Molecular Phylogenetics and Evolution xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species ☆

Yuchi Zheng [a,b], John J. Wiens [b,*]

[a] Department of Herpetology, Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu 610041, China
[b] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721-088, USA

ARTICLE INFO

Article history:
Received 26 May 2015
Revised 30 September 2015
Accepted 8 October 2015
Available online xxxx

Keywords:
Missing data
Phylogenomic
Phylogeny
Squamata
Supermatrix

ABSTRACT

Two common approaches for estimating phylogenies in species-rich groups are to: (i) sample many loci for few species (e.g. phylogenomic approach), or (ii) sample many species for fewer loci (e.g. supermatrix approach). In theory, these approaches can be combined to simultaneously resolve both higher-level relationships (with many genes) and species-level relationships (with many taxa). However, fundamental questions remain unanswered about this combined approach. First, will higher-level relationships more closely resemble those estimated from many genes or those from many taxa? Second, will branch support increase for higher-level relationships (relative to the estimate from many taxa)? Here, we address these questions in squamate reptiles. We combined two recently published datasets, one based on 44 genes for 161 species, and one based on 12 genes for 4161 species. The likelihood-based tree from the combined matrix (52 genes, 4162 species) shared more higher-level clades with the 44-gene tree (90% vs. 77% shared). Branch support for higher level-relationships was marginally higher than in the 12-gene tree, but lower than in the 44-gene tree. Relationships were apparently not obscured by the abundant missing data (92% overall). We provide a time-calibrated phylogeny based on extensive sampling of genes and taxa as a resource for comparative studies.

1. Introduction

The sequence data that are available to resolve phylogenies are becoming increasingly abundant, but also increasingly heterogeneous across taxa. For example, large-scale sequencing projects have been undertaken for many important clades, in which hundreds of loci are sequenced (e.g. Dunn et al., 2008; Regier et al., 2010; Kocot et al., 2011; Chiari et al., 2012; Struck et al., 2011; Smith et al., 2012; Jarvis et al., 2014; Weigert et al., 2014). Yet, at the same time, most species in many clades may still have data for no more than a few genes each (see below). How one should deal with this heterogeneity when addressing higher-level relationships is a critical but unresolved question in phylogenetics. One extreme approach is to focus on resolving higher-level relationships by sampling large numbers of genes in a relatively small sample of species: this is the common approach in phylogenomic studies (e.g. Chiari et al., 2012; Jarvis et al., 2014; Weigert et al., 2014). Another extreme approach is to include hundreds or thousands of species but for a smaller number of genes: this is the typical approach in supermatrix studies (e.g. Pyron and Wiens, 2011; Jetz et al., 2012; Pyron et al., 2013). Whether it is better to sample more taxa or more genes to resolve higher-level phylogenies was once a major debate in systematics (e.g. Graybeal, 1998; Rannala et al., 1998; Poe and Swofford, 1999; Rosenberg and Kumar, 2001; Zwickl and Hillis, 2002; Poe, 2003; Heath et al., 2008). Although the topic is less hotly debated now, the question is still highly relevant and very much unresolved.

It is also possible to combine these two extreme strategies (e.g. Wiens et al., 2005). For example, if data are available for many genes for a few taxa and for many taxa for a few genes, one could combine all these data in the same matrix. Although such a matrix would have extensive missing data for most taxa for most genes, the many genes would (in theory) provide better support for higher-level relationships, while still resolving all relationships at the species level. A few studies have utilized this combined approach, but so far incorporating relatively limited numbers of genes and/or taxa. For example, Wiens et al. (2005) applied this

☆ This paper was edited by the Associate Editor K.H. Kozak.

* Corresponding author.
E-mail addresses: zhengyc@cib.ac.cn (Y. Zheng), wiensj@email.arizona.edu (J.J. Wiens).

http://dx.doi.org/10.1016/j.ympev.2015.10.009
1055-7903/

approach to hylid frogs, whereas Cho et al. (2011) utilized this approach in lepidopteran insects. The findings of these studies did not support the idea that the extensive missing data necessarily generated misleading results. Wiens et al. (2005), for example, found that support values for the placement of individual species in the combined analysis was strongly related to the support for their placement in separate analyses of the most widely sampled gene among species (mitochondrial ribosomal 12S). In contrast, support was not related to the amount of missing data that these species contained in the combined matrix. These results are supported by simulation studies, which suggest that the mere presence of missing data may not itself be misleading, beyond the reduced amount of data included (e.g. Wiens, 2003; Philippe et al., 2004; Wiens and Morrill, 2011).

What is less clear is whether this combined approach is necessarily beneficial. Both Wiens et al. (2005) and Cho et al. (2011) presented anecdotal observations suggesting that estimates from the combined approach were preferable to those from their analyses with more taxa but fewer genes. Wiens et al. (2005) noted that some higher taxa were non-monophyletic in trees based on many taxa but few genes, but that these taxa were supported as monophyletic when all genes were included (even though these genes had non-missing data in only some species). Cho et al. (2011) found that the topology was almost identical after adding the incompletely sampled genes, but stated that support values seemed to improve for deeper nodes (although without explicit statistical tests).

We argue that some fundamental questions about this combined approach have yet to be addressed. First, will relationships estimated by the combined approach more closely resemble those from the analyses of many genes with few taxa or those from many taxa but few genes? It seems logical that higher-level relationships would be determined primarily by the larger set of genes in the well-sampled taxa. However, if some aspects of the higher-level relationships were not accurately reconstructed (e.g. due to long-branch attraction caused by limited taxon sampling), adding more taxa could improve estimation (e.g. Poe, 2003), and potentially overturn the relationships based on more genes. Simulations and empirical analyses show that even highly incomplete taxa can potentially subdivide long branches and lead to more accurate phylogenetic estimates when there is long-branch attraction due to limited taxon sampling (e.g. Wiens, 2005; Wiens and Tiu, 2012; Roure et al., 2013).

Second, will adding the set of many incomplete genes improve branch support (e.g. bootstrap values) for the higher-level relationships, relative to the analysis with many taxa but few genes? This idea was suggested by Wiens et al. (2005) and Cho et al. (2011), but not explicitly tested. The idea that adding genes with data in only some taxa can improve accuracy has been supported in recent analyses (e.g. Jiang et al., 2014; Zheng and Wiens, 2015). Thus, one might expect bootstrap support to also increase. On the other hand, bootstrap support for higher-level relationships might be limited by the uncertain placement of some taxa. If these taxa are not among those in the many-genes dataset, then adding genes may have only limited positive impacts on bootstrap support. We note that bootstrap support from concatenated analyses is not a substitute for accuracy. However, some evidence suggests that concatenated bootstrap support can be correlated with support from species-tree analyses, and with accuracy (Streicher et al., in press).

Here, we test these ideas with empirical data in squamate reptiles. Squamate reptiles include all lizards and snakes, with more than 9700 species currently recognized (Uetz and Hošek, 2015). Two recent studies have generated two very different points in a continuum of potential sampling strategies for genes and taxa across this group. First, Wiens et al. (2012) generated a dataset of 44 nuclear protein-coding genes for 161 squamate species, representing most families and subfamilies. The data matrix was 84% complete (i.e. some genes were lacking in some taxa), and included a total alignment of 33,717 base pairs. Second, Pyron et al. (2013) compiled a dataset of GenBank sequences for 12 genes for 4161 squamate species including all families. The data included 7 nuclear and 5 mitochondrial genes (12,896 aligned base pairs), but the matrix was only 19% complete. Four nuclear genes were shared between the two datasets (BDNF, NT3, R35, and RAG1).

We combined these two datasets, analyzed the data with maximum likelihood (RAxML; Stamatakis, 2006), and then asked the following questions. First, does the tree from the combined dataset share more nodes with the tree from extensive sampling of genes (Wiens et al., 2012) or from that based on extensive sampling of species (Pyron et al., 2013)? In other words, is greater sampling of genes or taxa more important in determining the higher-level relationships in the combined analysis? Second, does adding the set of incomplete genes for a limited set of taxa actually increase support values across the tree, and more specifically, for the higher-level relationships that are addressed in both of the separate datasets? Overall, our study further addresses the utility of this combined-data approach.

In addition to addressing these general questions, we also present the most extensive analysis of squamate phylogeny to date (in terms of sampling of both genes and species). We also time calibrate this tree, to make it more useful for a broad range of comparative studies, including studies of biogeography, diversification, and rates of character evolution.

2. Materials and methods

2.1. Combining datasets

The two datasets were combined using SequenceMatrix version 1.7.8 (Vaidya et al., 2011). When the same gene was included in both datasets, this program retained the longest sequence for each gene for each taxon. For RAG1, the gene fragment used in the 171-taxon dataset was nested within the much longer fragment used in the 4162-taxon dataset (although most taxa lacked data for this gene, or were sequenced for only a small portion of it). Therefore, data combination and alignment for this gene were manually corrected. Alignments of the four overlapping genes were nearly identical between the two datasets. We utilized the alignments from the 4162-taxon dataset because of the more comprehensive taxon sampling.

One species was present in the 171-taxon dataset but absent in the 4162-taxon dataset. For this species, *Smaug mossambicus* (formerly *Cordylus mossambicus*), sequences for non-overlapping genes of the 4162 taxon dataset were obtained from Stanley et al. (2011). These were then manually inserted into the 4172-taxon dataset, taking advantage of the dense taxon sampling of the genus *Smaug* in the study by Stanley et al. (2011).

The combined dataset contained 4172 taxa, including 10 outgroup species and 4162 squamate species, 52 genes, and 43,593 sites, with 92.03% missing data overall (mostly because of genes that lack sequence data in some taxa, but also including gaps in alignments within genes). The taxa included and GenBank numbers of their sequences are listed in Table S1. Full names of the 52 genes are given in Table S2. For each taxon, the number of genes ranged from 1 to 52 with a mean of 5.1 genes per species, and the length of sequences in each taxon ranged from 273 to 42,229 bp per species, with a mean of 3934 bp. For each squamate backbone taxon (the 161 taxa from the analysis of Wiens et al., 2012), the number of genes ranged from 6 to 52 with a mean of 42.1, and the length of their concatenated sequences ranged from 3502 to

Please cite this article in press as: Zheng, Y., Wiens, J.J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. Mol. Phylogenet. Evol. (2015), http://dx.doi.org/10.1016/j.ympev.2015.10.009

42,229 bp with a mean of 31,545 bp. The combined data matrix is available on Dryad (http://dx.doi.org/10.5061/dryad.tv055).

Taxonomy generally followed the Reptile Database (Uetz and Hošek, 2015). Several species that were treated as separate taxa by Pyron et al. (2013) are treated as synonymous in that database. These were tentatively retained as separate terminal taxa here (putative different species), but were listed as conspecific (Table S1).

2.2. Data partitioning

We determined the best-fit partition scheme for the combined dataset using PartitionFinder version 1.1.1 (Lanfear et al., 2012), using the Bayesian Information Criterion. Branch lengths were linked across partitions. The set of potential substitution models was restricted to the GTR+Γ model. RAxML applies only the GTR +Γ and GTR+I+Γ models, and only GTR+Γ is recommended by the developer of RAxML (i.e. because the I and Γ parameters are partially overlapping). The greedy search option was used in PartitionFinder. An analysis including all 4162 taxa and 52 genes was not computationally feasible. Therefore, the analysis was conducted on a reduced, representative dataset of 58 species (Table S3). These species were chosen both to be relatively complete (mean = 45.4 genes and 34,419 bp per species), and to represent as many squamate families as possible. Nearly all extant squamate families were included. The best-fit partitioning scheme divided the data into 28 partitions (Table S4), including separate partitions for the mitochondrial rRNA, mitochondrial protein coding genes, and nuclear protein coding genes.

2.3. Maximum-likelihood tree estimation

The combined dataset was analyzed using RAxML version 8 (Stamatakis, 2014a), utilizing CIPRES. Data were analyzed using the GTRCAT approximation (Stamatakis, 2006) with final GTR+Γ optimization. The rapid hill-climbing algorithm was used. We first conducted a set of six preliminary analyses to determine the initial rearrangement settings, specifically, whether to use automatic determination or a fixed value of 10. The value of 10 is sufficiently large and efficient for many datasets (Stamatakis, 2014b). In four of the six replicates, the fixed setting resulted in a higher final GAMMA-based likelihood. It also resulted in the highest final GAMMA-based likelihood across all six replicates. Consequently, the initial rearrangement setting was set to 10. For the GTRCAT approximation, the number of categories used was the default value of 25. We did not test alternative values of this parameter setting because of limited computational power. Furthermore, compared with the initial rearrangement setting, it appears to have less impact on the final results (Stamatakis, 2014b). In the final analysis, 200 inferences were executed to find the optimal tree. Then, 300 rapid bootstrap (Stamatakis et al., 2008) replicates were conducted using the GTRCAT approximation. SH-like nodal support values (Guindon et al., 2010) were also calculated with the GTR+Γ model. Model parameters were estimated up to an accuracy of 0.1 log likelihood units (the default value) for all analyses.

In order to compare support values between our combined-data tree and those of previous studies, we generated bootstrap values for the dataset of Pyron et al. (2013). These are presented in Fig. S1. For the dataset of Pyron et al. (2013), 300 rapid bootstrap replicates were executed using the GTRCAT approximation. The number of categories for the GTRCAT approximation was set to 25. For consistency, the partition scheme used by Pyron et al. (2013) was used. This consisted of 32 partitions, with the 10 protein coding genes partitioned by codon position and two partitions for the two ribosomal RNA genes (12S and 16S). We note that alternative designs would be to separately estimate the optimal partitioning scheme for these 12 genes alone or to use the same partitioning scheme as for the 52 genes but with 40 genes removed. None of these solutions is perfect, and we simply note that the use of somewhat different partitioning schemes between datasets may have had some influence on the results (e.g. Kainer and Lanfear, 2015), along with the differences in numbers of taxa and characters. However, we do not think that there should be a consistent bias in the overall results associated with these differences in partitioning schemes.

We also re-analyzed the dataset of Pyron et al. (2013) to ensure that an optimal tree was found from that study, which only used 10 inferences to find the optimal tree. In addition, we subdivided the data into clades to ensure that there were no artifacts associated with failing to find the optimal tree within clades, given the very large number of taxa overall. The methods and results are described in Appendix S1 and Table S5. In short, combining the results of these subdivided searches does not suggest that the overall tree was strongly suboptimal. These results also imply that 200 searches should be adequate for our analyses of the combined dataset of 4162 taxa.

2.4. Comparison of support for higher-level relationships

We compared bootstrap values for comparable clades between the tree generated here and those of Wiens et al. (2012) and Pyron et al. (2013). These comparisons focused only on higher-level relationships (given the limited sampling of species within higher-level clades in the dataset of Wiens et al., 2012). All the higher-level taxa (families and subfamilies) recognized in Fig. 1 of Pyron et al. (2013) were considered, as well as the clades relating these higher taxa. Also, three genera treated by Pyron et al. (2013) as *incertae sedis* were considered when comparing results of that study and the present study (these taxa were not included by Wiens et al., 2012). These taxa were the scincid lizard genus *Ateuchosaurus* and the lamprophiid snake genera *Micrelaps* and *Oxyrhabdium*. A total of 108 nodes were therefore included for the comparison of bootstrap support values between these two trees.

We also compared SH-like support values between our tree and that of Pyron et al. (2013). A total of 107 higher-level nodes were included in this comparison. SH-like supports are computed for a nearest neighbor interchange (NNI) optimal tree, which is generated during the process and can be slightly different from the initial tree (Anisimova and Gascuel, 2006). For example, for the optimal tree from our combined dataset, the position of Calabariidae was slightly different from that in the NNI-optimized tree. Consequently, this node was excluded, leaving 107 higher-level nodes instead of the 108 used in other comparisons.

We tested whether support values for higher-level relationships differed significantly between trees using the non-parametric Mann–Whitney *U* test, using SPSS version 12.0. The non-parametric test was used because normal distributions for all sets of support values were rejected by the Kolmogorov–Smirnov test. We also used the non-parametric Wilcoxon signed rank test, to compare support for paired (shared) nodes between trees.

2.5. Shimodaira–Hasegawa (SH) test

For the full dataset of 4162 taxa, we also tested whether adding the 44 genes for 161 species significantly changed relationships across the entire tree, relative to the analysis of 12 genes alone (Pyron et al., 2013). Therefore, we tested whether the dataset of Pyron et al. (2013) rejected the tree generated from the combined dataset, and whether the combined dataset rejected the tree of Pyron et al. (2013). We used the SH test (Shimodaira and Hasegawa, 1999) implemented in RAxML to compare different tree topologies.

Please cite this article in press as: Zheng, Y., Wiens, J.J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. Mol. Phylogenet. Evol. (2015), http://dx.doi.org/10.1016/j.ympev.2015.10.009

2.6. Estimating divergence dates

The optimal tree inferred from the combined dataset was used in estimating divergence times with treePL version 1.0 (Smith and O'Meara, 2012), which is an implementation of the penalized likelihood method (Sanderson, 2002) for very large datasets. Penalized likelihood (Sanderson, 2002) uses a tree with branch lengths and age constraints without prior parametric distributions. We utilized treePL because most other approaches to estimating divergence times (e.g. the uncorrelated lognormal relaxed clock approach in BEAST; Drummond et al., 2006; Drummond and Rambaut, 2007) would not be practical given the large number of taxa and genes analyzed here. For the treePL analysis, the farthest outgroup, Mammalia, was pruned because of the uncertainty in the corresponding branch lengths. Following Mulcahy et al. (2012), 13 fossil-based age constraints were used, including 11 minimum age constraints and 2 with both minimum and maximum ages (Table S6). The additive penalty function was used (Sanderson, 2002; Mulcahy et al., 2012), and the analysis was set to be thorough. The numbers of penalized likelihood replicates and cross validation simulated annealing iterations were set to 200,000 and 50,000, respectively. A "priming" analysis was first conducted to determine the best optimization parameters. Based on the results of this analysis, the values of gradient-based, auto-differentiation based, and auto-differentiation cross-validation-based optimizers were all set to 1. The random subsample and replicate cross-validation (RSRCV) analyses were conducted from 0.001 to 100,000 (in 10-fold increments) to determine the best smoothing value (which was found to be 100). RSRCV produces similar results to those using standard cross-validation (i.e. removing one taxon), but is capable of handling trees of thousands of taxa within a reasonable time frame (Smith and O'Meara, 2012).

3. Results

3.1. Comparison of topologies and support

The maximum likelihood analysis of the combined dataset (52 genes, 4162 squamate species) yielded a tree that is summarized in Fig. 1, available in full as Fig. S2, and available in Newick format in Appendix S2. The time-calibrated tree is summarized in Fig. 2, available in full as Fig. S3, and in Newick format in Appendix S3.

The tree more closely resembled the higher-level phylogeny based on 44 genes and 161 species (Wiens et al., 2012) than that based on 12 genes and 4161 species (Pyron et al., 2013). Among the 84 higher-level nodes from Wiens et al. (2012), 76 of them are shared with the combined tree here (90.5%). In contrast, 64 nodes (76.2%) are shared with the higher-level tree of Pyron et al. (2013). Among the 108 higher-level nodes from Pyron et al. (2013), 83 of them are shared with the combined tree here (76.8%). Thus, even though the matrix of Pyron et al. (2013) contained substantially more non-missing data (10,195,449 vs. 4,559,887 cells), the data matrix with more genes was still more influential for resolving higher-level relationships. We discuss specific differences in the higher-level relationships in the next section.

Comparison of mean bootstrap values for comparable nodes showed that adding the set of 44 genes increased mean likelihood bootstrap support, but not significantly. For all comparable higher-level nodes, the mean bootstrap support for the combined-data tree was 74.1% ($n = 108$; Fig. S2), whereas the mean support for comparable clades in the tree of Pyron et al. (2013) was 71.2% ($n = 108$; Fig. S1). This difference is not significant according to a two-tailed Mann–Whitney U test ($P = 0.434$). In contrast, the mean bootstrap support for comparable clades on the tree of Wiens et al. (2012) was 88.9% ($n = 84$). This is significantly higher than the comparable support values from this study and that of Pyron et al. (2013), based on a two-tailed Mann–Whitney U test ($P < 0.0001$). Similarly, support values for the SH-like test were greater for the higher-level clades after adding 44 genes (relative to those in the tree of Pyron et al., 2013), but not significantly so (87.4% vs. 85.5%; two-tailed Mann–Whitney U test, $P = 0.111$, $n = 107$ nodes; see Fig. S4 for values estimated for the higher-level tree in this study).

Comparison of paired nodes between trees using the Wilcoxon signed rank test yielded similar results. Specifically, bootstrap support was not significantly different between higher-level nodes from the combined-data tree estimated here and the tree of Pyron et al. ($n = 83$ shared nodes, mean support 83.7 vs. 82.8; $P = 0.508$). Support was significantly higher for nodes shared between the tree of Wiens et al. (2012) and the combined tree here ($n = 76$ shared nodes, mean support 93.3 vs. 83.2; $P < 0.0001$). Again, support values for the SH-like test were greater for the higher-level clades after adding 44 genes (relative to those from Pyron et al., 2013), but not significantly ($n = 83$ nodes; 92.4% vs. 91.5%; $P = 0.181$).

3.2. Comparison of higher-level relationships

In the following paragraphs, we describe how the tree from the combined analysis here is similar to and differs from the separate analyses based on many genes (Wiens et al., 2012) and many taxa (Pyron et al., 2013). Importantly, we do not attempt to review all of the previous literature on the phylogenetics of these taxa (much of which was already reviewed by Pyron et al., 2013). Note that we present bootstrap values for the tree of Pyron et al. (2013) to ensure that support values are comparable across all three studies (Fig. S1), instead of the SH-like test used exclusively by Pyron et al. (2013).

In the combined tree here (Fig. 1), dibamids are placed as the sister group to all other squamates. The clade uniting all squamates above Dibamidae is well-supported (bs = 90%). These are the same relationships found in the tree of Pyron et al. (2013). In contrast, the tree of Wiens et al. (2012) placed dibamids with Gekkota with moderate support (bs = 76%), and these two clades were together the sister group to all other squamates. It is notable that the analysis here and that of Wiens et al. (2012) both included many outgroups outside of Squamata, whereas Pyron et al. (2013) included only *Sphenodon*. Thus, our results here suggest that placement of Dibamidae as sister to all other squamates was not simply an artifact of limited outgroup sampling by Pyron et al. (2013).

Within Gekkota, relationships are similar to those of these two previous studies (and others; e.g. Gamble et al., 2012) in dividing the group into two clades: one including Carphodactylidae, Diplodactylidae, and Pygopodidae, and the other containing Eublepharidae, Gekkonidae, Sphaerodactylidae, and Phyllodactylidae. In the former clade, the relationships are similar to those of Pyron et al. (2013) in placing carphodactylids with pygopodids (with weak support in both studies), whereas Wiens et al. (2012) placed diplodactylids with pygopodids with strong support (bs = 90%). Relationships within the other clade are similar, although Phyllodactylidae was not included by Wiens et al. (2012).

Relationships among the families of Scincoidea (Xantusiidae, Gerrhosauridae, Cordylidae, Scincidae) are identical across the three analyses. However, there are some differences in higher-level relationships within Cordylidae and Scincidae. Within Cordylidae, the results here show Platysaurinae as paraphyletic with respect to Cordylinae, whereas Platysaurinae is monophyletic in the tree of Pyron et al. (2013). In Scincidae, higher-level relationships are generally similar, except that in the tree here Scincinae is paraphyletic with respect to Lygosominae and *Ateuchosaurus* is nested inside of Lygosominae, whereas in the tree of Pyron et al.

Please cite this article in press as: Zheng, Y., Wiens, J.J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. Mol. Phylogenet. Evol. (2015), http://dx.doi.org/10.1016/j.ympev.2015.10.009

Fig. 1. Summary of relationships among higher-level squamate clades estimated in this study, with numbers at nodes indicating bootstrap support values. The full species-level tree is shown in Fig. S2, and is available in Newick format in Appendix S2.

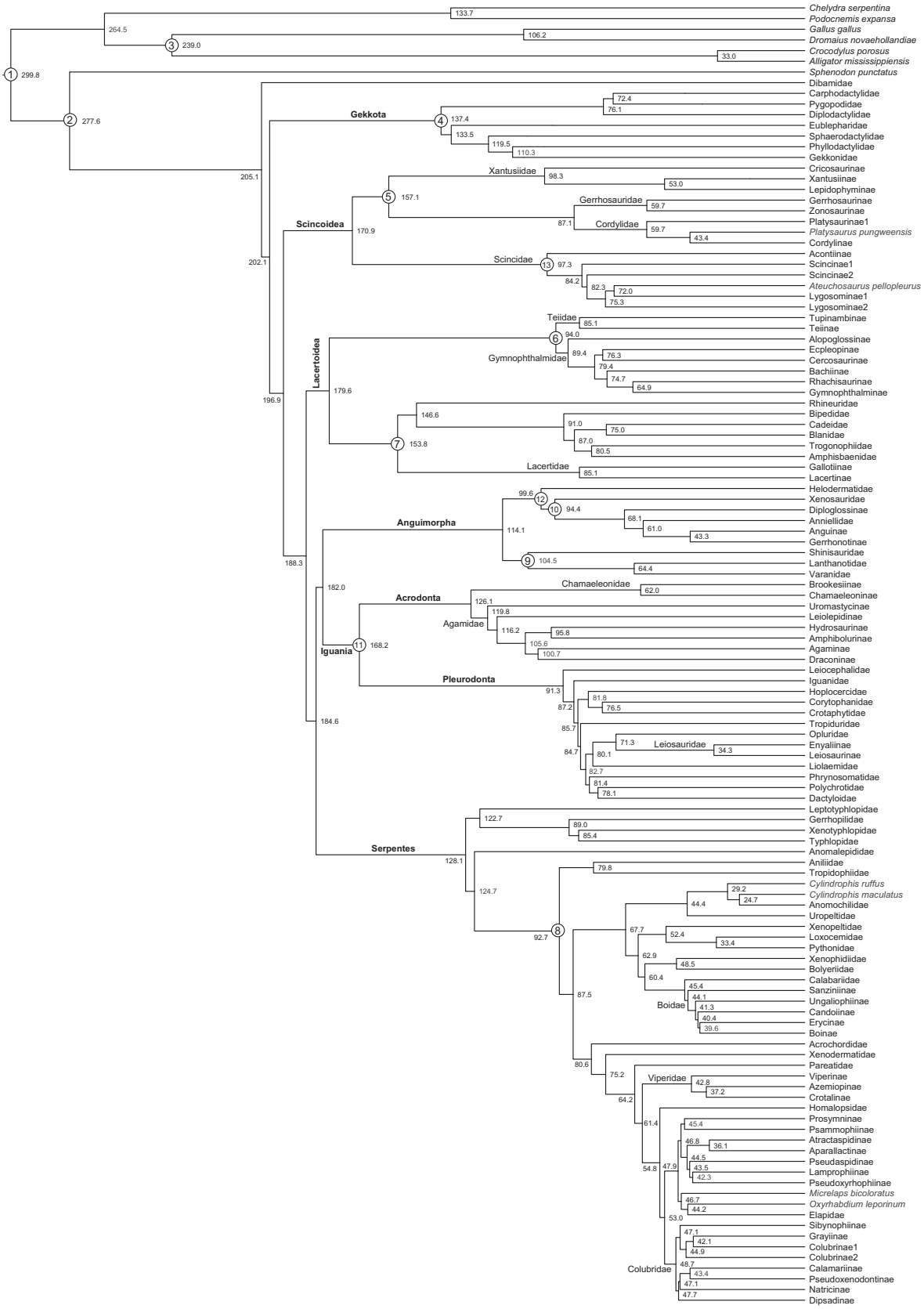

Fig. 2. Summary of time-calibrated phylogeny of higher-level squamate clades estimated in this study, with numbers at nodes indicating ages of clades. The full species-level tree is shown in Fig. S3, and is available in Newick format in Appendix S3.

Please cite this article in press as: Zheng, Y., Wiens, J.J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. Mol. Phylogenet. Evol. (2015), http://dx.doi.org/10.1016/j.ympev.2015.10.009

(2013) Scincinae is monophyletic and *Ateuchosaurus* is sister to Lygosominae. The tree of Wiens et al. (2012) also has Scincinae paraphyletic with respect to Lygosominae, but the relevant clades are not strongly supported, as is the case here. Monophyly of Scincinae is also supported in coalescent-based species-tree analyses of 10 and 44 loci with more limited taxon sampling (Lambert et al., 2015).

Relationships are also similar among the families and subfamilies of Lacertoidea (Gymnophthalmidae, Teiidae, Lacertidae, Amphisbaenia) across the three analyses. Relationships differ in Amphisbaenia, in that here Blanidae and Cadeidae are sister taxa (bs = 81%), whereas in the tree of Pyron et al. (2013), Blanidae is sister to a weakly supported clade consisting of Cadeidae, Amphisbaenidae, and Trogonophidae (note that Cadeidae was not included by Wiens et al., 2012). The relationships found here are more consistent with previous studies of the placement of Cadeidae (i.e. Vidal et al., 2007).

In the combined analyses here (Fig. 1), snakes are placed as the sister group to Anguimorpha + Iguania, as in other recent studies (e.g. Wiens et al., 2012; Pyron et al., 2013). However, the support for this clade is very weak here (bs = 37%, versus 67% by Wiens et al., 2012).

Within Anguimorpha, higher-level relationships are generally similar across studies, and differences are resolved in favor of the analysis with more genes (Wiens et al., 2012) rather than more species (Pyron et al., 2013). Specifically, in the analysis of Pyron et al. (2013), Xenosauridae is placed as sister to Helodermatidae + Anguidae with moderately strong support for Helodermatidae + Anguidae (bs = 89%). Here, and in Wiens et al. (2012), Helodermatidae is sister to a weakly supported clade of Xenosauridae + Anguidae. Within Anguidae, Diploglossinae is here placed as sister to the clade of Anniellinae (Anguinae + Gerrhonotinae), as in Wiens et al. (2012), both with moderate support for the latter clade (bs = 52% and 73%, respectively). In contrast, Pyron et al. (2013) placed Anniellinae as sister to Diploglossinae (Anguinae + Gerrhonotinae), with moderately strong support for the latter clade (bs = 84%). The results here do not support recognition of Anniellidae as a separate family distinct from Anguidae, since this renders Anguidae as non-monophyletic.

Within Iguania, all analyses support the clades Acrodonta and Pleurodonta. Within Acrodonta, the analyses here resolve the chamaeleonid subfamily Brookesiinae as monophyletic, whereas Brookesiinae was paraphyletic with respect to Chamaeleoninae in the tree of Pyron et al. (2013). Relationships among the subfamilies of Agamidae are largely identical between this study and that of Pyron et al. (2013). However, Wiens et al. (2012) placed Hydrosaurinae as sister to the strongly supported clade (bs = 100%) of Amphibolurinae (Agaminae + Draconinae), whereas the results here and those of Pyron et al. (2013) show weak support for placing Hydrosaurinae as sister to Amphibolurinae (bs = 50% and 49%, respectively).

The relationships among the families of pleurodont iguanians differ considerably across the three studies, but are generally very weakly supported. However, several clades are shared between the results here and those of Wiens et al. (2012) that differ from those of Pyron et al. (2013), including: (a) placing Leiocephalidae as sister to all other pleurodonts, instead of placing Tropiduridae as sister to other pleurodonts as in Pyron et al. (2013), (b) placing Corytophanidae and Crotaphytidae as sister taxa here, instead of placing Corytophanidae and Dactyloidae as sister taxa, (c) placing Phrynosomatidae as sister to Dactyloidae + Polychrotidae here, instead of placing phrynosomatids with crotaphytids and polychrotids with hoplocercids. All three analyses agree on a clade consisting of Liolaemidae + (Leiosauridae + Opluridae), but only the clade of Leiosauridae + Opluridae is strongly supported in all three analyses.

There are also several relationships within snakes that are shared by the tree here and that of Wiens et al. (2012), and that differ from those in Pyron et al. (2013). Here, and in the analysis of Wiens et al. (2012), the clade including Leptotyphlopidae, Typhlopidae, Xenotyphlopidae, and Gerrhopilidae is the sister to all other snakes, with Anomalepididae sister to the remaining snakes. The clade placing Anomalepididae with other snakes is strongly supported here and by Wiens et al. (2012), with bootstrap values of 100% and 95%, respectively. In the tree of Pyron et al. (2013), the position of these two clades is reversed, with Anomalepididae as sister to all other snakes, but with only weak support for the latter clade (bs = 40%). In the tree of Pyron et al. (2013), Xenophidiidae is placed as sister to a weakly supported clade consisting of all snakes exclusive of "scolecophidians" (Anomalepididae, Leptotyphlopidae, Typhlopidae, Xenotyphlopidae, Gerrhopilidae) and the clade of Aniliidae and Tropidophiidae. Here, Xenophidiidae is placed as sister to Bolyeriidae, but support is not strong (bs = 55%). This relationship is consistent with earlier studies (e.g. Lawson et al., 2004). Moreover, Pyron et al. (2013) placed Bolyeriidae as the sister taxon to the clade including Boidae, Calabariidae, Xenopeltidae, Loxocemidae, Pythonidae, Anomochilidae, Cylindrophiidae, and Uropeltidae, but with weak support for the latter clade (bs = 36%). Here, bolyeriids and xenophidiids are placed as the sister to Calabariidae and Boidae (as in Wiens et al., 2012). Furthermore, Calabariidae is here sister to Boidae, whereas Calabariidae is nested inside Boidae in the tree of Pyron et al. (2013). In the tree estimated here (Fig. 1), the clade of Cylindrophiidae, Anomochilidae, and Uropeltidae is sister to the clade of Boidae, Calabariidae, Xenopeltidae, Loxocemidae, and Pythonidae. In contrast, in the tree of Pyron et al. (2013), the clade of Cylindrophiidae, Anomochilidae, and Uropeltidae is sister to the clade of Xenopeltidae, Loxocemidae, and Pythonidae. In the tree estimated here, Anomochilidae is nested inside of Cylindrophiidae, whereas Anomochilidae is sister to Cylindrophiidae in the tree of Pyron et al. (2013). However, it should be noted that many of these relationships are weakly supported in all three studies.

Within the advanced snakes, relationships here generally follow the strongly supported relationships estimated by Wiens et al. (2012), rather than those estimated by Pyron et al. (2013). For example, here Acrochordidae is sister to a monophyletic Colubroidea (as in Wiens et al., 2012), which is strongly supported (bs = 92%). In contrast, in the tree of Pyron et al. (2013), Acrochordidae is sister to Xenodermatidae (bs = 93%), rendering Colubroidea paraphyletic. Here, the relationships are (Xenodermatidae (Pareatidae (Viperidae (Homalopsidae (Colubridae (Lamprophiidae + Elapidae)))))), as in the tree of Wiens et al. (2012). In the tree of Pyron et al. (2013), Homalopsidae is sister to Lamprophiidae + Elapidae, but with weak support for this clade (bs = 33%). Here, Homalopsidae is sister to Colubridae + (Lamprophiidae + Elapidae), and the clade including these latter three families is strongly supported (bs = 90%). Interestingly, most relationships among colubroid families found here were also found by Pyron et al. (2014) in species-tree analyses of 330 loci, in contrast to those of Pyron et al. (2013).

Relationships within the clade of Lamprophiidae + Elapidae differ somewhat between the tree here and that of Pyron et al. (2013), mostly due to different placements of the genera *Micrelaps* and *Oxyrhabdium*. Relationships among the subfamilies of Lamprophiidae are weakly supported in both studies (and most of these subfamilies were not included by Wiens et al., 2012).

Similarly, relationships among colubrid subfamilies differ between those estimated here and those in Pyron et al. (2013), but the relevant clades are very weakly supported. For example, Pyron et al. (2013) supported monophyly of Colubrinae, and placed Natricinae with Dipsadinae. Here, Colubrinae is paraphyletic with respect to Grayiinae, and Natricinae is sister to the clade of

Please cite this article in press as: Zheng, Y., Wiens, J.J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. Mol. Phylogenet. Evol. (2015), http://dx.doi.org/10.1016/j.ympev.2015.10.009

Table 1
Comparison of estimated divergence dates for selected major squamate clades from three recent studies and the present analysis. Estimates from Mulcahy et al. (2012) include those from BEAST and penalized likelihood (PL).

Clade	Mulcahy et al. (2012)-BEAST	Mulcahy et al. (2012)-PL	Pyron and Burbrink (2014)	Zheng and Wiens (2015)	Present study
Squamate root	180.0	191.8	174.1	212.7	205.1
Gekkota + Unidentata	173.4	189.5	168.8	197.7	202.1
Gekkota	80.8	102.1	86.5	59.9	137.4
Unidentata	162.8	184.6	168.8	183.2	196.9
Scincoidea	123.3	167.9	151.8	132.0	170.9
Episquamata	149.1	174.0	162.2	165.7	188.3
Lacertoidea	135.0	163.9	154.0	145.6	179.6
Toxicofera	140.8	169.8	160.6	155.1	184.6
Anguimorpha	107.9	110.2	118.0	111.8	114.1
Iguania	86.9	147.0	146.4	94.5	168.2
Serpentes	113.0	115.5	131.1	118.3	128.1

Calamariinae + Pseudoxenodontinae rather than Dipsadinae. More specifically, the analyses here place the strongly-supported clade of Asian arboreal colubrines (*Ahaetulla*, *Chrysopelea*, and *Drendrelaphis*) as sister to Grayiinae (with weak support, bs = 38%), instead of sister to all other colubrines (which together form a well-supported clade, bs = 92%). Pyron et al. (2013) mentioned that this former colubrine clade might need to be recognized as a separate subfamily (i.e. Ahaetullinae), and that members of this clade share potentially diagnostic phenotypic traits related to jumping and gliding.

The higher-level relationships found here are generally supported in recent analyses that combine the 44-gene dataset of 161 living taxa (plus two additional genes, both included here) with a morphological dataset including >600 characters and 49 fossil taxa (Reeder et al., 2015). The trees differ in the placement of Dibamidae (sister to Gekkota rather than to all other squamates as in our study), but higher-level relationships estimated here within anguimorphs and snakes, that differ from those in Pyron et al. (2013), are largely supported in the analyses incorporating morphological data and fossil taxa. These relationships include: (a) placement of Helodermatidae as sister to Xenosauridae and Anguidae, and not Anguidae, (b) Diploglossinae as sister to Anniellinae (Anguinae + Gerrhonotinae), and not Anguinae + Gerrhonotinae, (c) placement of the clade including Leptotyphlopidae and Typhlopidae as sister to all other snakes, instead of Anomalepididae, (d) Xenophidiidae as sister to Bolyeriidae, (e) Xenophidiidae + Bolyeriidae as sister to Boidae + Calabariidae, (f) Calabariidae as sister to Boidae, not nested inside it, (g) Acrochordidae as sister to Colubroidea, and not Xenodermatidae, and (h) Homalopsidae as sister to Colubridae (Lamprophiidae + Elapidae), and not Lamprophiidae + Elapidae.

3.3. Comparison of lower-level relationships

Despite these differences in higher-level relationships across the tree, relationships within these higher level clades are generally similar to those of Pyron et al. (2013). Specifically, Pyron et al. (2013) included a total of 4162 species, and a total of 4159 (4162–3) nodes can therefore be compared. Among them, 3637 (87.45%) nodes were found on both the tree of Pyron et al. (2013) and the tree inferred from our combined dataset. Conversely, a total of 522 (12.55%) nodes found on the tree of Pyron et al. (2013) were not observed on our combined-data tree. Note that all nodes are considered here, regardless of their levels of support.

Although these trees differ in a minority of nodes, the differences in overall relationships are statistically significant. Shimodaira–Hasegawa tests of both the combined dataset here and the dataset of Pyron et al. (2013) reject the other topology. Specifically, using the dataset of Pyron et al. (2013), the tree estimated here received a likelihood score of −2613067.59. This was significantly less optimal than the Pyron et al. (2013) tree (−2612683.80) at the 5% level (SH test, SD = 160.77). Similarly, based on the combined dataset used here, the tree of Pyron et al. (2013) received a likelihood score of −3441994.22. This was significantly less optimal than the score for the tree estimated here (−3440044.27) at the 1% level (SH test, SD = 186.13).

3.4. Time-calibrated phylogeny

We also time-calibrated the tree estimated here (summary in Fig. 2; full tree in Fig. S3). The tree is available in Newick format in Appendix S3. The dates estimated for major clades are broadly similar to those estimated in other recent studies (Mulcahy et al., 2012; Pyron and Burbrink, 2014; Zheng and Wiens, 2015), but several are older (Table 1). All four studies used very similar fossil calibration points. Along the backbone of squamate phylogeny, many of our estimates for higher-level clades are within ~20 Myr of those estimated by Zheng and Wiens (2015), an analysis utilizing BEAST and 20 nuclear loci with very little missing data, and our dates are both younger (for the squamate root) or somewhat older (for younger clades like Toxicofera). Our age estimates are substantially older for other clades, like Gekkota, Scincoidea, and Iguania. This may reflect the limited taxon sampling of Zheng and Wiens (2015), or possibly a tendency for penalized likelihood to estimate older clade ages than BEAST (e.g. Mulcahy et al., 2012). Our age estimates are also generally older than those estimated by Mulcahy et al. (2012) in their BEAST analysis and those estimated by Pyron and Burbrink (2014), using methods similar to those used here. However, the younger age estimates in the BEAST analysis of Mulcahy et al. (2012) may be (at least in part) an artifact of setting a narrow range on the calibration age priors (p. 979; Mulcahy et al., 2012). The analysis of Zheng and Wiens (2015) may therefore better reflect BEAST age estimates for these clades (since that study otherwise used very similar data and methods to those used by Mulcahy et al., 2012). Overall, the general similarity between dates estimated here and those in analyses with negligible missing data (e.g. Zheng and Wiens, 2015) reinforces the idea that large amounts of missing data need not be problematic for divergence-time estimation (e.g. Filipski et al., 2014; Zheng and Wiens, 2015).

4. Discussion

The data available for molecular phylogenetic analyses are becoming increasingly heterogeneous for many groups of organisms, from single genes that have been sequenced for thousands

Please cite this article in press as: Zheng, Y., Wiens, J.J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. Mol. Phylogenet. Evol. (2015), http://dx.doi.org/10.1016/j.ympev.2015.10.009

of species to whole genomes that are sequenced for only dozens. This heterogeneity raises the question of whether it is better to estimate phylogeny using many genes from fewer taxa or many taxa with fewer genes, and whether it is possible to combine these two approaches. Here, we explore this combination of supermatrix and phylogenomic approaches in squamate reptiles (although we note that the number of genes needed to be considered "phylogenomic" can be a moving target). We addressed two general questions about this combined approach. First, will higher-level relationships resemble those from many genes with fewer taxa or those estimated from many taxa but fewer genes? Second, will adding many genes for a few taxa increase branch support for higher-level relationships?

We found that higher-level relationships estimated from the combined approach were more similar to those based on many genes in fewer taxa, with 90% of clades shared relative to only 77% of higher-level clades shared with the tree based on fewer genes but more taxa. There were still some relationships that were resolved in favor of the tree of Pyron et al. (2013), such as the placement of Dibamidae as the sister group to all other squamates, as opposed to the clade of Dibamidae + Gekkota in Wiens et al. (2012). Nevertheless, the overall pattern was for more clades to be shared with the analysis based on many genes. This was especially apparent within snakes and within anguimorphs, where higher-level relationships were generally resolved in favor of the dataset based on many genes. Remarkably, the overall tree of 4162 ingroup taxa was significantly different (based on the SH test) from the tree for 4161 ingroup taxa from Pyron et al. (2013), even though new data were added in only 161 taxa.

An important question concerning these conflicts and their resolution is: which relationships are most likely to be correct? It makes intuitive sense that higher-level relationships will be better resolved by the set of many slow-evolving nuclear genes rather than a smaller, less complete dataset containing many fast-evolving mitochondrial genes. This intuition is also supported by the finding that higher-level relationships have stronger statistical support in the separate analysis based on many genes (mean bs = 89%; Wiens et al., 2012), significantly stronger than those based on many taxa but fewer genes (mean bs = 71%; Pyron et al., 2013). It is also notable that the higher-level relationships found here are generally supported in analyses that combine the 44-gene dataset with data from morphology and fossils (Reeder et al., 2015).

We also found that bootstrap values for higher-level relationships were higher after adding the dataset with many genes (mean bs = 74% versus 71%), but that this increase was not statistically significant. Moreover, we found that support values for higher-level relationships were significantly lower in the combined dataset relative to the more complete dataset of 44 nuclear genes alone (mean bs = 74% versus mean bs = 89%). Why are the bootstrap values for higher-level relationships here lower than those from the 44-gene dataset? We think that one potential explanation for this pattern is that the dataset with many taxa contained several higher taxa of highly uncertain placement, most likely due (at least in part) to their relatively limited data. Thus, even though the combined matrix overall contained 52 genes, some key higher taxa were still placed by a very small number of genes. For example, the snake *Xenophidion schaefferi* (the sole representative of the poorly known Xenophidiidae) was included based on a single mitochondrial gene, a gene that is relatively sparsely sampled among squamates (cytochrome *b*). Almost all nodes pertaining to the specific placement of this species within snakes had relatively low support values, both in our tree and that of Pyron et al. (2013). Thus, the inclusion of this taxon may have reduced support values widely in this portion of snake phylogeny, beyond its immediate sister group. A simple thought experiment further illustrates this idea. A taxon with no data whatsoever could be placed anywhere in the tree with equal support. Thus, adding such a taxon should lead to low bootstrap values throughout the tree, regardless of whether the initial dataset to which it was added included only 2 genes and generally had weak support values, or was based on 100 genes and had strong support for every node. Thus, taxa of uncertain placement should tend to equalize support values between datasets.

Another potential explanation is that conflict between the datasets may reduce support in the tree from the combined datasets. For example, the 44-gene dataset and 12-gene dataset each show strong support for conflicting relationships for some higher-level relationships within advanced snakes (e.g. placement of Xenodermatidae). Even though these conflicts are resolved in favor of the 44-gene dataset, support values may be lowered (relative to those from the 44-gene dataset) by the inclusion of a set of genes that (when combined) strongly favor a quite different set of relationships. Note that these two explanations are not mutually exclusive.

Overall, our results suggest tempered optimism for the combined sampling approach. We found that higher-level relationships were typically resolved as we might expect and hope (i.e. favoring the dataset with more genes) and that support values for these relationships were increased. Our optimism is tempered in that the increase in support values was moderate, and support values were significantly lower than those obtained from analyzing the dataset with many genes alone. We offered two potential explanations for why this may be the case.

More broadly, we show here that the large amounts of missing data do not seem to be an impediment to this combined approach (see also Wiens et al., 2005; Cho et al., 2011). Here, we analyzed a matrix with 92% missing data overall. The results that we obtained at broader phylogenetic scales were extremely similar to those from analyzing 44 genes alone (with less than 16% missing data), with 90% of nodes shared. More broadly, even though the tree differed significantly from that of Pyron et al. (2013), 87.5% of the 4159 nodes were shared with that tree. Thus, there is little basis for arguing that the increased amounts of missing data led to a radically different topology in this case. Although the dataset of Pyron et al. (2013) had considerable missing data to begin with (81%), the topology from that study was largely consistent with previous taxonomy at the level of genera, subfamilies, and families (and most deviations from that taxonomy had antecedents in earlier studies with less missing data; see extensive review in that paper). Our results here are also largely consistent with previous taxonomy (i.e. most genera, subfamilies, and families are monophyletic). For example, according to the Reptile Database (Uetz and Hošek, 2015), a total of 881 genera were included in our analyses, and 510 of these were represented by two or more species (and thus could be supported as monophyletic or not). In the tree of Pyron et al. (2013), 407 genera were monophyletic (79.8%). In our tree here, 403 genera were monophyletic (79.0%), and 399 genera were monophyletic in both trees (78.2%). Most higher taxa were monophyletic in both studies, and only a handful of higher taxa differed in their status as monophyletic vs. non-monophyletic between our study and that of Pyron et al. (2013), including some taxa found to be non-monophyletic only here (e.g. Colubrinae, Platysaurinae, Scincinae), and some only supported as monophyletic here (e.g. Boidae, Brookesiinae) but not supported by Pyron et al. (2013).

The idea that the amount of missing data is not an impediment to combining data matrices (and including incomplete taxa and characters) has strong precedents in earlier theoretical studies (review in Wiens and Morrill, 2011), and has important implications for future studies. Given our results, we suggest that it should be possible to combine datasets from sparsely sampled matrices with many taxa and few genes with those containing hundreds or even thousands of genes. In short, if theory and the empirical

Please cite this article in press as: Zheng, Y., Wiens, J.J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. Mol. Phylogenet. Evol. (2015), http://dx.doi.org/10.1016/j.ympev.2015.10.009

results here show that massive amounts of missing data are not (in themselves) problematic, we see no reason why this should change simply because more genes (and more missing data) are included.

Finally, we believe that our results provide an improved estimate of squamate phylogeny. Our results are generally similar to those of Pyron et al. (2013) throughout the tree, but are significantly different overall, and are more similar to the 44-gene results of Wiens et al. (2012) for higher-level relationships. We have also time-calibrated this tree to facilitate its use in comparative studies that require a temporal component (e.g. analyses of biogeography, diversification, and rates of trait evolution). This tree is available here in Newick format (Appendix S3).

Acknowledgments

Y.Z. was supported by grants from the National Natural Science Foundation of China (NSFC-31372181) and Chinese Academy of Sciences (Y1C3051100). We thank two anonymous reviewers for helpful comments on the manuscript.

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.2015.10.009.

References

Anisimova, M., Gascuel, O., 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. 55, 539–552.
Chiari, Y., Cahais, V., Galtier, N., Delsuc, F., 2012. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol. 10, 65.
Cho, S., Zwick, A., Regier, J.C., Mitter, C., Cummings, M.P., Yao, J., Du, Z., Zhao, H., Kawahara, A.Y., Weller, S., Davis, D.R., Baixeras, J., Brown, J.W., Parr, C., 2011. Can deliberately incomplete gene sample augmentation improve a phylogeny estimate for the advanced moths and butterflies (Hexapoda: Lepidoptera)? Syst. Biol. 60, 782–796.
Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.
Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4, e88.
Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., Sørensen, M.V., Haddock, S.D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R.M., Wheeler, W.D., Martindale, M.Q., Giribet, G., 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745–749.
Filipski, A., Murillo, O., Freydenzon, A., Tamura, K., Kumar, S., 2014. Prospects for building large timetrees using molecular data with incomplete gene coverage among species. Mol. Biol. Evol. 31, 2542–2550.
Gamble, T., Greenbaum, E., Jackman, T.R., Russell, A.P., Bauer, A.M., 2012. Repeated origin and loss of adhesive toepads in geckos. PLoS ONE 7, e39429.
Graybeal, A., 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. 47, 9–17.
Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.
Heath, T.A., Hedtke, S.M., Hillis, D.M., 2008. Taxon sampling and the accuracy of phylogenetic analyses. J. Syst. Evol. 46, 239–257.
Jarvis, E.D., Mirarab, S., Aberer, J., Li, B., Houde, P., Li, C., et al., 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346, 1320–1331.
Jetz, W., Thomas, G.H., Joy, J.B., Hartmann, K., Mooers, A.O., 2012. Global diversity of birds in space and time. Nature 491, 444–448.
Jiang, W., Chen, S.-Y., Wang, H., Li, D.-Z., Wiens, J.J., 2014. Should genes with missing data be excluded from phylogenetic analyses? Mol. Phylogenet. Evol. 80, 308–318.
Kainer, D., Lanfear, R., 2015. The effects of partitioning on phylogenetic inference. Mol. Biol. Evol. 32, 1611–1627.
Kocot, K.M., Cannon, J.T., Todt, C., Citarella, M.R., Kohn, A.B., Meyer, A., Santos, S.R., Schander, C., Moroz, L.L., Leib, B., Halanych, K.M., 2011. Phylogenomics reveals deep molluscan relationships. Nature 477, 452–456.
Lambert, S.M., Reeder, T.W., Wiens, J.J., 2015. When do species tree and concatenated estimates disagree? An empirical analysis with higher-level scincid lizard phylogeny. Mol. Phylogenet. Evol. 82, 146–155.
Lanfear, R., Calcott, B., Ho, S.Y., Guindon, S., 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29, 1695–1701.
Lawson, R., Slowinski, J.B., Burbrink, F.T., 2004. A molecular approach to discerning the phylogenetic placement of the enigmatic snake *Xenophidion schaeferi* among the Alethinophidia. J. Zool. 263, 285–294.
Mulcahy, D.G., Noonan, B.P., Moss, T., Townsend, T.M., Reeder, T.W., Sites Jr., J.W., Wiens, J.J., 2012. Estimating divergence dates and evaluating dating methods using phylogenomic and mitochondrial data in squamate reptiles. Mol. Phylogenet. Evol. 65, 974–991.
Poe, S., Swofford, D.L., 1999. Taxon sampling revisited. Nature 398, 299–300.
Poe, S., 2003. Evaluation of the strategy of long branch subdivision to improve accuracy of phylogenetic methods. Syst. Biol. 52, 423–428.
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.
Pyron, R.A., Wiens, J.J., 2011. A large-scale phylogeny of Amphibia including over 2,800 species, and a revised classification of extant frogs, salamanders, and caecilians. Mol. Phylogenet. Evol. 61, 543–583.
Pyron, R.A., Burbrink, F.T., Wiens, J.J., 2013. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. BMC Evol. Biol. 13, 93.
Pyron, R.A., Burbrink, F.T., 2014. Early origin of viviparity and multiple reversions to oviparity in squamate reptiles. Ecol. Lett. 17, 13–21.
Pyron, R.A., Hendry, C.R., Chou, V.M., Lemmon, E.M., Lemmon, A.R., Burbrink, F.T., 2014. Effectiveness of phylogenomic data and coalescent species-tree methods for resolving difficult nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia). Mol. Phylogenet. Evol. 81, 221–231.
Rannala, B., Huelsenbeck, J.P., Yang, Z., Nielsen, R., 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. 47, 702–710.
Reeder, T.W., Townsend, T.M., Mulcahy, D.G., Noonan, B.P., Wood, P.L., Sites Jr., J.W., Wiens, J.J., 2015. Integrated analyses resolve conflicts over squamate reptile phylogeny and reveal unexpected placements for fossil taxa. PLoS ONE 10, e0118199.
Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W., Cunningham, C.W., 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463, 1079–1083.
Rosenberg, M.S., Kumar, S., 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. U.S.A. 98, 10751–10756.
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.
Sanderson, M.J., 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19, 101–109.
Shimodaira, H., Hasegawa, M., 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16, 1114–1116.
Smith, S.A., O'Meara, B.C., 2012. TreePL: divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics 28, 2689–2690.
Smith, S.A., Wilson, N.G., Goetz, F., Feehery, C., Andrade, S.C.S., Rouse, G.W., Giribet, G., Dunn, C.W., 2012. Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature 480, 364–367.
Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.
Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol. 57, 758–771.
Stamatakis, A., 2014a. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.
Stamatakis, A., 2014b. The RAxML v8.1.X Manual. <https://github.com/stamatak/standard-RAxML/blob/master/manual/NewManual.pdf> (accessed 09.11.14).
Stanley, E.L., Bauer, A.M., Jackman, T.R., Branch, W.R., Mouton, P.L.F.N., 2011. Between a rock and a hard polytomy: rapid radiation in the rupicolous girdled lizards (Squamata: Cordylidae). Mol. Phylogenet. Evol. 58, 53–70.
Streicher, J.W., Schulte, J.A., Wiens, J.J., 2015. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. Syst. Biol. (in press). http://dx.doi.org/10.1093/sysbio/syv058.
Struck, T.H., Paul, C., Hill, N., Hartmann, S., Hösel, C., Kube, M., Lieb, B., Meyer, A., Tiedemann, R., Purschke, G., 2011. Phylogenomic analyses unravel annelid evolution. Nature 471, 95–98.
Uetz, P., Hošek, J. (Eds.), 2015. The Reptile Database. <http://www.reptile-database.org> (accessed 18.02.15).
Vaidya, G., Lohman, D.J., Meier, R., 2011. SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. Cladistics 27, 171–180.
Vidal, N., Azvolinsky, A., Cruaud, C., Hedges, S.B., 2007. Origin of tropical American burrowing reptiles by transatlantic rafting. Biol. Lett. 4, 115–118.
Weigert, A., Helm, C., Meyer, M., Nickel, B., Arendt, D., Hausdorf, B., Santos, S.R., Halanych, K.M., Purschke, G., Bleidorn, C., Struck, T.H., 2014. Illuminating the base of the annelid tree using transcriptomics. Mol. Biol. Evol. 31, 1391–1401.
Wiens, J.J., 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52, 528–538.
Wiens, J.J., 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst. Biol. 54, 731–742.
Wiens, J.J., Morrill, M.C., 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst. Biol. 60, 719–731.
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.

Please cite this article in press as: Zheng, Y., Wiens, J.J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. Mol. Phylogenet. Evol. (2015), http://dx.doi.org/10.1016/j.ympev.2015.10.009

Wiens, J.J., Fetzner, J.W., Parkinson, C.L., Reeder, T.W., 2005. Hylid frog phylogeny and sampling strategies for speciose clades. Syst. Biol. 54, 719–748.
Wiens, J.J., Hutter, C.R., Mulcahy, D.G., Noonan, B.P., Townsend, T.M., Sites Jr., J.W., Reeder, T.W., 2012. Resolving the phylogeny of lizards and snakes (Squamata) with extensive sampling of genes and species. Biol. Lett. 8, 1043–1046.
Zheng, Y., Wiens, J.J., 2015. Do missing data influence the accuracy of divergence-time estimation with BEAST? Mol. Phylogenet. Evol. 85, 41–49.
Zwickl, D.J., Hillis, D.M., 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51, 588–589.

Please cite this article in press as: Zheng, Y., Wiens, J.J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. Mol. Phylogenet. Evol. (2015), http://dx.doi.org/10.1016/j.ympev.2015.10.009