited practical interest.

We also looked at a question that few authors have addressed before: how does the inclusion versus exclusion of incomplete genes impact poorly resolved nodes in particular? We found that the impacts on poorly supported nodes mirrored those for the overall data set, but were generally more dramatic. Thus, for the yeast data, increases in accuracy from including incomplete genes were generally greater for the poorly supported nodes under those conditions where adding incomplete genes increased accuracy (Figs. 3 and 4). Conversely, the decreases were greater for those conditions under which there were decreases. But again, the decreases were never >5% and increases could be >20% (on average). Similarly, for the Rosales data (in which adding incomplete genes only increased or had no effect on mean accuracy), increases in accuracy were dramatically higher when considering the two poorly resolved nodes only (Fig. 5). An obvious interpretation of these results is that these poorly resolved nodes are the ones most likely to be impacted either positively or negatively by the inclusion or exclusion of genes with missing data (other branches are simply unchanged, one way or the other). These results for poorly resolved nodes further confirm our general conclusions that (at least for these datasets) adding incomplete genes tends to either increase accuracy or else have a negligible impact on accuracy (whether positive, negative, or none).

Our results also address how missing data impact branch length estimation. Overall, we find little evidence that missing data strongly bias these estimates. For the yeast data, branch lengths were sometimes significantly longer when genes with missing data were included (i.e. when few genes were sampled overall and missing data were extensive; Table 3). However, excluding the

Table 3
Comparison (for the yeast data) of mean branch lengths (across all branches in each tree) averaged across all 100 replicates for each set of conditions, with branch lengths from trees from the complete data, from those including the incomplete genes (I), and from those excluding these genes (E).

	5 Genes				10 Genes			
Proportion incomplete genes	Missing data (%)	Branch length (complete data)	Branch length (I)	Branch length (E)	Missing data (%)	Branch length (complete data)	Branch length (I)	Branch length (E)
20% Incomplete genes	25	0.2207	0.2202	0.2236	24.93	0.2193	0.2183	0.2164
	50	0.2325	0.2328	0.2370	49.64	0.2235	0.2237	0.2243
	75	0.2234	0.2229	0.2231	75.18	0.2158	0.2196	0.2187
	87.5	0.2306	0.2340	0.2340	87.5	0.2235	0.2251	0.2251
40% Incomplete genes	24.63	0.2330	0.2344	0.2486	24.88	0.2174	0.2182	0.2162
	50.23	0.2390	0.2470	0.2476	50.43	0.2199	0.2210	0.2315
	75.26	0.2225	0.2249	0.2269	75.04	0.2217	0.2267	0.2277
	87.5	0.2299	0.2491	0.2491	87.5	0.2213	0.2222	0.2222
60% Incomplete genes	24.9	0.2177	0.2244	0.2451*	25.09	0.2174	0.2187	0.2336*
	50.04	0.2142	0.2145	0.2187	50.05	0.2209	0.2184	0.2270
	75.09	0.2228	0.2360	0.2371	74.75	0.2214	0.2250	0.2329
	87.5	0.2264	0.2283	0.2287	87.5	0.2258	0.2276	0.2276
80% Incomplete genes	25.29	0.2305	0.2264	0.2597*	24.86	0.2196	0.2234	0.2495*
	49.61	0.2294	0.2368	0.2590*	49.88	0.2198	0.2129	0.2341
	74.76	0.2247	0.2641*	0.2843*	75.13	0.2158	0.2313	0.2418*
	87.5	0.2244	0.2798*	0.2803*	87.5	0.2236	0.2543*	0.2554*
	20 Genes				50 Genes			
20% Incomplete genes	25.23	0.2185	0.2187	0.2202	24.86	0.2152	0.2145	0.2162
	50.12	0.2157	0.2145	0.2174	49.58	0.2155	0.2149	0.2152
	75.08	0.2131	0.2148	0.2152	75.04	0.2143	0.2134	0.2137
	87.5	0.2131	0.2141	0.2141	87.5	0.2149	0.2143	0.2143
40% Incomplete genes	24.93	0.2123	0.2118	0.2151	25.07	0.2147	0.2135	0.2170
	50.07	0.2173	0.2178	0.2210	50.2	0.2138	0.2159	0.2182
	74.67	0.2168	0.2184	0.2194	74.46	0.2099	0.2059	0.2043*
	87.5	0.2166	0.2145	0.2145	87.5	0.2147	0.2152	0.2152
60% Incomplete genes	24.83	0.2176	0.2156	0.2262	25.03	0.2154	0.2136	0.2208
	50.22	0.2177	0.2168	0.2239	50.29	0.2159	0.2164	0.2188
	75.25	0.2180	0.2217	0.2242	74.97	0.2154	0.2192	0.2208
	87.5	0.2139	0.2215	0.2218	87.5	0.2136	0.2173	0.2175
80% Incomplete genes	25.08	0.2167	0.2135	0.2322*	24.93	0.2159	0.2127	0.2119
	49.92	0.2151	0.2115	0.2242	50.15	0.2142	0.2154	0.2232*
	74.8	0.2149	0.2245	0.2248	75.07	0.2159	0.2152	0.2119
	87.5	0.2141	0.2227	0.2238	87.5	0.2140	0.2190	0.2200

* Mean branch lengths that differ significantly from the branch lengths for the complete data under these conditions.

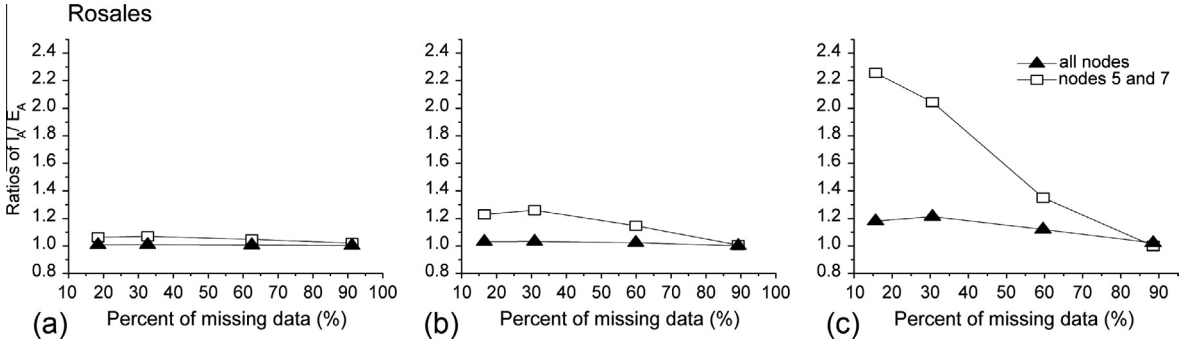

Fig. 5. Impacts of including versus excluding incomplete genes on the accuracy of concatenated maximum likelihood analysis for 10 nuclear and plastid genes from Rosales, with accuracy analyzed for all 14 nodes (□) and two difficult nodes only: nodes 5 and 7 (▲) in Fig. 2. Ratios >1 indicate that including incomplete genes (I_A) increases accuracy relative to excluding these genes (E_A). The *x*-axis shows four different levels of missing data (~15%, ~30%, ~60%, and 90%) in the data matrices. Each point is the average from 100 replicates. Results are shown separately when (a) 2, (b) 4, and (c) 6 of the 10 genes are incomplete.

incomplete genes under these conditions actually led to significantly longer branch lengths more frequently (Table 3). For the Rosales data, branch lengths were not significantly different from the complete data when genes with missing data were included, although there did appear to be a non-significant trend for mean branch lengths to be somewhat underestimated given extensive missing data. These results are consistent with other empirical studies in showing no consistent, significant biases in estimated branch lengths from data matrices with extensive missing data (e.g. Pyron et al., 2011; Wiens and Tiu, 2012). Nevertheless, further study of this issue would be desirable. We also acknowledge that this lack of consistent bias does not guarantee that branch lengths will be correctly estimated in each case (e.g., underestimates and overestimates may balance each other out across different branches on the tree and/or across different replicates, leading to no significant bias in one direction or the other).

Table 4
Percent replicates (out of 100) in which the estimated topology is the "true" topology, both including incomplete genes and excluding incomplete genes, using the data from Rosales.

Number of incomplete genes	Percent missing data (%)	Percent replicates estimating correct tree (including incomplete genes)	Percent replicates estimating correct tree (excluding incomplete genes)
2 Incomplete genes	18.43	100	88
	32.67	99	86
	62.46	97	88
	91.17	91	87
4 Incomplete genes	16.49	98	62
	30.93	89	53
	59.88	81	59
	89.23	63	62
6 Incomplete genes	15.65	94	11
	30.56	84	11
	59.62	40	16
	88.52	8	6

Table 5
Comparison (for the Rosales data) of mean branch lengths (across all branches in each tree) averaged across all 100 replicates for each set of conditions, including trees from the complete data, including the incomplete genes with missing data, and excluding these genes.

Number of incomplete genes	Missing data (%)	Branch length (complete data)	Branch length (including incomplete genes)	Branch length (excluding incomplete genes)
2 Incomplete genes	18.43	0.01814	0.01808	0.01838
	32.67	0.01814	0.01795	0.01788
	62.46	0.01814	0.01793	0.01833
	91.17	0.01814	0.01784	0.01804
4 Incomplete genes	16.49	0.01814	0.01798	0.01806
	30.93	0.01814	0.01780	0.01795
	59.88	0.01814	0.01756	0.01850
	89.23	0.01814	0.01782	0.01858
6 Incomplete genes	15.65	0.01814	0.01790	0.01898
	30.56	0.01814	0.01763	0.01745
	59.62	0.01814	0.01694	0.01808
	88.52	0.01814	0.01690	0.01772

We think that our results provide useful insights on an important and widespread issue in phylogenetics. Nevertheless, our study also has several limitations that should be mentioned. First, because we are using empirical data sets from the natural world, the true phylogenies of these groups are not actually known. However, it seems reasonable that deviations from the tree based on the complete data most likely represent errors, and this general approach of treating the tree from the complete data as correct is widely used in other studies of missing data (e.g., Philippe et al., 2004; Fulton and Strobeck, 2006; Burleigh et al., 2009; Rubin et al., 2012; Wiens and Tiu, 2012; Roure et al., 2013). Furthermore, our results are broadly concordant with those from simulations. Second, a similar issue with empirical data sets is that it is generally not possible to examine the full range of conditions that can potentially be encountered in other empirical data sets. Third, we acknowledge that our results are based on only two datasets, and very different results might be obtained in other empirical studies. However, similar results have been obtained in vertebrates (Wiens et al., 2005) and insects (Cho et al., 2011). We also note that our two datasets span very different sets of branch lengths, with relatively long branches in yeast and shorter branches in Rosales.

Many other studies now suggest that data matrices with large amounts of missing data are not generally misleading (e.g., Wiens and Morrill, 2011; Rubin et al., 2012; Roure et al., 2013), even if they did not examine the costs and benefits of adding genes with missing data as we did. Nevertheless, additional studies of this question would still be useful. In general, we predict that the greatest benefits to including genes with missing data will be seen when trees are poorly supported due to a paucity of genes being sampled, and when the genes added have enough non-missing data to at least be informative (>3 taxa). We expect greater benefits from adding genes with less missing data, even though adding many incomplete genes may lead to including large amounts of missing data overall. Clearly, the added genes should also have similar (or higher) levels of phylogenetic information, and equal or lesser homoplasy (i.e., adding incomplete genes may not be helpful if the added genes are problematic for other reasons besides their missing data).

We also emphasize that there are several related questions that should also be addressed in future studies. We focused only on concatenated analyses, and many analyses now estimate trees from multi-locus data using explicit species-tree methods (e.g., Edwards et al., 2007; Heled and Drummond, 2010). More studies on the impact of missing data on species-tree methods are needed (e.g., Hovmöller et al., 2013), especially the question of whether to include or exclude genes with missing data. Even for concatenated analyses, analyses of the impacts of missing data on estimates of branch support (e.g., bootstrapping, Bayesian posterior probabilities) are also needed (e.g., de la Torre-Bárcena et al., 2009). The impacts of including vs. excluding genes with missing data should also be tested for methods for estimating divergence dates (e.g., the Bayesian uncorrelated lognormal approach in BEAST; Drummond et al., 2006).

Finally, a critically important point that should be noted from our study is that, based on our results, excluding genes with missing data can potentially lead to less accurate phylogenies than can be obtained from including these characters. In fact, these decreases from excluding these incomplete genes appear to be

both more common and of greater impact than potential increases from excluding them. Therefore, we see no basis for treating the exclusion of genes with missing data as a necessarily safer or more conservative approach. We see no argument for why an approach that seems to decrease accuracy under realistic conditions should be considered safer or more conservative.

Acknowledgments

This work was supported by National Key Basic Research Program of China (Grant No. 2014CB954100), Key Research Program of the Chinese Academy of Sciences (Grant No. KJZD-EW-L07), the National Natural Science Foundation of China (Grant No. 40830209) and The Science and Technology Program of Yunnan Province (Grant No. 2008GA029). We thank Lei Zhao of the Kunming Institute of Botany for assistance in running some of the analyses. We thank Rokas Antonis and Shu-Dong Zhang for their kindly providing their datasets.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2014.08.006.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Aut. Control 19, 716–723.
Burleigh, J.G., Hilu, K., Soltis, D., 2009. Inferring phylogenies with incomplete data sets: a 5-gene, 567-taxon analysis of angiosperms. BMC Evol. Biol. 9, 61.
Cho, S., Zwick, A., Regier, J.C., Mitter, C., Cummings, M.P., Yao, J., Du, Z., Zhao, H., Kawahara, A.Y., Weller, S., 2011. Can deliberately incomplete gene sample augmentation improve a phylogeny estimate for the advanced moths and butterflies (Hexapoda: Lepidoptera)? Syst. Biol. 60, 782–796.
Cranston, K.A., Hurwitz, B., Ware, D., Stein, L., Wing, R.A., 2009. Species trees from highly incongruent gene trees in rice. Syst. Biol. 58, 489–500.
de la Torre-Bárcena, J.E., Kolokotronis, S.-O., Lee, E.K., Stevenson, D.W., Brenner, E.D., Katari, M.S., Coruzzi, G.M., DeSalle, R., 2009. The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data. PLoS ONE 4, e5764.
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.
Edwards, S.V., Liu, L., Pearl, D.K., 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA 104, 5936–5941.
Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.
Fulton, T.L., Strobeck, C., 2006. Molecular phylogeny of the Arctoidea (Carnivora): effect of missing data on supertree and supermatrix analyses of multiple gene data sets. Mol. Phylogenet. Evol. 41, 165–181.
Heled, J., Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27, 570–580.
Hovmöller, R., Knowles, L.L., Kubatko, L.S., 2013. Effects of missing data on species tree estimation under the coalescent. Mol. Phylogenet. Evol. 69, 1057–1062.
Huelsenbeck, J.P., 1991. When are fossils better than extant taxa in phylogenetic analysis? Syst. Biol. 40, 458–469.
Lanfear, R., Calcott, B., Ho, S.Y.W., Guindon, S., 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29, 1695–1701.
Lemmon, A.R., Brown, J.M., Stanger-Hall, K., Lemmon, E.M., 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58, 130–145.
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.
Posada, D., Buckley, T.R., 2004. Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53, 793–808.
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.
Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.
Roure, B., Baurain, D., Philippe, H., 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30, 197–214.
Rubin, B.E.R., Ree, R.H., Moreau, C.S., 2012. Inferring phylogenies from RAD sequence data. PLoS ONE 7, e33394.
Sanderson, M.J., McMahon, M.M., Steel, M., 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evol. Biol. 10, 155.
Sanderson, M.J., McMahon, M.M., Steel, M., 2011. Terraces in phylogenetic tree space. Science 333, 448–450.
Schaefer, H., Renner, S.S., 2011. Phylogenetic relationships in the order Cucurbitales and a new classification of the gourd family (Cucurbitaceae). Taxon 60, 122–138.
Soltis, D.E., Smith, S.A., Cellinese, N., Wurdack, K.J., Tank, D.C., Brockington, S.F., Refulio-Rodriguez, N.F., Walker, J.B., Moore, M.J., Carlsward, B.S., et al., 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Am. J. Bot. 98, 704–730.
Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst. Biol. 57, 758–771.
Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.
Sytsma, K.J., Morawetz, J., Pires, J.C., Nepokroeff, M., Conti, E., Zjhra, M., Hall, J.C., Chase, M.W., 2002. Urticalean rosids: circumscription, rosid ancestry, and phylogenetics based on rbcL, trnL-F, and ndhF sequences. Am. J. Bot. 89, 1531–1546.
Wang, H., Moore, M., Soltis, P., Bell, C., Brockington, S., Alexandre, R., Davis, C., Latvis, M., Manchester, S., Soltis, D., 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc. Natl. Acad. Sci. USA 106, 3853–3858.
Wiens, J.J., Moen, D., 2008. Missing data and the accuracy of Bayesian phylogenetics. J. Syst. Evol. 46, 307–314.
Wiens, J.J., Morrill, M.C., 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst. Biol. 60, 719–731.
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., Tiu, J., 2012. Highly incomplete taxa can rescue phylogenetic analyses from the negative impacts of limited taxon sampling. PLoS ONE 7, e42925.
Wiens, J.J., Fetzner Jr, J.W., Parkinson, C.L., Reeder, T.W., 2005. Hylid frog phylogeny and sampling strategies for speciose clades. Syst. Biol. 54, 778–807.
Wiens, J.J., Kuczynski, C.A., Smith, S.A., Mulcahy, D.G., Sites Jr., 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., Townsend, T., Reeder, T.W., Mulcahy, D.G., Sites, J.W., 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., 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: is there a 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.
Zhang, S.D., Soltis, D.E., Yang, Y., Li, D.Z., Yi, T.S., 2011. Multi-gene analysis provides a well-supported phylogeny of Rosales. Mol. Phylogenet. Evol. 60, 21–28.
Zwick, A., Regier, J.C., Mitter, C., Cummings, M.P., 2011. Increased gene sampling yields robust support for higher-level clades within Bombycoidea (Lepidoptera). Syst. Entomol. 36, 31–43.