Syst. Biol. 65(1):128–145, 2016

DOI:10.1093/sysbio/syv058
Advance Access publication September 1, 2015

How Should Genes and Taxa be Sampled for Phylogenomic Analyses with Missing Data? An Empirical Study in Iguanian Lizards

JEFFREY W. STREICHER[1,2,*], JAMES A. SCHULTE II[3], AND JOHN J. WIENS[1]

[1]*Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA;* [2]*Department of Life Sciences, The Natural History Museum, London SW7 5BD, UK and* [3]*Department of Biology, Clarkson University, Potsdam, NY 13699, USA;* **Correspondence to be sent to: Department of Life Sciences, The Natural History Museum, London SW7 5BD, UK; E-mail: j.streicher@nhm.ac.uk*

Received 14 November 2014; reviews returned 29 July 2015; accepted 4 August 2015
Associate Editor: John McCormack

Abstract.—Targeted sequence capture is becoming a widespread tool for generating large phylogenomic data sets to address difficult phylogenetic problems. However, this methodology often generates data sets in which increasing the number of taxa and loci increases amounts of missing data. Thus, a fundamental (but still unresolved) question is whether sampling should be designed to maximize sampling of taxa or genes, or to minimize the inclusion of missing data cells. Here, we explore this question for an ancient, rapid radiation of lizards, the pleurodont iguanians. Pleurodonts include many well-known clades (e.g., anoles, basilisks, iguanas, and spiny lizards) but relationships among families have proven difficult to resolve strongly and consistently using traditional sequencing approaches. We generated up to 4921 ultraconserved elements with sampling strategies including 16, 29, and 44 taxa, from 1179 to approximately 2.4 million characters per matrix and approximately 30% to 60% total missing data. We then compared mean branch support for interfamilial relationships under these 15 different sampling strategies for both concatenated (maximum likelihood) and species tree (NJst) approaches (after showing that mean branch support appears to be related to accuracy). We found that both approaches had the highest support when including loci with up to 50% missing taxa (matrices with ~40–55% missing data overall). Thus, our results show that simply excluding all missing data may be highly problematic as the primary guiding principle for the inclusion or exclusion of taxa and genes. The optimal strategy was somewhat different for each approach, a pattern that has not been shown previously. For concatenated analyses, branch support was maximized when including many taxa (44) but fewer characters (1.1 million). For species-tree analyses, branch support was maximized with minimal taxon sampling (16) but many loci (4789 of 4921). We also show that the choice of these sampling strategies can be critically important for phylogenomic analyses, since some strategies lead to demonstrably incorrect inferences (using the same method) that have strong statistical support. Our preferred estimate provides strong support for most interfamilial relationships in this important but phylogenetically challenging group. [Missing data; phylogenomics; Reptilia; species tree; Squamata; taxon sampling; UCEs.]

High-throughput DNA sequencing is leading to exciting new data sets for phylogenetic analyses, potentially including hundreds or even thousands of loci (e.g., Crawford et al. 2012; McCormack et al. 2013; Leaché et al. 2015). At the same time, these approaches are raising (and intensifying) long-standing questions and debates about how phylogenetic studies should be designed. For example, should sampling of taxa be prioritized over sampling of characters, or vice versa (e.g., Graybeal 1998; Rannala et al. 1998; Poe and Swofford 1999; Rosenberg and Kumar 2001; Zwickl and Hillis 2002; Heath et al. 2008)? Should loci or taxa with missing data cells be included or excluded (e.g., Lemmon et al. 2009; Wiens and Morrill 2011; Wiens and Tiu 2012; Hovmöller et al. 2013; Roure et al. 2013)? Should the exclusion of missing data take priority over greater sampling of characters or taxa (e.g., Wiens 2005; Jiang et al. 2014)?

These issues are exemplified by the targeted-sequence capture approach. This approach has been proposed as the most promising for studying ancient groups with phylogenomic data (Faircloth et al. 2012; Lemmon and Lemmon 2013). However, when utilizing this approach, increased sampling of taxa and characters is often accompanied by increased levels of missing data (Fig. 1). This occurs because hybrid enrichment libraries often do not yield even enrichment across all taxa and loci (see Zhou and Holliday 2012; McCormack et al. 2013). But how this ragged sampling of genes and taxa should influence study design remains unclear. A large body of empirical and theoretical work has explored the consequences of including and excluding characters and taxa with missing data in phylogenetic analyses (e.g., Wiens 1998, 2003, 2005; Driskell et al. 2004; Philippe et al. 2004; Lemmon et al. 2009; Wiens and Morrill 2011; Wiens and Tiu 2012; Hovmöller et al. 2013; Roure et al. 2013; Jiang et al. 2014). Much of this work suggests that while too much missing data may sometimes rob characters or taxa of their potential benefits, the mere presence of missing data is not necessarily problematic. Moreover, the practice of excluding taxa or characters simply because they contain some missing data cells may itself have negative consequences for phylogenetic accuracy (e.g., Wiens and Tiu 2012; Wagner et al. 2013; Huang and Knowles 2014; Jiang et al. 2014).

Some recent analyses have explored the impact of missing data on phylogenomic anlayses, but have primarily focused on RADseq data. These analyses suggest that exclusion of missing data may sometimes be problematic. Huang and Knowles (2014) demonstrated that avoiding missing data cells in RADseq data can actually lead to excluding the most variable and potentially informative loci. Simulation studies have found that allowing higher proportions of missing data in RADseq alignments did not adversely affect the accuracy of estimated trees (Rubin et al. 2012). Empirical

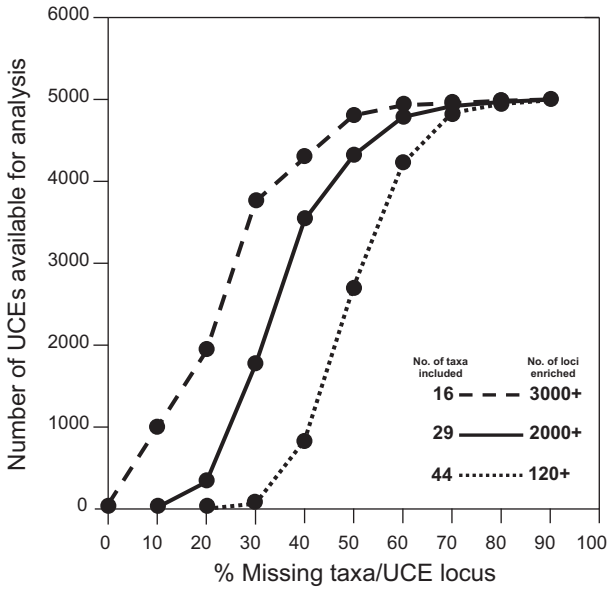

FIGURE 1. Relationship between the number of UCEs included in a data set and the maximum number of missing taxa allowed per UCE locus for that locus to be included, based on an empirical analysis of 44 iguanian lizards. Allowing more missing data per locus allows inclusion of more loci for a given taxon sampling regime. Each datapoint represents the whole matrix of UCEs generated by *phyluce* 2.0.0 when allowing different levels of missing taxa per individual UCE. Similarly, allowing smaller minimum numbers of UCEs per taxon allows inclusion of more taxa. The three sampling regimes depicted correspond to sets of taxa that were enriched for more than 3000 loci (16 taxa), 2000 loci (29 taxa), and 120 loci (all 44 taxa).

studies have found that allowing missing data led to higher nodal support on the same branches estimated by more complete data matrices (e.g., Streicher et al. 2014). However, the RADseq approach has a genealogical sampling bias that limits the number of loci that can be detected between divergent taxa (Arnold et al. 2013; Cariou et al. 2013). Thus, results from RADseq analyses may not apply to other types of data, nor to the deeper phylogenetic questions typically addressed with targeted-sequence capture data.

More generally, it is not clear how sampling for targeted-sequence capture studies should be designed (given finite resources). Should studies try to obtain large numbers of loci for a more limited set of taxa? Or more taxa and fewer loci? Should taxa or loci with missing data be excluded? What amount of missing data should be allowed? Do the answers to these questions change when applying concatenated versus species tree approaches? These fundamental questions have barely been addressed. Interestingly, most targeted-sequence capture studies have only presented phylogenetic results for data sets including approximately 30% missing data or less (e.g., Crawford et al. 2012; Faircloth et al. 2012, 2013; Lemmon et al. 2012; Leaché et al. 2014; Smith et al. 2014; Xi et al. 2014). In some cases, authors have literally halved the number of loci included in order to avoid including missing data cells (e.g., Leaché et al. 2014), but whether excluding these loci increases or decreases accuracy remains unclear.

Here, we take an empirical approach to the issue of optimal sampling design for targeted-sequence capture studies. We generate and explore a large data set with considerable missing data to address the conditions under which mean branch support for uncertain relationships is maximized (i.e., the combination of characters and taxa included). Although branch support may not be perfectly correlated with accuracy (i.e., the similarity between the estimated and true phylogeny; Hillis and Bull 1993), sampling approaches that lead to consistently weak support are clearly suboptimal. We also show that branch support for these uncertain relationships is correlated with those for well-established clades, and that branch support can be positively correlated with the number of well-established clades that are recovered (suggesting that this index is indeed related to accuracy).

We focus on pleurodont iguanians, an ancient rapid radiation that has repeatedly resisted attempts at phylogenetic resolution using traditional sequencing approaches. Pleurodont iguanians (*sensu* Frost et al. 2001) are one of the dominant lizard clades in the Western Hemisphere, with more than 1020 species (Sites et al. 2011; Uetz and Hošek 2014). This clade include many familiar and intensively studied taxa, such as *Anolis* (Dactyloidae), iguanas (Iguanidae), spiny and horned lizards (*Sceloporus, Phrynosoma*; Phrynosomatidae), basilisks (*Basiliscus;* Corytophanidae), the diverse South American genus *Liolaemus* (Liolaemidae), and the biogeographically enigmatic oplurids of Madagascar. Pleurodont iguanians are thought to have originated approximately 80 Ma (e.g., Townsend et al. 2011; Blankers et al. 2013) and are currently divided among 12 families. The monophyly of each family is well-supported, especially after recognition of Dactyloidae as a separate family from Polychrotidae (e.g., Townsend et al. 2011; Wiens et al. 2012; Blankers et al. 2013; Pyron et al. 2013; Reeder et al. 2015). In contrast, relationships among the families have been highly uncertain. These relationships are generally only weakly supported, and typically inconsistent between studies (Fig. 2). The only consistently well-supported interfamilial relationship is the clade uniting Opluridae and Leiosauridae (e.g., Schulte et al. 2003; Noonan and Chippindale 2006; Townsend et al. 2011; Wiens et al. 2012; Blankers et al. 2013; Pyron et al. 2013; Reeder et al. 2015). In many analyses, Liolaemidae is then sister to this clade (e.g., Wiens et al. 2012; Blankers et al. 2013), sometimes with strong support (e.g., Townsend et al. 2011; Pyron et al. 2013).

Here, we use our data to address two interrelated questions. First, what sampling strategy for genes and taxa (and associated level of missing data) maximizes branch support for concatenated and species-tree analyses? Second, what are the relationships among the families of pleurodont lizards? To address these questions, we generated and analyzed data matrices with 3–4921 loci and different levels of taxon sampling

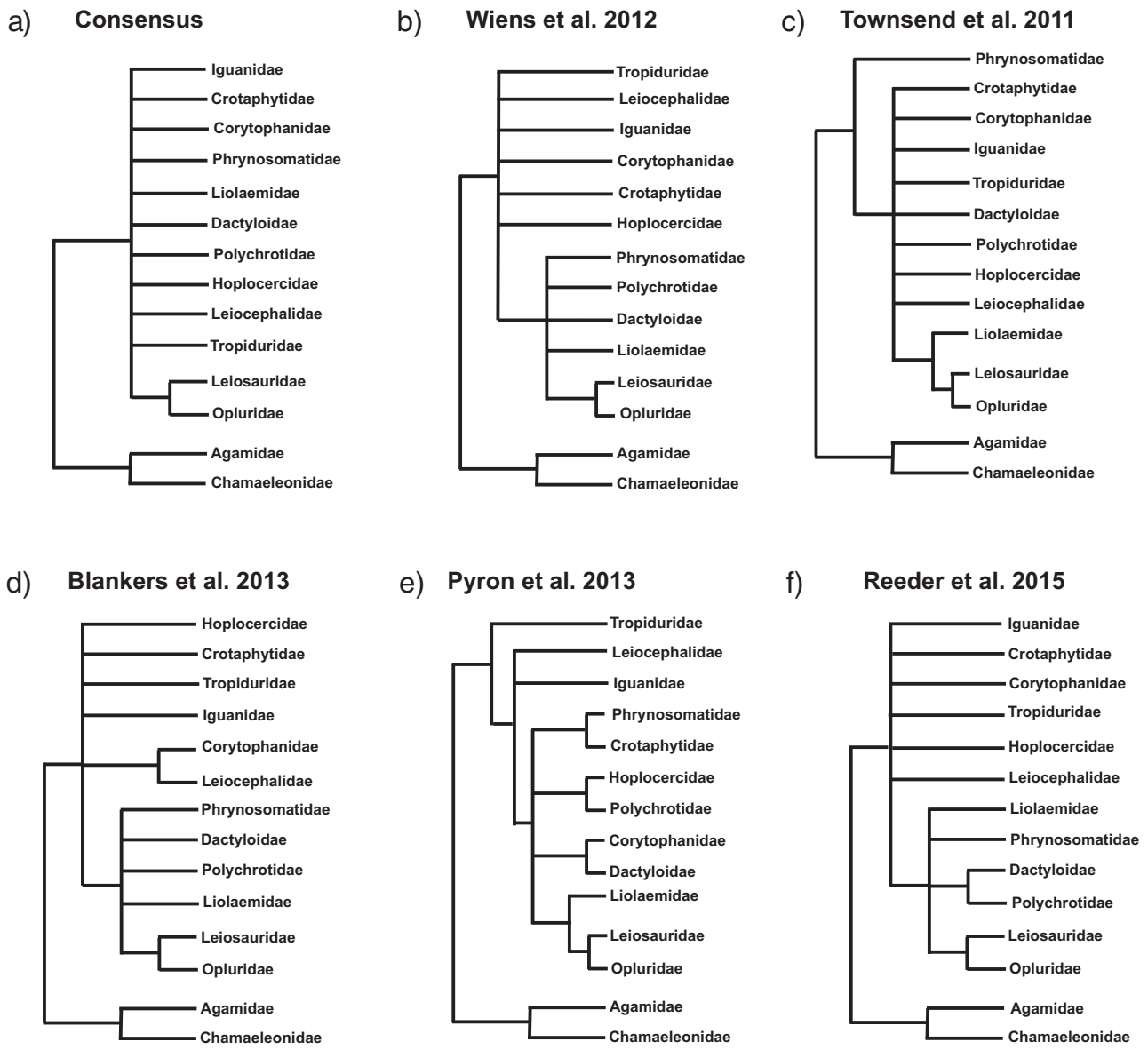

FIGURE 2. Summary of recent hypotheses of between-family relationships in iguanian lizards. a) Strict consensus tree of relationships in b–f, showing that apart from the sister relationship between Opluridae and Leiosauridae, there is little consensus among studies for family-level relationships. b) Wiens et al. (2012; their Fig. 1, from concatenated likelihood analysis of 44 nuclear loci). c) Townsend et al. (2011; their Fig. 3, from concatenated likelihood and Bayesian analyses of 29 nuclear loci). d) Blankers et al. (2013; their Fig. 1, based on concatenated likelihood analysis of 30 mostly nuclear genes). e) Pyron et al. (2013; their Fig. 1, based on concatenated likelihood analysis of 12 nuclear and mitochondrial genes). f) Reeder et al. (2015; their Fig. 1 based on concatenated likelihood analysis of 46 mostly nuclear genes and morphological data). All branches with less than 75% support have been arbitrarily collapsed to show weak support.

(16, 29, and 44 taxa), and missing data (missing data in up to 20–60% of taxa per locus, yielding ~30–60% missing data in each matrix overall).

METHODS

Targeted Sequence Capture

We used ultraconserved elements (UCEs) to construct a phylogenomic data set for iguanian lizards. UCEs are highly conserved regions of genomes observed across distantly related taxa (Bejerano et al. 2004; Sandelin et al. 2004; Miller et al. 2007). These regions range in size from ca. 300–1000 bp (Bejerano et al. 2004). They are ideal for targeted sequence capture because flanking regions on either side of each "ultraconserved" region targeted by RNA probes can vary both among and within species (Faircloth et al. 2012; Smith et al. 2014). UCEs have been used successfully to conduct phylogenetic studies in several major vertebrate groups (e.g., mammals, McCormack et al. 2012; fish, Faircloth

et al. 2013; birds, McCormack et al. 2013; Sun et al. 2014; reptiles, Crawford et al. 2012; 2015; Leaché et al. 2015) and within other animal clades (e.g., arthropods, Faircloth et al. 2014).

Taxon Sampling and UCE Enrichment Protocol

We selected 43 iguanian species that included multiple representatives of all acrodont and pleurodont families (Table 1 and Fig. 2). Vouchers are listed in Supplementary Table S1, available on Dryad at http://dx.doi.org/10.5061/dryad.13t18. Most of these taxa (and individuals) were already included by Townsend et al. (2011) and Wiens et al. (2012), but we added species to augment sampling of Dactyloidae, Leiocephalidae, and Polychrotidae (which were each represented by only one species each in the previous studies), along with additional species

TABLE 1. Summary of UCE enrichment results for the data set generated in this study

Family	Species	Number of trimmed reads	Number of contigs	N50 value	Number of UCE duplicates	Number of UCEs enriched
Pleurodonta (Ingroup)						
Corytophanidae	*Basiliscus basiliscus*	1,211,018	3642	365	120	2295
Corytophanidae	*Corytophanes cristatus*	730,537	4116	335	291	2213
Corytophanidae	*Laemanctus serratus*	313,931	3710	328	13	877
Crotaphytidae	*Crotaphytus collaris*	801,057	4682	392	279	2951
Crotaphytidae	*Gambelia wislizenii*	2,239,228	8827	412	662	3743
Dactyloidae	*Anolis carolinensis**	N/A	N/A	N/A	57*	4328*
Dactyloidae	*Anolis olssoni*	426,022	3356	449	49	2687
Dactyloidae	*Anolis sagrei*	99,472	476	325	2	397
Hoplocercidae	*Enyalioides laticeps*	404,395	1365	298	19	954
Hoplocercidae	*Morunasaurus annularis*	861,027	3461	337	156	2006
Iguanidae	*Brachylophus fasciatus*	853,365	3100	324	83	1318
Iguanidae	*Dipsosaurus dorsalis*	776,045	4608	452	115	3310
Iguanidae	*Sauromalus ater*	1,305,668	5385	486	133	3536
Leiocephalidae	*Leiocephalus barahonensis*	1,237,472	5174	544	168	3681
Leiocephalidae	*Leiocephalus carinatus*	1,419,944	5686	610	145	3650
Leiocephalidae	*Leiocephalus personatus*	725,247	4731	519	116	3618
Leiosauridae	*Leiosaurus catamarcensis*	11,193,933	101,008	292	627	3679
Leiosauridae	*Pristidactylus torquatus*	446, 635	2173	313	82	1439
Leiosauridae	*Urostrophus vautieri*	2,236,444	6906	361	748	2786
Liolaemidae	*Liolaemus bellii*	182,943	1038	275	21	699
Liolaemidae	*Liolaemus magellanicus*	8,899,098	16,463	579	278	3726
Liolaemidae	*Phymaturus palluma*	636,235	3066	349	108	2010
Liolaemidae	*Phymaturus somuncurensis*	295,428	2252	425	34	1850
Opluridae	*Chalarodon madagascariensis*	4,776,610	13,195	382	676	3195
Opluridae	*Oplurus cyclurus*	816,027	4260	409	229	2715
Phrynosomatidae	*Petrosaurus mearnsi*	9,781,841	15,649	425	1055	3733
Phrynosomatidae	*Phrynosoma platyrhinos*	2,891,114	8234	488	553	3704
Phrynosomatidae	*Sceloporus variabilis*	35,079	148	276	2	120
Phrynosomatidae	*Uma scoparia*	2,925,795	6658	563	214	3460
Phrynosomatidae	*Uta stansburiana*	6,327,548	15,625	404	1046	3993
Polychrotidae	*Polychrus acutirostris*	421,524	3277	453	66	2641
Polychrotidae	*Polychrus marmoratus*	285,518	2121	387	24	1784
Tropiduridae	*Microlophus atacamensis*	240,539	1569	389	23	1290
Tropiduridae	*Stenocercus guentheri*	82,380	657	266	9	496
Tropiduridae	*Plica plica*	3,556,741	8122	374	637	2839
Tropiduridae	*Uranoscodon superciliosus*	1,557,595	5076	527	135	3567
Acrodonta (Outgroup)						
Agamidae	*Calotes emma*	1,272,693	4860	386	217	2625
Agamidae	*Draco blanfordii*	811,004	3094	339	118	1914
Agamidae	*Hydrosaurus* sp.	3,298,704	6779	398	501	2947
Agamidae	*Leiolepis belliana*	1,299,762	4768	383	271	2577
Agamidae	*Physignathus cocincinus*	299,728	1803	271	36	708
Agamidae	*Trapelus agilis*	481,714	2693	359	51	1910
Chamaeleonidae	*Brookesia brygooi*	387,490	2341	299	184	1398
Chamaeleonidae	*Chamaeleo calyptratus*	647,718	3781	450	138	2731

Notes: Column definitions are as follow: the number of trimmed reads is the total number of Illumina reads following the removal of adaptor contamination and poor quality sequences in illumiprocessor. The number of contigs is the total number of contigs generated by Velvet assembly that mapped to UCE loci. The N50 value is the length for which all contigs of that length or longer contain at least half of the sum of the lengths of all contigs produced by Velvet. The number of UCE duplicates is the total number of UCE loci removed by phyluce for matching more than one Velvet contig (large numbers may relate to contamination). The number of UCEs enriched is the total number of unique loci enriched for this study. Note that data for *Anolis carolinensis* were taken from Alföldi et al. (2011).

of Corytophanidae, Liolaemidae, and Tropiduridae. In addition to generating new sequence data, we downloaded a previously sequenced set of UCEs for the ingroup species *Anolis carolinensis* (Alföldi et al. 2011; Crawford et al. 2012). We sampled from a set of eight genera of acrodont iguanians for use as outgroups, including representatives of both families (Agamidae, Chamaeleonidae), both subfamilies of Chamaeleonidae, and most subfamilies within Agamidae. Acrodonts are strongly supported as the sister group to pleurodonts (e.g., Townsend et al. 2011; Wiens et al. 2012; Pyron et al. 2013; Reeder et al. 2015). Preliminary analyses including more distant outgroups had relatively little impact on topologies within Pleurodonta, and so were not included in this study.

We followed the protocol described by Faircloth et al. (2012; available at http://ultraconserved.org), but with some modifications. We used a dsDNA high sensitivity assay kit and a Qubit fluorometer (Life Technologies) to measure the amount of DNA template. For each sample, we enzymatically sheared ca. 100–150 ng of DNA using NEBNext dsDNA Fragmentase (New England Biolabs) at 37°C for 25 min. Given the dilute concentrations of our samples, we did not check fragment size distribution after shearing. We ordered 88 oligonuclotides to construct 43 uniquely barcoded adapters and 1 universal adapter (Sigma-Aldrich). These oligonuclotides are listed in Supplementary Table S2, available on Dryad at http://dx.doi.org/10.5061/dryad.13t18. We used a × 10 annealing buffer to construct adaptors (500 mM NaCl, 100 mM Tris–HCL, 10 mM EDTA). We used a master-mix library prep kit to repair fragmented samples, A-tail, and ligate adaptors (NEBNext DNA library prep Master Mix Set). All sample cleaning between treatments was performed using magnetic beads in × 1.8 ratio of beads to sample (Rohland and Reich 2012; Sera-Mag Speedbeads, Fisher Scientific). We pooled ligated samples prior to size selection. We used a Pippin Prep electrophoresis system (Sage Science) and 1.5% agarose cartridges to size select fragments between 438 and 538 bp. Size selection was performed on pools of 12 ligated samples that each had a concentration of 10–30 ng/μL (total of 100–300 ng of DNA/pool). Following size selection we combined all 48 samples into a single pool. We amplified shotgun libraries from this pool using a high fidelity Phusion polymerase (NEB) and Illumina Truseq primers PCR1 (5′-AAT GAT ACG GCG ACC ACC GAG A-3′) and PCR 2 (5′-CAA GCA GAA GAC GGC ATA CGA G-3′) in a 12 cycle PCR (10 s at 98°C, 30 s at 65°C, 30 s at 72°C). To capture UCEs from our shotgun libraries, we used the 5000 tetrapod probe set (available from www.ultraconserved.org) and ordered a custom Sure Select XT target enrichment kit (Agilent). We used this kit with Dynabeads (M-270 Streptavidin; Life Technologies) to hybridize probes to our pooled library and isolate fragments of UCEs.

We deviated from the standard UCE protocol by not diluting the SureSelect probe mix and performing a single sequence capture on 48 pooled individuals (some not included in this study). We used this approach (versus the recommended 1–24 individuals per capture) because we were interested in maximizing the number of individuals per sequence capture. An experiment with fewer individuals per sequence capture (25) suggested that pooling 48 samples did not strongly decrease the number of loci obtained per individual. The mean number of UCE loci per taxon was 2091 for two captures with 48 individuals each and 2258 for the single capture with 25 individuals (Streicher J.W. and Wiens J.J., unpublished data).

A post-hybridization PCR was conducted using Phusion enzyme and Truseq primers for 18 cycles using the same cycling profile described above. Enriched libraries were visualized for fragment size distribution and concentration using Bioanalyzer 7500 DNA chip sets (Agilent). We sequenced the capture library (48 individuals at a time) using a 600 cycle sequencing paired-end run (i.e., PE300) on an Illumina MiSeq at the University of Texas at Arlington genomics core facility (Arlington, TX, USA; http://gcf.uta.edu/).

UCE Quality Control and Alignment

Sequences were converted to fastq and demultiplexed using BaseSpace. To process our raw demultiplexed sequence data we used the program *illumiprocessor* 2.0.2 (Faircloth et al. 2013; Bolger et al. 2014). Specifically, we trimmed low-quality ends and removed adaptor contamination from all reads. We assembled reads 1 and 2 (along with any high-quality singleton reads identified by Illumiprocessor 2.0.2) using Velvet 1.2.10 (Zerbino and Birney 2008) with a kmer length of 75 and a coverage cutoff of 10. To identify and align UCE loci from contigs assembled in Velvet 1.2.10 we used the software *phyluce* 2.0.0 (phylogenetic estimation from UCEs; Faircloth et al. 2012; available online at http://github.com/faircloth-lab/phyluce). We aligned contigs in *phyluce* 2.0.0 using the MUSCLE alignment algorithm (Edgar 2004) with default settings (including those for trimming). We deposited raw Illumina data used to identify UCEs in the NCBI sequence read archive (Supplementary Table S1, available on Dryad at http://dx.doi.org/10.5061/dryad.13t18) and all alignments and trees in the Dryad Digital Repository (www.datadryad.org).

Concatenated Phylogenetic Analyses

We concatenated UCE data using the convert feature of *phyluce* 2.0.0. On each concatenated data set, we performed a maximum-likelihood analysis (single tree search) in RAxML 8.0 (Stamatakis 2014) applying a single GTRCAT model across the alignment. In theory, we could have applied partitions to different subsets of the data. However, searching for the best-fitting partitions would have been computationally problematic given the large number of loci involved, and common partitions in other DNA data sets (e.g., codon positions) are largely lacking in UCE data (Bejerano et al. 2004). Moreover, any

existing heterogeneity in rates should be accounted for (at least in part) by including the gamma distribution of rates among sites (which RAxML calculates as a final step when running the GTRCAT model). We used the cyberinfrastructure for phylogenetic research science gateway (CIPRES; Miller et al. 2010) to run concatenated analyses. We used rapid botstrapping feature with an automatic cutoff. Using bootstrap values obtained from rapid bootstrapping with the GTRCAT model is controversial because they can potentially be inflated or misleading (e.g., Simmons and Norton 2014). Given the large sizes of many of our data sets (i.e., computational limitations) and a desire to employ the same methodology when generating nodal support for all concatenated analyses, we had little choice but to use this methodology. However, we also ran a subset of our analyses with RAxML using the GTR+Γ model for 100 bootstrap pseudoreplicates (focusing on those conditions that maximized support in GTRCAT model-based analyses). Comparing bootstrap values did not reveal a pattern of consistent inflation (see Results).

Species-Tree Analysis

Species-tree analysis can be problematic for UCE data sets in that the large number of loci precludes use of many coalescent-based species-tree methods (e.g., *BEAST; Heled and Drummond 2010). Further, most species-tree methods that are capable of analyzing thousands of loci in a reasonable amount of time (i.e., less than 2 weeks) do not allow one to include genes or taxa with missing data (e.g., STAR: Liu et al. 2009; MP-EST: Liu et al. 2010; SNAPP: Bryant et al. 2012). An important exception is NJst (Liu and Yu 2011), a summary statistic method that uses a group of unrooted gene trees as input. To estimate species trees using NJst, we first generated gene trees for each UCE locus using likelihood analyses in RAxML 8.0. We applied a single GTR+Γ model to each gene for gene-tree estimation and conducted tree searches (*sensu* Liu and Yu 2011), using the –m GTRGAMMA command. We assessed nodal support for NJst analyses via the method of Seo (2008). Specifically, we generated 100 bootstrap samples for each UCE using RAxML 8.0 and used these as input for NJst. We ran all NJst analyses using the Species Tree Analysis Web Server (Shaw et al. 2013).

Recently, an additional species-tree method that can use gene trees with missing taxa was described (ASTRAL; Mirabab et al. 2014). However, in the original description it was not compared with NJst. Here, as an exploratory analysis, we ran ASTRAL 4.7.6 for the conditions that maximized support in the NJst analyses (described in further detail below), and compared the results. As in NJst analyses, nodal support for ASTRAL analyses was obtained using the method of Seo (2008).

Comparison of Sampling Strategies

We obtained an average of 2413 UCEs (± 168.62 SE; range: 120–3993; median: 2641) for each species newly sequenced for this study (Table 1). No single UCE locus was enriched for all species. By allowing for increasing numbers of missing taxa per locus, we were able to include thousands of UCE loci (Fig. 1). We explored three taxon sampling strategies. The first (hereafter "44 taxa") allowed all taxa to be included (all 36 ingroup and 8 outgroup species), even those that had a relatively low number of UCE loci (as few as 120 loci enriched). The second (hereafter "29 taxa") included only taxa enriched for more than 2000 loci. This criterion allowed inclusion of 25 ingroup species and all 12 families (along with four outgroup taxa). The third (hereafter "16 taxa") included only taxa enriched for more than 3000 loci and included 15 ingroup species, representing nine families and one outgroup taxon. For each of the three taxon sampling schemes, we then generated five data sets that allowed for a maximum of up to 20%, 30%, 40%, 50%, and 60% missing taxa per locus using the "align adjust; –min taxa" option in *phyluce* 2.0.0. In total, we generated 15 unique data sets, each based on a slightly different sampling strategy for including taxa and loci. We then performed a concatenated and NJst species tree analysis on each data set.

It is important to clarify here the terminology with regards to "missing taxa for a given locus" and "overall missing data in the matrix". With UCE data sets, one typically includes or excludes loci based on the proportion of taxa that are missing data for that locus. Thus, one can be relatively strict (e.g., only allowing loci that have missing data for 20% or fewer of the taxa in the overall matrix) or more liberal (e.g., including loci with missing data in up to 60% of the taxa). The maximum percent of missing taxa per locus is clearly related to the overall percentage of missing data cells in the matrix, but (at least in our study) only for a given level of taxon sampling. The overall amount of missing data was also related to the number of taxa included, with inclusion of more taxa leading to more missing data in the matrix overall. Including fewer taxa meant that more loci were (on average) more complete. In other words, for 16 taxa, the maximum amount of missing data was less (40% or less) than for 44 taxa (58% or less), even though the maximum percentage of missing taxa was the same.

We used the data sets with 50% missing taxa per locus (across all sampling strategies) to examine the distribution of segregating sites, alignment length, and number of taxa (Supplementary Fig. S1, available on Dryad at http://dx.doi.org/10.5061/dryad.13t18; Table 2). We calculated the number of segregating sites using the R package *ape* (Paradis et al. 2004). We calculated alignment lengths and number of taxa for each UCE using Geneious 8.05 (BioMatters; www.geneious.com). We found that the average number of segregating sites and maximum alignment length decreased slightly when allowing less-complete taxa in the matrix (Table 2). We also observed more right-skewed distributions in the number of taxa per locus when using strategies that allow more missing data (Supplementary Fig. S1, available on Dryad at http://dx.doi.org/10.5061/dryad.13t18).

TABLE 2. Mean and range data for the number of segregating sites, alignment length, and the number of taxa for UCE data generated for this study

Sampling strategy	Number of UCEs	Number of segregating sites	Alignment length (base pairs)	Number of taxa/ alignment
		Mean (Range)	Mean (Range)	Mean (Range)
16 taxa	4789	86.0 (3–295)	502.3 (141–1075)	11.8 (8–16)
29 taxa	4319	84.9 (5–300)	427.0 (134–931)	18.9 (15–28)
44 taxa	2716	82.7 (10–267)	400.8 (159–849)	24.7 (22–36)

Note: All strategies were summarized using an up to 50% missing taxa per locus criterion.

The results of both concatenated and species-tree analyses were congruent with previous studies in generally supporting monophyly of all pleurodont families, but with variable support for relationships among familes (Figs 3 and 4). Given the questionable relationships among pleurodont families, we focused on the support for these between-family nodes to compare sampling strategies. These included 10 nodes in the 29- and 44-taxon data sets and 7 nodes in the 16-taxon data set. Thus, we summarized the mean support across these sets of nodes for each method for each set of conditions. We address how this measure might be related to accuracy in a separate section below.

In order to explore the causes of variation in the results among sampling strategies, we also tested for bivariate relationships between variables such as branch support, the percentage of missing taxa allowed per locus, the number of characters included, the number of loci included, and the overall amount of missing data in each matrix. We assessed the total percentage of missing data in each matrix from RAxML outputs (i.e., "proportion of gaps and completely undetermined characters in this alignment"). We used the results of each set of 15 analyses as the units for these comparisons (i.e., 16, 29, and 44 taxa, 20–60% missing taxa per locus). For comparisons between variables that were normally distributed (assessed by visualizing histograms and by use of Shapiro–Wilk tests) we used least squares regression (r^2). For variables failing to meet the assumption of normality, we used nonparametric two-sided Spearman's rank correlation coefficients (ρ). These analyses were conducted in SYSTAT 11 (Systat Inc.).

We primarily evaluated sampling strategies to determine the conditions under which mean support for between-family nodes was maximized for each method (concatenated, species-tree). We also evaluated the conditions under which congruence between these methods was maximized. To assess the similarity of trees from each method, we tallied the number of identical between-family nodes and divided by the overall number of these nodes (either 7 [16 taxa] or 10 [29, 44 taxa]).

We also evaluated the conditions under which these methods are congruent. Specifically, we tested whether concatenated and species-tree methods tend to be more congruent on longer branches (e.g., Lambert et al. 2015). To do this, we first visually categorized branches as being either congruent or incongruent between the concatenated and species-tree estimates for a given set of conditions. We then tested for differences between the estimated branch lengths in these two categories using a nonparametric Kruskal–Wallis one-way analysis of variance (given the non-normality of the data) and basic descriptive statistics. We focused on those conditions under which each method had highest support (see below), and separately tested the sets of branch lengths estimated by each method.

We note that branch lengths in NJst species trees are defined as the average number of internodes between two species across gene trees and that this metric is not considered equivalent to branch lengths generated by other species-tree methods (Liu and Yu 2011). Some subsequent authors have reported that NJst branch lengths are indicative of patterns that are of questionable biological meaning (e.g., the number of times a group appears in an input tree; Shaw et al. 2013). Thus, we decided to test if congruent branch lengths had a strong positive relationship between concatenated and NJst species-tree analyses, as would be expected if these methods produce similar branch lengths when using the same underlying data set. We regressed natural log-transformed branch lengths of congruent branches from matched NJst and concatenated likelihood analyses for the conditions under which support was maximized for each method (concatenated: 44-taxa, 50% missing taxa; NJst: 16-taxa, 50% missing taxa; see Results). We found a strong positive relationship between congruent branch lengths in all cases (e.g., 44-taxa strategy, 50% missing taxa; $r^2 = 0.81$, $P < 0.001$; 16 taxa strategy, 50% missing taxa; $r^2 = 0.90$, $P < 0.001$). We therefore considered NJst branch lengths to be potentially meaningful.

Testing the Test: Relating Mean Support and Accuracy

In this study, we used mean branch support for interfamilial relationships as our criterion for comparing sampling strategies. A potential issue is that the support for these highly uncertain relationships may not reflect the accuracy of a method or sampling strategy. To address this issue, we took advantage of the fact that the monophyly of all iguanian families is now well established, as is the split between acrodonts and pleurodonts. For example, most families are supported by morphological data (e.g., Frost and Etheridge 1989), by coalescent-based species-tree analyses of 29 loci (Townsend et al. 2011), and by concatenated analyses with extensive taxon sampling (e.g., Blankers et al. 2013; Pyron et al. 2013). We considered these relationships to be effectively known, and used the proportion

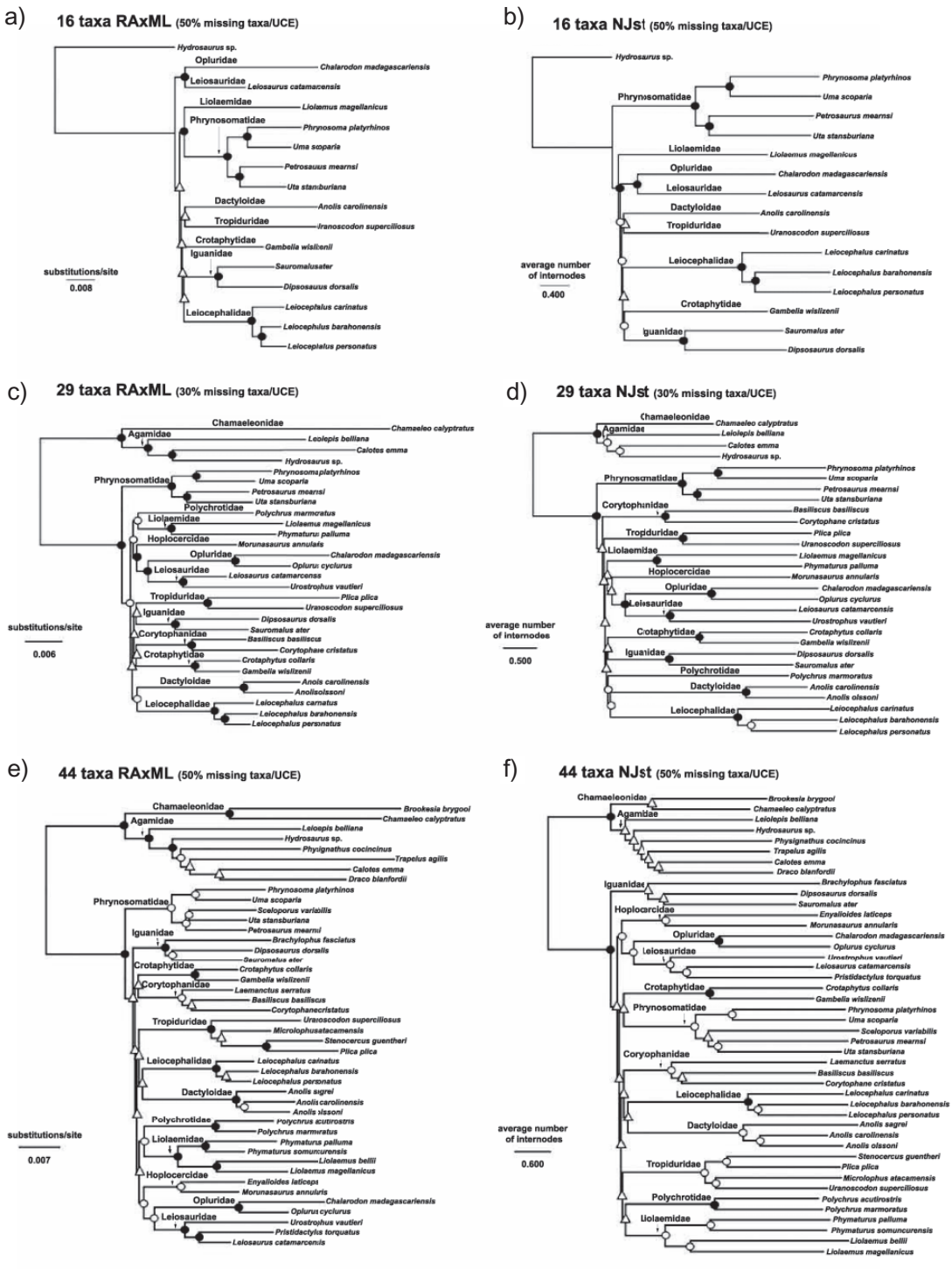

FIGURE 3. For each phylogenetic method (NJst, concatenated likelihood) and level of taxon sampling (16, 29, and 44 taxa), five data sets were generated with different numbers of loci included, each associated with different levels of missing taxa per gene (including genes with up to 20%, 30%, 40%, 50%, and 60% missing taxa per gene). Each tree shown represents one of the five data sets chosen to illustrate salient differences and similarities between phylogenetic methods. These select data sets ranged in size from 1762 to 4319 UCEs. a) Concatenated analysis of 16 taxa (up to 50% missing taxa per gene), b) NJst species tree analysis of 16 taxa (50% missing taxa), c) concatenated analysis of 29 taxa (30% missing taxa), d) NJst species tree analysis of 29 taxa (30% missing taxa), e) concatenated analysis of 44 taxa (50% missing taxa), and f) NJst species-tree analysis of 44 taxa (50% missing taxa). Support values are summarized across the five data sets as follows: black circles indicate nodes that were recovered in all analyses of the five data sets with support more than 90%. White circles indicate nodes that were recovered in all analyses of the five data sets but where support was sometimes less than 90%. White triangles indicate nodes where the depicted relationship was not recovered in one or more of the analyses of the five data sets. See Supplementary Figures S2–S31, available on Dryad at http://dx.doi.org/10.5061/dryad.13t18, for topologies and support values from individual analyses. Support from 44 taxa, 20% missing taxa data sets (Supplementary Figs. S22 and S27, available on Dryad at http://dx.doi.org/10.5061/dryad.13t18) is not considered here because *Draco blanfordii* and *Sceloporus variabilis* were not included in these data sets (due to too few loci).

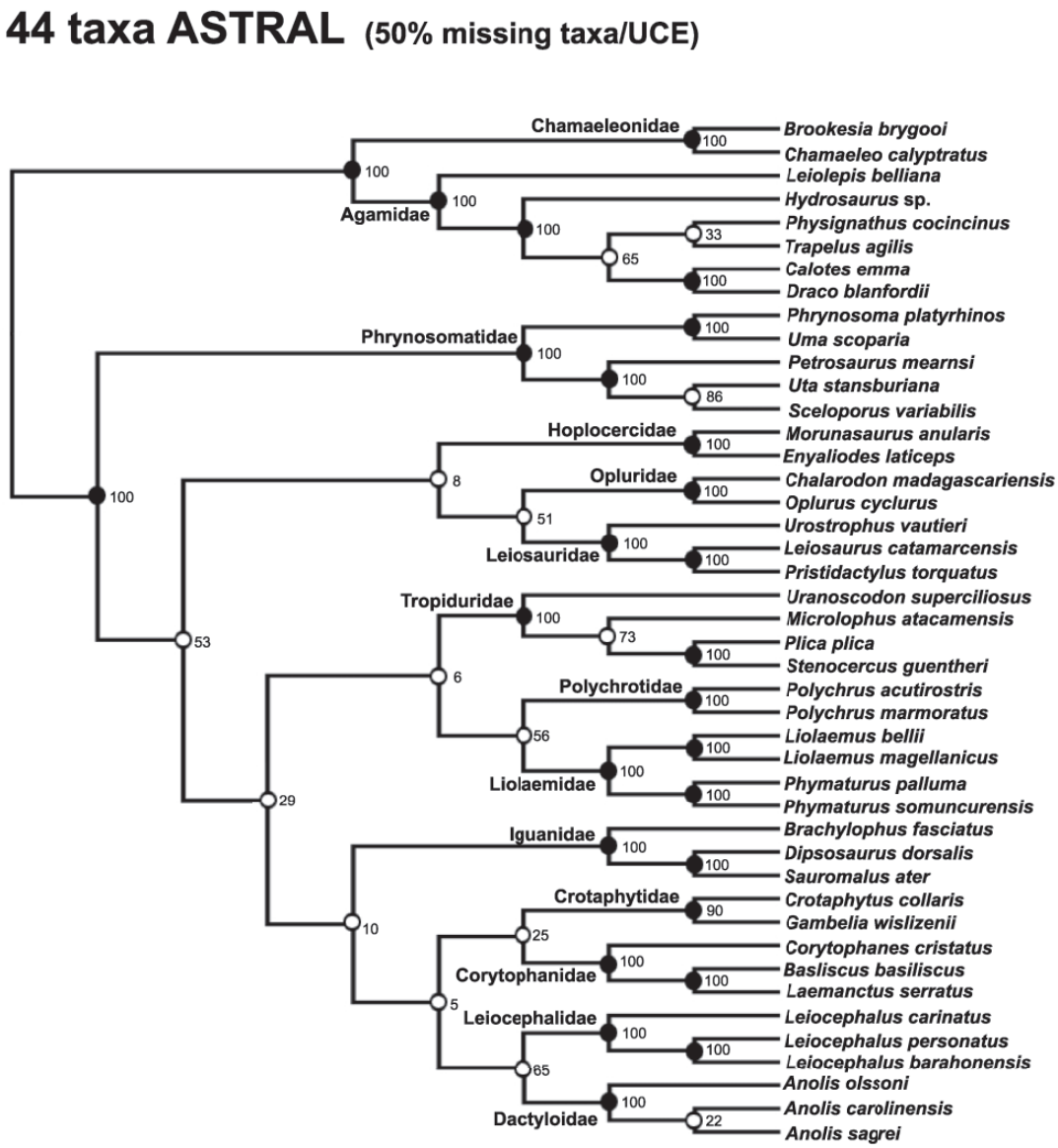

FIGURE 4. Species-tree analysis of 2716 UCEs (44 taxa, up to 50% missing taxa per gene tree) based on ASTRAL. Bootstrap support values appear adjacent to branches. In contrast to NJst, ASTRAL branch lengths all equal 1, and we therefore present the resulting species tree as a cladogram. Black circles indicate nodal support more than 90% and white circles indicate nodes where support was less than 90%.

of these known nodes recovered in an analysis as an index of accuracy (e.g., Wiens and Tiu 2012; an approach with a long history for assessing accuracy of methods: e.g., Hillis 1995). Therefore, for a given method, we tested whether mean support for the interfamilial relationships was correlated with accuracy, across the 15 sampling strategies. We also tested whether mean branch support for known branches (monophyly of families) was correlated with mean support for unknown branches (relationships among families) across the different conditions examined. We argue that a method and sampling strategy that is accurate under a given set of conditions should yield strong support for known relationships. For these analyses, the number of known nodes varied somewhat, depending on the taxon sampling (e.g., with 16 taxa, most families are represented by a single species), but this was not problematic because we used the proportion of known nodes recovered, not the absolute number. Note that we did not use this index of accuracy for comparing strategies in general because most well-established clades were correctly recovered under most conditions.

These analyses upheld the use of mean support for interfamilial relationships as an optimality criterion for comparing sampling strategies. Support for interfamilial relationships was significantly and positively correlated with accuracy for NJst ($\rho = 0.60$; $P = 0.020$). This relationship was positive but not significant for concatenated likelihood ($\rho = 0.43$; $P = 0.107$), apparently

because accuracy was very high and therefore almost completely invariant for most conditions. Both methods showed strong, positive correlations between mean branch support for well-established clades and those for interfamilial relationships (concatenated likelihood: $\rho = 0.70$; $P = 0.004$; NJst: $\rho = 0.84$, $P < 0.001$). Thus, we assume that analyses that yield stronger support for interfamilial relationships are more likely to yield more accurate results overall (in addition to better resolving the relationships of interest).

Results

Results of the phylogenetic analyses are summarized in Figs. 3 and 4. All trees (and branch support values) are shown individually in Supplementary Figs. S2–S37, available on Dryad at http://dx.doi.org/10.5061/dryad.13t18. Patterns of support, character sampling, gene sampling, missing data, and toplogical congruence between methods across the different sampling strategies are summarized in Fig. 5. We compare our results to those of other recent studies of pleurodont phylogeny in the Discussion.

Comparison of trees across methods and conditions shows that the choice of these sampling strategies is critically important. For example, species-tree analyses with 16 taxa show strong support (97–100%) for placing Phrynosomatidae as sister to all other pleurodonts (Fig. 3b and Supplementary Figs. S7–S11, available on Dryad at http://dx.doi.org/10.5061/dryad.13t18), as do many other species-tree and concatenated analyses. In contrast, species-tree analyses with 44 taxa can show strong support (up to 90%) for placing Iguanidae as sister to all other pleurodonts instead (Fig. 5b and Supplementary Figs. S29–S31, available on Dryad at http://dx.doi.org/10.5061/dryad.13t18). Thus, regardless of which topology is actually correct, at least one sampling strategy clearly leads to estimating an incorrect tree with strong branch support (even when using the same phylogenetic method), because both cannot be right.

Maximizing Support in Concatenated Analyses

Among concatenated analyses, likelihood bootstrap support for relationships among families was maximized with up to 50% missing taxa per locus, and with 44 taxa sampled overall (Fig. 5a). The next highest mean nodal support was observed with 44 taxa and up to 60% missing taxa per locus. Importantly, these are not the conditions with the most characters or genes sampled (Fig. 5c). In fact, branch support was maximized with less than half the maximum number of characters included (1,053,473 bp versus the maximum of 2,422,778 bp), suggesting that the benefits of greater taxon sampling outweigh those of greater character sampling for these concatenated analyses. Interestingly, support was highest in the analytical treatments that included close to the maximum amount of missing taxa (50%), but not the maximum amount (60%; Fig. 5a,e).

Remarkably, rather than finding that increasing missing data weakened branch support, we found a strong positive relationship between missing data (% total missing data cells in each matrix) and mean nodal support values ($r^2 = 0.36$, $P = 0.018$). This relationship occurs (at least in part) because of a strong positive relationship between the total number of characters included in each matrix and the maximum percentage of missing taxa allowed for a gene to be included ($\rho = 0.72$, $P = 0.002$). Although not significant, we also observed positive relationships between mean nodal support and the total number of characters ($\rho = 0.44$, $P = 0.102$) and between the total percentage of missing data per matrix and the total number of characters in that matrix ($\rho = 0.22$, $P = 0.43$).

We ran RAxML GTR+Γ with nonparametric bootstrapping on the conditions that maximized support in the GTRCAT analyses (50% missing taxa [29, 44 taxa], 60% missing taxa [16 taxa]; Supplementary Figs. S32–S34, available on Dryad at http://dx.doi.org/10.5061/dryad.13t18). In comparing support between GTRCAT and GTR+Γ search strategies, we found that GTR+Γ had slightly lower mean support than GTRCAT and that support was maximized under the same conditions (Table 3). However, these differences were not statistically significant (Wilcoxon signed-rank tests [critical values at $P \leqslant 0.05$]: 16 taxa: $W = 7$, critical value $= 0$; 29 taxa: $W = 13$, critical value $= 5$; 44 taxa: $W = 30$, critical value $= 17$). Thus, there was no clear evidence of inflated support values in GTRCAT analyses.

Maximizing Support in Species-Tree Analyses

For species-tree analyses (here using NJst), mean branch support was maximized when including loci with up to 50% missing taxa each, but with minimal taxon sampling (16 species; Fig. 5b). This treatment included the second highest number of genes (4,789 UCEs; Fig. 5d). The next highest mean support was observed under the same conditions but including genes with up to 60% missing taxa each (which included the largest number of genes). Mean support for all 16-taxon analyses was similar across surprisingly large differences in gene sampling (1831–4921 UCEs).

There was a strong positive relationship between species-tree support values and number of genes included ($\rho = 0.83$, $P < 0.001$), regardless of missing data or taxa sampled. Across all 15 conditions, mean support was significantly lower than in concatenated analyses (nonparametric Wilcoxon signed-rank test: $Z = -2.669$, $P = 0.008$; Fig. 5a,b). With the exception of the 16-taxon case, mean nodal support generally increased when allowing more missing taxa per locus. Thus, there was an overall positive relationship between missing data and mean branch support ($\rho = 0.57$, $P = 0.027$). We also

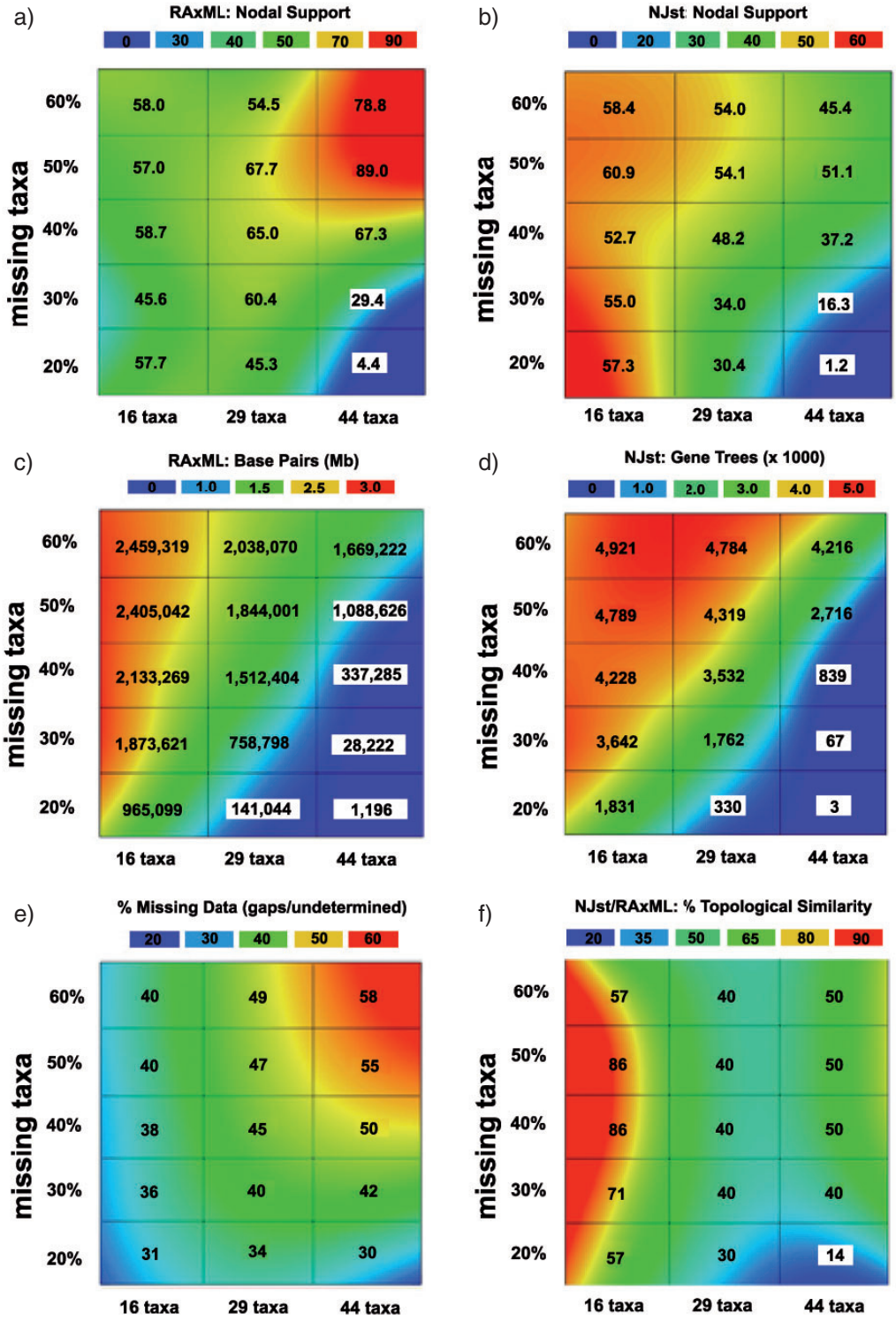

FIGURE 5. Heat maps for the 15 data sets (including genes with up to 20%, 30%, 40%, 50%, and 60% missing taxa per gene and including 16, 29, and 44 taxa overall) analyzed in this study, using two phylogenetic methods (concatenated likelihood: RAxML; species tree: NJst). Heat maps show increasing values of (a) mean nodal support for maximum-likelihood concatenated analyses in RAxML for 7 (16 taxa) and 10 between-family focal nodes (29, 44 taxa), (b) as in A, but showing mean nodal support for species-tree analyses in NJst, (c) total number of nucleotides (in base pairs) analyzed for each concatenated data set, (d) total number of gene trees (individual UCEs) analyzed for each species-tree data set, (e) total amount of missing data in percentage of gaps and undetermined sites in each concatenated data set, and (f) percentage of focal clades that are congruent between matched concatenated and species tree analyses. White boxes have been added around some values to aid with visualization.

TABLE 3. Comparison of select analyses with UCE sequences using concatenated and species tree methodologies

Method	Mean support for focal nodes		
	16 taxa	29 taxa	44 taxa
RAxML GTRCAT + rapid bootstrapping	57.0	67.7	89.0
RAxML GTR+Γ nonparametric bootstrapping	59.6	64.3	88.0
NJst Species-tree analysis	60.9	54.1	51.1
ASTRAL Species-tree analysis	51.9	49.5	43.6

Note: All analyses allowed for up to 50% missing taxa/UCE locus.

found a strong positive relationship between missing data and the number of genes included ($\rho = 0.79$, $P < 0.001$).

For each taxon sampling strategy, we ran ASTRAL on the conditions that maximized support for NJst analyses (50% missing taxa per locus; Fig. 4 and Supplementary Figs. S35–S37, available on Dryad at http://dx.doi.org/10.5061/dryad.13t18). In comparing support levels between NJst and ASTRAL, we found that ASTRAL consistently had lower support than NJst (Table 3). However, support was maximized under the same conditions using both methods (16 taxon strategy).

Comparing Support and Congruence between Concatenated and Species-Tree Analyses

Across the different conditions examined, support values were significantly and positively related between concatenated and species-tree analyses ($r^2 = 0.47$, $P = 0.005$). Thus, conditions that yielded high mean support for one method tended to yield high support for the other.

Congruence between RAxML and NJst (in terms of the proportion of between-family clades that were identical; Fig. 5f) was maximized under conditions in which the species-tree method had highest average support (16 taxa, maximum of 40–50% missing taxa per UCE). Congruence also was relatively high under conditions when concatenated analyses had the highest support (44 taxa, maximum of 40–60% missing taxa per UCE). However, congruence was moderate under almost all conditions, and never higher than 57% for the between-family nodes. We found that branches on which concatenated and NJst species-tree analyses were concordant were significantly longer than branches on which they disagreed (Table 4). This was true using branch lengths estimated by both RAxML and NJst, and for the conditions under which each method had maximum mean branch support.

DISCUSSION

Sampling Strategies for Phylogenomic Analysis

Phylogenomic analyses are creating exciting opportunities to resolve important phylogenetic questions (McCormack and Faircloth 2013). These methods also are raising (or exacerbating) questions about study design, including the relative importance of increasing sampling of characters and taxa and minimizing missing data. In this article, we have explored the consequences of different sampling designs for branch support for a difficult phylogenetic problem (between-family relationships of pleurodont lizards, an ancient, rapid radiation).

Our analyses suggest that the best sampling design depends on the phylogenetic method used (Fig. 5), but that a strong general pattern emerges nonetheless. Specifically, for both concatenated and species-tree analyses, branch support was maximized when we allowed up to 50% missing taxa per gene (Fig. 5a,b). Moreover, we found a positive relationship between the amount of missing data per matrix and mean branch support values, for both methods. This does not occur because missing data cells are beneficial in any way, but instead because allowing more missing data allows for the inclusion of more genes and more taxa. We suggest that empirical researchers should examine the consequences of including and excluding genes and taxa with missing data on their results, rather than simply excluding these data a priori without justification (e.g., Leaché et al. 2014). Our results here, and those of many other empirical and simulation studies based on much smaller data sets (e.g., Wiens and Morrill 2011; Wiens and Tiu 2012; Roure et al. 2013; Jiang et al. 2014), do not support the practice of excluding genes and taxa solely because they are not complete.

In theory, highly incomplete taxa might contribute to weak support for the interfamilial relationships, given their potential for unstable placement. However, weak support does not appear to be caused by the unstable placement of taxa with few genes. For example, the most incomplete taxon in our data set (*Sceloporus variabilis*, with only 120 loci) is consistently placed in the appropriate family (Phrynosomatidae) with strong branch support. Therefore, weak support for relationships among families clearly is not being caused by this highly incomplete taxon.

On the other hand, our results suggest that branch support is not generally maximized with the maximum amount of missing data either. In some cases, there were decreases in branch support when including genes with up to 60% missing taxa relative to results including genes with a maximum of 50% missing taxa. These decreases in mean branch support occurred despite substantial increases in the number of genes and characters sampled, especially for 44 taxa. Thus, our results suggest that support is maximized when allowing for a large, but still intermediate, level of missing data (up to 50% of taxa missing per gene).

TABLE 4. Comparison of branch lengths of branches that are either congruent or incongruent between estimated trees from concatenated and species-tree approaches

Condition	Branch category	*N*	Mean length (± SE)	χ^2	*P*
16 taxa, 50% missing taxa concatenated	Congruent	23	0.017 ± 0.002	13.80	< 0.0001
	Incongruent	6	0.001 ± 0.004		
16 taxa, 50% missing taxa, species tree	Congruent	23	1.250 ± 0.109	13.80	< 0.0001
	Incongruent	6	0.031 ± 0.214		
44 taxa, 50% missing taxa, concatenated	Congruent	76	0.010 ± 0.001	21.95	< 0.0001
	Incongruent	10	0.001 ± 0.002		
44 taxa, 50% missing taxa, species tree	Congruent	76	1.434 ± 0.088	23.86	< 0.0001
	Incongruent	10	0.076 ± 0.242		

Notes: Results are based on conditions where support was maximized for each method (16 taxa for species tree, 44 taxa for concatenated, up to 50% missing taxa per gene for both methods). For each set of conditions, the upper set of results are based on branch lengths estimated by the concatenated likelihood analysis, whereas the lower set of results are based on branch lengths estimated by the species-tree method (NJst). Statistical results are from the Kruskal–Wallis test.

Despite the general concordance between the methods (i.e., support maximized at 50% missing taxa per gene), there were also important differences. Specifically, for concatenated analyses, mean branch support (maximum-likelihood bootstrap values for between-family relationships) was maximized when including the maximum number of taxa and an intermediate number of genes (44 taxa, 2716 genes; Fig. 5). For the species-tree method used here (NJst), support was maximized with minimal taxon sampling and extensive sampling of genes (16 taxa, 4769 genes; Fig. 5). While some might find it dissatisfying that the exact same conditions are not optimal for all methods, we think that this is an important caution for study design.

Our results raise many important issues for future studies. First, our results suggest that the best sampling design for maximizing clade support may depend on the method used, which raises the question: which method should be used? Given that many simulation studies have suggested that species-tree methods are more accurate than concatenated analyses (e.g., Edwards et al. 2007; Heled and Drummond 2010; Leaché and Rannala 2011; Liu and Yu 2011), it might seem that the only consideration in sampling design should be finding the conditions that maximize performance of the species-tree method. However, many species-tree methods have only been tested under limited conditions, and their performance relative to concatenated analyses remain uncertain. For example, Liu and Yu (2011) tested the relative accuracy of Bayesian concatenated analysis and the NJst species-tree method used here, but only in the five-taxon case and with internal branches that were similar in length to the terminal branches (i.e., not an ancient, rapid radiation with short internal branches and deep terminal branches). Even under these relatively "easy" conditions, concatenated analysis was more accurate in some cases (i.e., longer branches). Furthermore, although NJst and ASTRAL are able to include genes with missing data (in principle and shown in practice here), it is not clear if they remain more accurate than concatenated analysis under these conditions. Interestingly, Mirarab et al. (2014) recently demonstrated that concatenated analyses can be more accurate than species-tree summary methods under scenarios where gene trees have poor phylogenetic signal.

Our results here may also shed some light on this issue. If we consider Pleurdonta and the pleurodont families as "known" clades, then the proportion of these clades that are supported in a given analysis can be used as an index of accuracy, along with the support for these clades (see Methods). In our study, concatenated likelihood analysis has slightly higher accuracy and significantly higher support for these established clades than does NJst (mean accuracy = 0.96 vs. 0.93; mean support = 95 vs. 91; using nonparametric Wilcoxon signed-rank test: accuracy: $Z = -1.342$, $P = 0.180$; support: $Z = -2.032$, $P = 0.042$). It is also notable that mean branch support for interfamilial relationships was substantially higher for concatenated analysis under its optimal conditions than for NJst under its best conditions (mean 89 vs. 61) and that support for known and unknown clades are significantly correlated for both approaches (see Methods).

The effect of limited taxon sampling on the accuracy of species-tree methods is another neglected but potentially critical issue for future studies. Studies addressing the accuracy of species-tree methods have generally focused on the sampling of loci and/or individuals within species (e.g., Edwards et al. 2007; Heled and Drummond 2010; Liu and Yu 2011). However, limited taxon sampling might also have important consequences for these methods, especially for studies of species-rich clades in which relatively few species are included overall. For example, in an empirical analysis of an ancient, species-rich clade of lizards (Scincidae), Lambert et al. (2015) found that a moderate reduction in taxon sampling (30%) led to reduced support values from a coalescent species-tree method (*BEAST), despite a large increase in loci sampled (from 10 to 44 genes). Limited taxon sampling in higher-level studies might lead to inaccurate gene trees, which might mislead both species-tree and concatenated analyses (Lambert et al. 2015). We also note that in our study congruence between the trees from these methods

is high in the 16-taxon case, but many of the congruent clades disappear with greater taxon sampling for both methods.

Conclusions from our study should be interpreted with several caveats in mind. First, because this is an empirical study of a group in which the relationships are highly uncertain, the true phylogeny is not known. Therefore, we focused on finding the sampling conditions under which branch support is maximized for each method (given that poorly supported branches may not be resolved correctly). We acknowledge that branch support and accuracy could be decoupled under some conditions (where we define accuracy as the similarity between estimated and true topologies), but our analyses focusing on well-established clades suggest that there is a strong relationship between support for "known" and "unknown" clades for the conditions examined here (see Methods). Nevertheless, we note at least one case where there is strong branch support for a clade that is then strongly rejected by the same method under a different sampling strategy (see above). Furthermore, the relationship between accuracy and support values for NJst has not been explicitly studied. Second, our sampling interval only ranged from 20% to 60% missing taxa per gene and 30% to 58% missing data cells per matrix. Thus, it is possible that our conclusions (and topologies) might change if we included data sets with higher or lower levels of missing data. We already show that branch support seems to decline somewhat with the highest levels of missing taxa, and we expect this trend would continue with even more missing data. For our data, reducing missing cells even further would require further decreasing the sampling of taxa and loci. In future studies, it would be interesting to begin with a very large, very complete UCE data set for a well-resolved group, experimentally subsample the data, and explore the ability of different sampling strategies to recover well-known relationships (although such well-known relationships might be the least relevant for most empirical studies). It also would be interesting to conduct similar analyses using other species-tree methods (but note that other methods may be limited in the number of loci they can include and whether they allow missing data).

Finally, we note that in this study we modified the standard UCE protocol (Faircloth et al. 2012) so as to obtain data from more samples per sequence capture. One concern is that this approach might explain some of the incompleteness of our largest data sets (i.e., those including the most genes and taxa). However, we note that we were able to generate data for all 48 samples used in each capture, and that if this protocol is refined for more even enrichment it would represent a substantial cost-reduction for targeted sequence capture (TSC) studies. As previously mentioned, our experiments using fewer individuals per sequence capture did not result in obtaining substantially more UCEs per sample. Instead, it seems more likely that other factors might explain some of the incompleteness in our data. For example, a recent experiment where we sheared DNA using a sonicator (Biorupter, Diagenode) instead of enzymatic shearing (used in this study) revealed that the average number of UCEs obtained per sample increased by almost 50%. Thus, it is possible that enzymatic shearing is suboptimal for targeted enrichment.

Implications for Pleurodont Phylogeny

Our study represents the first attempt to resolve pleurodont relationships with next-generation sequencing methods. The results might be considered somewhat disappointing in that relationships among families are often weakly supported, and differ considerably depending on the combination of phylogenetic method and sampling strategy used. Moreover, even using the same method, relationships that are strongly supported under one sampling strategy may be contradicted (with strong support) under another. Nevertheless, several interesting patterns appear in most analyses. Furthermore, if we use our results on selecting the "best" sampling strategy to choose our estimate of pleurodont relationships (following from our tests of accuracy, and the criterion of highest mean support used throughout the paper), then most pleurodont relationships are resolved with relatively strong support, with many nodes agreed on by concatenated and species-tree anayses (Fig. 6). We describe these patterns below.

We find five major clades that emerge with some consistency across the analyses. (1) We find that the southern South American family Leiosauridae and the Madagascar family Opluridae are supported as sister taxa in almost every analysis, often with strong support. This relationship has appeared in nearly every recent phylogenetic analysis of pleurodont relationships (e.g., Townsend et al. 2011; Wiens et al. 2012; Blankers et al. 2013; Pyron et al. 2013). (2) We find Hoplocercidae is often placed as the sister group to Leiosauridae + Opluridae, with strong support in concatenated and species-tree analyses with extensive taxon sampling (29 and 44 loci). In many previous analyses, Liolaemidae is instead placed as sister to this clade (e.g., Wiens et al. 2012; Blankers et al. 2013), sometimes with strong support (Townsend et al. 2011; Pyron et al. 2013). (3) In our analyses, Liolaemidae is often placed with Polychrotidae (concatenated and species-tree analysis with 44 taxa), and these two families are often placed with the clade of Leiosauridae, Opluridae, and Hoplocercidae. These relationships are largely unprecedented based on previous analyses. (4) In almost all analyses with 29 and 44 taxa (both concatenated and species tree), the anoles (Dactyloidae) are placed with Leiocephalidae. These relationships are largely unprecedented based on previous studies (but see Schulte et al. 2003). (5) Finally, many analyses suggest that Phrynosomatidae is sister to all other pleurodonts. These include concatenated analyses with 29 and 44 taxa, the NJst species-tree analyses with 16 and 29 taxa, and the ASTRAL species-tree analyses

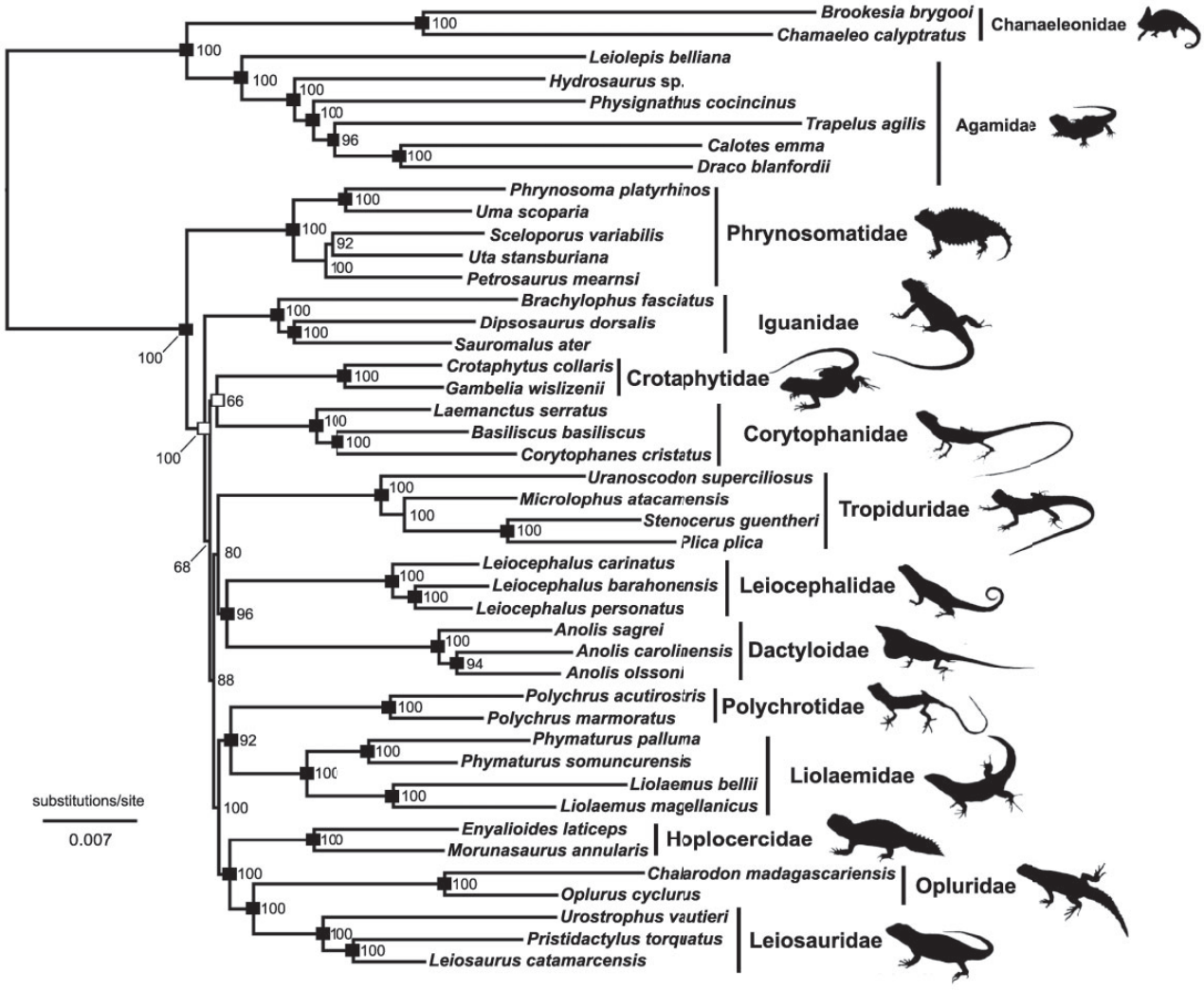

FIGURE 6. Preferred estimate of iguanian lizard relationships from this study. The tree is a phylogram from the concatenated likelihood analysis of the 44-taxon data set, including loci with up to 50% missing taxa per locus (2716 UCEs, total of 1,088,626 base pairs in the concatenated alignment). Numbers at nodes indicate bootstrap support. Black squares indicate clades that were also recovered by the matched NJst species-tree analysis, with 44 taxa and allowing up to 50% missing taxa per locus (Fig. 3f). White squares indicate clades that were recovered in other NJst species-tree analyses (but not in the one based on 44 taxa and up to 50% missing taxa per locus).

with 16, 29, and 44 taxa. This relationship was strongly supported in concatenated analyses by Townsend et al. (2011) with 29 nuclear loci, but was not supported in their species-tree analyses nor subsequent concatenated analyses (e.g., Wiens et al. 2012; Blankers et al. 2013; Pyron et al. 2013). We note that when examining the number of duplicate UCE contigs removed by *phyluce* (a putative measure of contamination), two of the samples with the greatest number of UCEs removed were phrynosomatids (*Petrosaurus* and *Uta*; Table 1). However, there were also tropidurids and leiosaurids with similar levels of UCE duplicates, and the strong support for monophyly of these families also argues against the impacts of contamination.

In terms of highest overall nodal support, our best estimate of pleurodont relationships is represented by the 44-taxon concatenated analysis with up to 50% missing taxa allowed per gene (Fig. 6). This tree provides relatively strong support for most relationships among families, with all 11 nodes with support $\geqslant$65%, 9 of 11 nodes $\geqslant$80%, and 7 of 11 nodes $\geqslant$90%. Furthermore, many of these relationships are also consistent with those in many species-tree analyses, including the five relationships described above. Intriguingly, these relationships are consistent with the hypothesis that the ancestor of pleurodonts dispersed from the Old World (i.e., where all acrodonts presently occur) into North and Middle America (i.e., where most phrynosomatids presently occur, along with crotaphytids, and many corytophanids and iguanids), and subsequently spread into South America (where hoplocercids, leiosaurids, liolaemids, polychrotids, and tropidurids occur), and

then to Madagascar (where oplurids are endemic), as discussed by Townsend et al. (2011).

Finally, although there is some disagreement among analyses, many relationships are highly congruent between concatenated and species-tree analyses. These include the monophyly of all 12 pleurodont families, many relationships within families (i.e., in the 44-taxon case), and most relationships among acrodont lineages (Figs. 3 and 4). These congruent nodes seem to be associated with longer branches, perhaps related to greater congruence between genes on longer branches (e.g., Wiens et al. 2012; Lambert et al. 2015).

Conclusions

In this study, we compare the results from different sampling strategies for loci and taxa in an empirical phylogenomic analysis of an ancient, rapid radiation. We show that allowing more missing data can increase the number of taxa and loci that are included, and increase support for estimated relationships (but that including the maximum amount of missing data does not necessarily maximize support). We also show that maximum support for inferred relationships is obtained from different sampling strategies for concatenated analyses (maximum taxon sampling, moderate sampling of loci) and species-tree analyses (minimum taxon sampling, extensive sampling of loci). We recognize that decisions about the optimal sampling strategy also should include simulations, where the true topology is known. Nevertheless, our study reveals an important finding that may not be widely appreciated: that different sampling strategies can have important impacts on estimated relationships, in some cases as much or more than the method itself. For example, using the same method (e.g., NJst) very different relationships can be obtained with different sampling strategies, each with strong support (e.g., iguanids vs. phrynosomatids as sister to other pleurodonts; Fig. 3b,d, and f). Thus, we show that some sampling strategies must be yielding incorrect but strongly supported results. While this sensitivity may be largely confined to short branches in this ancient, rapid radiation, it is just such branches that phylogenomic data may be needed to resolve. Finally, our study provides concrete recommendations regarding optimal sampling strategies, and provides a generally well-supported hypothesis of pleurodont relationships.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.13t18.

Acknowledgments

The authors thank the numerous individuals and institutions whose provision of tissue samples made this study possible, including T. Reeder, L. Avila, S. Poe, S. Blair Hedges, B. Noonan, J. Campbell, R. Etheridge, J. McGuire, J. Kolbe, T. Sanger, K. Nicholson, J. Valladares, and the American Museum of Natural History, California Academy of Sciences, Louisiana State University Museum of Zoology, Museum of Vertebrate Zoology (University of California at Berkeley), University of Kansas Museum of Natural History, University of Michigan Museum of Zoology, and the Yale Peabody Museum. They thank G. Mathews for laboratory assistance. J. Castoe provided invaluable assistance with adapter design and high-throughput library preparation. They thank S. Lambert for discussion and the University of Arizona for financial support. P. Foster kindly provided assistance with computing resources at the Natural History Museum. They sincerely thank F. Anderson, B. Faircloth, J. McCormack, and an anonymous reviewer for comments and suggestions that improved the quality of the manuscript.

References

Alföldi J., Di Palma F., Grabherr M., Williams C., Kong L., Mauceli E., Russell P., Lowe C.B., Glor R.E., Jaffe J.D., Ray D.A., Boissinot S., Shedlock A.M., Botka C., Castoe T.A., Colbourne J.K., Fujita M.K., Godinez Moreno R., ten Hallers B.F., Haussler D., Heger A., Heiman D., Janes D.E., Johnson J., de Jong P.J., Koriabine M.Y., Lara M., Novick P.A., Organ C.L., Peach S.E., Poe S., Pollock D.D., de Queiroz K., Sanger T., Searle S., Smith J.D., Smith Z., Swofford R., Turner-Maier J., Wade J., Young S., Zadissa A., Edwards S.V., Glenn T.C., Schneider C.J., Losos J.B., Lander M.B., Ponting C.P., Lindblad-Toh K. 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature 477:587–591.

Arnold B., Corbett-Detig R.B., Hartl D., Bomblies K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol. Ecol. 22:3179–3190.

Bejerano G., Pheasant M., Makunin I., Stephen S., Kent W.J., Mattick J.S., Haussler D. 2004. Ultraconserved elements in the human genome. Science 304:1321–1325.

Blankers T., Townsend T.M., Pepe K., Reeder T.W., Wiens J.J. 2013. Contrasting global-scale evolutionary radiations: Phylogeny, diversification, and morphological evolution in the major clades of iguanian lizards. Biol. J. Linn. Soc. 108:127–143.

Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1932.

Cariou M., Duret L., Charlat S. 2013. Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. Ecol. Evol. 3:846–852.

Crawford N.G., Faircloth B.C., McCormack J.E., Brumfield R.T., Winker K., Glenn T.C. 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. 8:783–786.

Crawford N.G., Parham J.F., Sellas A.B., Faircloth B.C., Glenn T.C., Papenfuss T.J., Henderson J.B., Hansen M.H., Simison W.B. 2015. A phylogenomic analysis of turtles. Mol. Phylogenet. Evol. 83:250–257.

Driskell A.C., Ane 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.

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

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

Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements anchor

thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61:717–726.
Faircloth B.C., Sorenson L., Santini F., Alfaro M.E. 2013. A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). PLoS One 8:e54848.
Faircloth B.C., Branstetter M.G., White N.D., Brady S.G. 2014. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Mol. Ecol. Res. 15:489–501.
Frost D.R., Etheridge R. 1989. A phylogenetic analysis and taxonomy of iguanian lizards (Reptilia: Squamata). Misc. Publ. Mus. Nat. Hist. Univ. Kansas 81:1–65.
Frost D.R., Etheridge R., Janies D., Titus T.A. 2001. Total evidence, sequence alignment, evolution of polychrotid lizards, and a reclassification of the Iguania (Squamata: Iguania). Am. Mus. Nov. 3343:1–38.
Graybeal A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. 47:9–17.
Heath T.A., Hedtke S.M., Hillis D.M. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J. Syst. Evol. 46:239–257.
Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.
Hillis D.M. 1995. Approaches for assessing phylogenetic accuracy. Syst. Biol. 44:3–16.
Hillis D.M., Bull J.J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42:182–192.
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.
Huang H., Knowles L.L. 2014. Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of RAD sequences. Syst. Biol. 10.1093/sysbio/syv046, in press.
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.
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.
Leaché A.D., Rannala B. 2011. The accuracy of species tree estimation under simulation: A comparison of methods. Syst. Biol. 60:126–137.
Leaché A.D., Wagner P., Linkem C.W., Böhme W., Papenfuss T.J., Chong R.A., Lavin B.R., Bauer A.M., Nielsen S.V., Greenbaum E., Rödel M-O., Schmitz A., LeBreton M., Ineich I., Chirio L., Ofori-Boateng C., Eniang E.A., Baha El Din S., Lemmon A.R., Burbrink F.T. 2014. A hybrid phylogenetic-phylogenomic approach for species tree estimation in African *Agama* lizards with applications to biogeography, character evolution, and diversification. Mol. Phylogenet. Evol. 79:215–230.
Leaché A.D., Chavez A.S., Jones L.N., Grummer J.A., Gottscho A.D., Linkem C.W. 2015. Phylogenomics of phrynosomatid lizards: Conflicting signals from sequence capture versus restriction site associated DNA sequencing. Genome Biol. Evol. 7:706–719.
Lemmon E.M., Lemmon A.R. 2013. High-throughput genomic data in systematics and phylogenetics. Annu. Rev. Ecol. Evol. Syst. 44: 99–121.
Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61:727–744.
Lemmon A.R., Brown J.M., Stranger-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.
Liu L., Yu L. 2011. Estimating species trees from unrooted gene trees. Syst. Biol. 60:661–667.
Liu L., Yu L., Edwards S.V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10:302.
Liu L., Yu L., Pearl D.K., Edwards S.V. 2009. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58:468–477.
McCormack J.E., Faircloth B.C. 2013. Next-generation phylogenetics takes root. Mol. Ecol. 22:19–21.
McCormack J.E., Faircloth B.C., Crawford N.G., Gowaty P.A., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species tree analysis. Genome Res. 22:746–754.
McCormack J.E., Harvey M.G., Faircloth B.C., Crawford N.G., Glenn T.C., Brumfield R.T. 2013. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PLoS One 8:e54848.
Miller W., Rosenbloom K., Hardison R.C., Hou M., Taylor J., Raney B., Burhans R., King D.C., Baertsch R., Kosakovsky-Pond S.L., Nekrutenko A., Giardine B., Harris R.S., Tyekucheva S., Diekhans M., Pringle T.H., Murphy J., Lesk A., Weinstock G.M., Linblad-Toh K., Gibbs R.A., Lander E.S., Siepel A., Haussler D., Kent W.J. 2007. 28-way vertebrate alignment and conservation track in the UCSC genome browser. Genome Res. 17:1797–1808.
Miller M.A., Pfeiffer W., Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Proceedings of the Gateway Computing Environments Workshop (GCE). New Orleans (LA): IEEE. p. 1–8.
Mirarab S., Bayzid M.S., Warnow T. 2014. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. 10.1093/sysbio/syv063, in press.
Mirabab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014. ASTRAL: Genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541–i548.
Noonan B.P., Chippindale P.T. 2006. Vicariant origin of Malagasy reptiles supports Late Cretaceous Antarctic landbridge. Am. Nat. 168:730–741.
Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.
Philippe H., Snell E.A., Bapteste E., Lopez P., Holland P.W.H., Casane D. 2004. Phylogenomics of eukaryotes: Impacts of missing data on large alignments. Mol. Biol. Evol. 21:1740–1752.
Poe S., Swofford D.L. 1999. Taxon sampling revisited. Nature 398: 299–300.
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.
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. Jr., Sites J.W. Jr., Wiens J.J. 2015. Integrated analyses resolve conflicts over squamate reptile phylogeny and reveal unexpected placements for fossil taxa. PLoS One 10:e0118199.
Rohland N., Reich D. 2012. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res. 22:939–946.
Rosenberg M.S., Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. USA. 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.
Rubin B.E.R., Ree R.H., Moreau C.S. 2012. Inferring phylogenies from RAD sequence data. PLoS One 7:e33394.
Sandelin A., Bailey P., Bruce S., Engström P.G., Klos J.M., Wasserman W.W., Ericson J., Lenhard B. 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics 5:99.
Schulte J.A. II, Valladares J.P., Larson A. 2003. Phylogenetic relationships within Iguanidae inferred using molecular and morphological data and a phylogenetic taxonomy of iguanian lizards. Herpetologica 59:399–419.
Seo T.K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol. Biol. Evol. 25:960–971.
Shaw T., Ruan Z., Glenn T., Liu L. 2013. STRAW: Species TRee Analysis Web Server. Nucleic Acids Res. 41:W230–W241.
Simmons M.P., Norton A.P. 2014. Divergent maximum-likelihood-branch-support values for polytomies. Mol. Phylogenet. Evol. 73: 87–96.

Sites J.W. Jr., Reeder T.W., Wiens J.J. 2011. Phylogenetic insights on evolutionary novelties in lizards and snakes: Sex, birth, bodies, niches, and venom. Ann. Rev. Ecol. Evol. Syst. 42:227–244.
Smith B.T., Harvey M.G., Faircloth B.C., Glenn T.C., Brumfield R.T. 2014. Target capture and massively parallel sequencing of ultraconserved elements (UCEs) for comparative studies at shallow evolutionary timescales. Syst. Biol. 63:83–95.
Stamatakis A. 2014. RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1315.
Streicher J.W., Devitt T.J., Goldberg C.S., Malone J.H., Blakmon H., Fujita M.K. 2014. Diversification and asymmetrical gene flow across time and space: Lineage sorting and hybridization in polytypic barking frogs. Mol. Ecol. 23:3273–3291.
Sun K., Meiklejohn K.A., Faircloth B.C., Glenn T.C., Braun E. L., Kimball R.T. 2014. The evolution of peafowl and other taxa with ocelli (eyespots): A phylogenomic approach. Proc. R. Soc. B 281: 20140.
Townsend T., Mulcahy D.G., Sites J.W. Jr., Kuczynski C.A., Wiens J.J., Reeder T.W. 2011. Phylogeny of iguanian lizards inferred from 29 nuclear loci, and a comparison of concatenated and species-tree approaches for an ancient, rapid radiation. Mol. Phylogenet. Evol. 61:363–380.
Uetz P., Hošek J. (eds). 2014. The Reptile Database. Available from: URL http://www.reptile-database.org (accessed 11 November 2014).
Wagner C.E., Keller I., Wittwer S., Selz O.M., Mwaiko S., Greuter L., Sivasundar A., Seehausen O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22:787–798.
Wiens J.J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst. Biol. 47:625–640.
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., Hutter C.R., Mulcahy D.G., Noonan B.P., Townsend T.M., Sites J.W. Jr., Reeder T.W. 2012. Resolving the phylogeny of lizards and snakes (Squamata) with extensive sampling of genes and species. Biol. Lett. 8:1043–1046.
Wiens J.J., Tiu J. 2012. Highly incomplete taxa can rescue phylogenetic analyses from the negative impacts of limited taxon sampling. PLoS One 7:42925.
Xi Z., Liu L., Rest J.S., Davis C.C. 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. Syst. Biol. 63:919–932.
Zerbino D.R., Birney E. 2008. Velvet: Algorithms for de novo short read assemble using de Bruijn graphs. Genome Res. 18:821–829.
Zhou L., Holliday J.A. 2012. Targeted enrichment of the black cottonwood (*Populus trichocapra*) gene space using sequence capture. BMC Genomics 13:703.
Zwickl D.J., Hillis D.M. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.