Syst. Biol. 0(0):1–22, 2020

DOI:10.1093/sysbio/syaa064

Do Alignment and Trimming Methods Matter for Phylogenomic (UCE) Analyses?

DANIEL M. PORTIK[1,2,*] AND JOHN J. WIENS[1]

[1]*Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA; and*
[2]*California Academy of Sciences, San Francisco, CA 94118, USA*
**Correspondence to be sent to: Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA; E-mail: daniel.portik@gmail.com.*

Received 15 May 2019; reviews returned 2 August 2020; accepted 3 August 2020
Associate Editor: Brant Faircloth

Abstract.—Alignment is a crucial issue in molecular phylogenetics because different alignment methods can potentially yield very different topologies for individual genes. But it is unclear if the choice of alignment methods remains important in phylogenomic analyses, which incorporate data from hundreds or thousands of genes. For example, problematic biases in alignment might be multiplied across many loci, whereas alignment errors in individual genes might become irrelevant. The issue of alignment trimming (i.e., removing poorly aligned regions or missing data from individual genes) is also poorly explored. Here, we test the impact of 12 different combinations of alignment and trimming methods on phylogenomic analyses. We compare these methods using published phylogenomic data from ultraconserved elements (UCEs) from squamate reptiles (lizards and snakes), birds, and tetrapods. We compare the properties of alignments generated by different alignment and trimming methods (e.g., length, informative sites, missing data). We also test whether these data sets can recover well-established clades when analyzed with concatenated (RAxML) and species-tree methods (ASTRAL-III), using the full data (∼5000 loci) and subsampled data sets (10% and 1% of loci). We show that different alignment and trimming methods can significantly impact various aspects of phylogenomic data sets (e.g., length, informative sites). However, these different methods generally had little impact on the recovery and support values for well-established clades, even across very different numbers of loci. Nevertheless, our results suggest several "best practices" for alignment and trimming. Intriguingly, the choice of phylogenetic methods impacted the phylogenetic results most strongly, with concatenated analyses recovering significantly more well-established clades (with stronger support) than the species-tree analyses. [Alignment; concatenated analysis; phylogenomics; sequence length heterogeneity; species-tree analysis; trimming]

Sequence alignment is a critical issue in molecular phylogenetic analyses. Numerous studies have shown that different alignment methods can yield very different topologies for individual genes, and that inaccurate alignments can lead to inaccurate topologies (e.g., Ogden and Rosenberg 2006; Smythe et al. 2006; Talavera and Castresana 2007; Liu et al. 2012; Mirarab et al. 2015; Nguyen et al. 2015).

Yet it is less clear whether alignment methods matter in phylogenomic studies, when hundreds or thousands of genes are included. One can imagine at least two extreme scenarios. First, that different alignment methods lead to systematic errors or biases that are amplified across many loci, and these can substantially impact the resulting phylogenetic estimates. Second, that any possible errors or biases associated with different alignment methods become inconsequential when dozens, hundreds, or thousands of loci are analyzed. An intermediate scenario is that results from different methods are not radically different, but that some alignment methods nevertheless produce higher quality alignments and improved phylogenetic estimates relative to others. Similarly, alignment methods might impact results, but only when data sets have relatively few loci, and not when hundreds or thousands of loci are used. To our knowledge, no previous studies have specifically focused on evaluating the impact of different alignment methods on phylogenomic analyses, and whether some methods might give better results than others. Yet, many workflows for phylogenomic data tend to offer relatively few options for alignment (Freyman 2015; Faircloth 2016; Andermann et al. 2018; Smith and Walker 2019). Overall, it seems urgently important to address how alignment methods may impact phylogenomic analyses.

A related, underexplored issue is that of trimming sequence alignments to remove poorly aligned regions and to reduce the amount of missing data in the alignment (Castresana 2000; Talavera and Castresana 2007; Dress et al. 2008; Capella-Gutiérrez et al. 2009; Wu et al. 2012). Given a set of orthologous sequences, alignment methods generally align highly conserved regions accurately, whereas regions containing many insertions and/or deletions are aligned less reliably (Edgar and Batzoglou 2006; Kemena and Notredame 2009; Thompson et al. 2011; Chatzou et al. 2016). In an effort to reduce noise and improve phylogenetic signal, various trimming methods can be used to identify and remove these unreliable alignment columns prior to analyses. In addition, many alignments are constructed from sequences that have different lengths due to different amounts of data recovered during data collection and processing. This heterogeneity is especially common in data sets from targeted-sequence capture, where heterogeneity can arise from library preparation, capture efficiency, sequencing, and bioinformatics processing (Bi et al. 2012; Faircloth et al. 2012; Lemmon et al. 2012; Portik et al. 2016; Schott et al. 2017; Andermann et al. 2020). Alignments constructed from sequences of different lengths typically yield a "core" portion of sequences, with ends that vary in length (e.g., Fig. 1a). These ends may have considerable missing data, but they may also contain phylogenetically informative sites (at least for

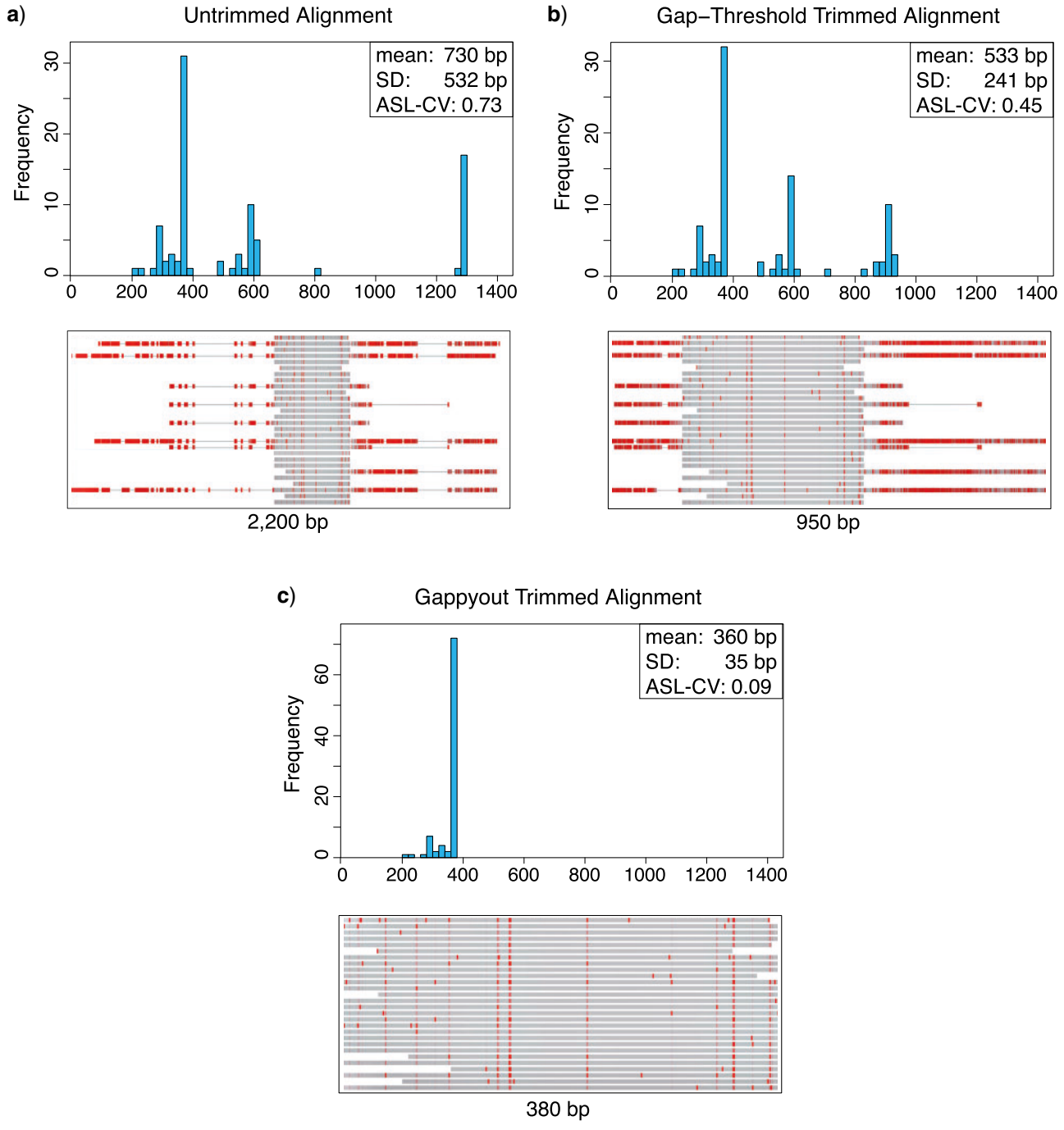

FIGURE 1. An illustration of the sequence-length heterogeneity in one UCE locus (UCE-734) under different trimming strategies for the squamate data set. In a)–c), the top panel shows the frequency distribution of sequence lengths among all 90 taxa for the alignment (from MAFFT-auto). The top panel also includes values for the mean ASL (aligned sequence length), standard deviation (SD), and the ASL coefficient-of-variation (ASL-CV). The bottom panel shows aligned sequences for the first 28 of 90 taxa. Sequences are shown for the a) untrimmed alignment, b) gap-threshold trimmed alignment, and c) gappyout trimmed alignment. Taxa are arranged alphabetically in the alignment visualizations, and therefore the same 28 taxa are shown in the same sequence order across trimming categories. Alignments are colored by frequency-based differences, in which infrequently occurring bases and columns containing a high degree of mismatches are colored red. Visualizations of the sequence alignments were accomplished using the NCBI Multiple Sequence Alignment Viewer v1.10 (https://www.ncbi.nlm.nih.gov/projects/msaviewer/).

some taxa). Trimming methods can be used to remove the ends of such alignments, reducing both sequence-length variation and the overall amount of missing data, but at the cost of decreasing the overall number of informative sites (e.g., Fig. 1b,c). Alignment trimming raises the widespread trade-off involved with missing data: is it better to eliminate these missing data, or to retain portions of the sequences with missing data cells and the phylogenetic information included (for those taxa with nonmissing data)? There have been many empirical and theoretical studies on the pros and cons of including missing data, including studies with phylogenomic data (Hosner et al. 2016; Streicher et al. 2016; Xi et al. 2016; Longo et al. 2017; Molloy and Warnow 2018; Nute et al. 2018). These studies suggest that including such characters can be beneficial, although the benefits may decrease as the amount of missing data increases. In a single-locus context, Tan et al. (2015) explicitly demonstrated that light trimming (removing up to 20% of alignment positions) had minimal impact on gene-tree reconstruction, whereas heavy trimming (removing >40%) tended to remove both phylogenetic noise and signal, leading to inaccurate topologies. However, their case study primarily focused on individual protein-coding genes. The effects on large-scale phylogenomic data sets remain uncertain. Furthermore, in one of the most commonly used data types in phylogenomic studies (ultraconserved elements: UCEs hereafter; Faircloth et al. 2012) the loci are often not protein-coding (especially in vertebrates; e.g., Bejerano et al. 2004; White and Braun 2019), and each locus typically contains a combination of highly conserved and more highly variable regions. Therefore, these data may be far more sensitive to different alignment and trimming methods than protein-coding genes (see below), and so may offer a better system to test the potential impact of these methods. Overall, further study is needed on the impacts of alignment trimming on phylogenomic analyses, particularly for vertebrate UCE data.

In this article, we explicitly test whether different alignment and trimming options impact phylogenomic analyses, specifically for UCE data. To do this, we assemble and analyze large empirical data sets of UCEs for squamate reptiles, birds, and tetrapods. UCEs contain a conserved core region surrounded by variable flanking regions (Faircloth et al. 2012). These flanking regions are particularly useful for phylogenetics (Faircloth et al. 2012). Different alignment methods may align these variable regions more or less accurately, and we investigate if these methods affect downstream phylogenetic analyses. Our primary focus is a squamate UCE data set that is derived from a combination of published genomes and targeted sequence-capture experiments (Leaché and Linkem 2015; Leaché et al. 2016; Linkem et al. 2016; Streicher et al. 2016; Streicher and Wiens 2016, 2017). Sequence-capture experiments normally generate considerable length variation from stochastic processes (Bi et al. 2012; Portik et al. 2016; Schott et al. 2017). In contrast, UCE sequences extracted from published genomes can be both longer and more homogeneous in length, with extended variable regions (Fig. 1). Our secondary focus is on bird and tetrapod UCE data sets derived solely from published genomes. These data sets are expected to contain less sequence-length heterogeneity, relative to the squamate data set. They also represent somewhat different timescales, with birds being the youngest (∼100 million year old [Myr] crown age; Jarvis et al. 2014), tetrapods the oldest (∼350 Myr; Irisarri et al. 2017), with squamates intermediate in age (∼200 Myr; Zheng and Wiens 2016). We recognize that UCE data are not necessarily representative of all types of phylogenomic data. However, they may be particularly sensitive to different alignment and trimming methods, given their conserved core and variable flanking regions. Therefore, if different methods show little impact on UCE data, then this conclusion may apply to other data types that may be more constrained in regards to evolutionary rates (e.g., exons; Hutter et al. 2019), all else being equal.

Our overall methods are as follow. We compare three alignment methods that are widely used in phylogenomic studies: Clustal-O (Sievers et al. 2011), MAFFT (Katoh et al. 2005; Katoh and Standley 2013), and Muscle (Edgar 2004). We also compare three trimming strategies (untrimmed and two methods implemented in TrimAl; Capella-Gutiérrez et al. 2009) and their interactions with different alignment methods. We first create per-locus alignments and concatenated alignments from each combination of methods, and compare the alignment lengths, number of informative sites, and percent missing data. We then compare the topologies and support levels for trees obtained using two commonly used phylogenetic methods for genomic data sets. Specifically, we compare the species-tree (i.e., gene-tree summary) method ASTRAL-III (Mirarab et al. 2014; Mirarab and Warnow 2015; Zhang et al. 2017) and concatenated maximum likelihood analysis with RAxML (Stamatakis 2014). Given that the true tree for each data set is unknown, we evaluate the accuracy of the different alignment and trimming methods using clades that are each supported by the combination of traditional morphology-based taxonomy (and/or morphological synapomorphies) and previous molecular analyses. We consider these clades to be sufficiently well-established to be treated as "known" for method comparison (e.g., Streicher et al. 2016, 2018). This congruence approach for assessing method performance is especially useful for our study because it is currently difficult to simulate realistic UCE data (especially with regards to length variability). Finally, we perform these analyses using the full set of loci from empirical data (∼5000 loci) and with smaller, subsampled data sets comprised of 10% and 1% of the total available loci. These latter analyses allow us to address whether the impacts of different alignment and trimming methods change with the number of loci analyzed.

TABLE 1. Summary statistics for the individual alignments (squamates: 4430 loci; birds: 4992 loci; tetrapods: 5024 loci) produced from each alignment and trimming combination, including alignment length (bp), number of informative sites, percent missing data, and the aligned-sequence lengths coefficient of variation (ASL-CV).

			Alignment length		Informative sites		Missing data		ASL-CV	
Dataset	Trimming category	Alignment method	Average	SD	Average	SD	Average (%)	SD	Average	SD
Squamates	Untrimmed	Clustal-O	1,307	75	677	177	46.7	19.7	0.54	0.31
		MAFFT-auto	1,500	161	549	166	53.3	17.5	0.58	0.34
		MAFFT-FNi	1,564	198	528	159	55.2	16.6	0.57	0.34
		Muscle	1,410	165	600	178	51.1	15.6	0.53	0.32
	Gap-threshold	Clustal-O	1,087	194	593	192	38.6	18.8	0.46	0.29
		MAFFT-auto	1,073	210	470	169	38.9	17.6	0.45	0.28
		MAFFT-FNi	1,073	221	447	161	39.4	16.9	0.45	0.28
		Muscle	1,096	198	528	179	39.3	18.0	0.46	0.29
	Gappyout	Clustal-O	565	367	239	257	8.6	11.3	0.11	0.14
		MAFFT-auto	558	343	189	195	9.0	12.4	0.12	0.15
		MAFFT-FNi	547	335	176	179	8.9	11.9	0.13	0.16
		Muscle	603	357	218	214	11.8	16.3	0.16	0.21
Birds	Untrimmed	Clustal-O	1,368	99	792	186	18.8	5.8	0.05	0.03
		MAFFT-auto	1,615	215	691	183	30.4	9.1	0.05	0.03
		MAFFT-FNi	1,677	242	693	190	32.8	9.5	0.04	0.03
		Muscle	1,673	221	675	183	32.8	8.9	0.04	0.03
	Gap-threshold	Clustal-O	1,193	53	754	173	8.3	4.2	0.04	0.03
		MAFFT-auto	1,158	47	656	165	6.2	4.4	0.04	0.03
		MAFFT-FNi	1,161	49	644	165	6.8	4.6	0.04	0.03
		Muscle	1,158	46	637	165	6.3	4.2	0.03	0.03
	Gappyout	Clustal-O	1,076	122	653	183	4.3	3.5	0.03	0.03
		MAFFT-auto	1,077	91	595	169	3.3	2.8	0.03	0.03
		MAFFT-FNi	1,078	94	583	166	3.7	2.9	0.03	0.03
		Muscle	1,083	85	580	165	3.5	2.9	0.03	0.03
Tetrapods	Untrimmed	Clustal-O	1,409	95	1,000	173	21.3	5.0	0.05	0.03
		MAFFT-auto	1,958	342	1,005	247	41.8	9.9	0.05	0.03
		MAFFT-FNi	2,037	365	1,027	260	44.0	10.1	0.05	0.03
		Muscle	1,913	275	1,011	246	41.0	8.7	0.04	0.02
	Gap-threshold	Clustal-O	1,262	61	957	162	13.1	3.8	0.05	0.03
		MAFFT-auto	1,257	80	888	187	14.3	6.0	0.05	0.03
		MAFFT-FNi	1,277	86	898	194	15.9	6.4	0.05	0.03
		Muscle	1,235	65	867	178	13.0	5.0	0.04	0.02
	Gappyout	Clustal-O	1,144	167	849	216	9.1	5.3	0.04	0.03
		MAFFT-auto	1,063	163	718	199	7.1	5.1	0.04	0.02
		MAFFT-FNi	1,049	186	697	212	7.6	5.6	0.03	0.02
		Muscle	1,072	128	723	185	6.6	4.2	0.03	0.02

MATERIALS AND METHODS

UCE Data

We used 130 published genomes to extract new UCE data for squamates, birds, and tetrapods (Supplementary File S1: Table S1; all Supplementary Files are available on Dryad at http://dx.doi.org/10.5061/dryad.p8cz8w9mh). For birds, we sampled 34 orders, and included two species per order when possible ($n = 19$).

For tetrapods, we sampled broadly across Amphibia (including the three major clades: Anura, Caudata, and Gymnophiona), Crocodylia (all three families), Squamata (14 families; representing all major clades, including Iguania, Serpentes, Gekkota, and Anguimorpha), the single species of Sphenodontia, and Testudines (13 of 14 families). In general, we obtained all available genomes for each of the above groups. For tetrapod groups with more genomes available, we selectively sampled representatives of major clades. Within Mammalia, we sampled Eutheria (including multiple species within Afrotheria, Euarchontoglires, Laurasiatheria, and Xenarthra), Metatheria (Dasyuromorphia and Didelphimorphia), and Prototheria (Monotremata). For the tetrapod data set, we subsampled one species per order for Aves. Complete details regarding our genome

sampling are provided in Supplementary File S1, Text S1 available on Dryad.

We downloaded genomes in fasta format and converted them to 2bit format using the faToTwoBit program of the UCSC Genome Browser (Kent et al. 2002). We used PHYLUCE (Faircloth 2016) to align the tetrapod 5k UCE probe set (5060 loci) to each 2bit genome file using LASTZ (Harris 2007). PHYLUCE was then used to extract all matching UCE sequences and retain up to 500 base pairs of flanking sequence (per side). The bird and tetrapod data sets were created from UCE data obtained exclusively from published genomes.

For squamates, we used published genomes and also UCE data generated from sequence-capture experiments (Supplementary File S1: Table S2 available on Dryad). We obtained sequence-capture UCE data from several published sources. We first downloaded UCE data from Streicher et al. (2016) and Streicher and Wiens (2016, 2017). These studies used the tetrapod 5k UCE probe set (Faircloth et al. 2012) to target up to 5060 UCEs. The data from these three studies contained a total of 95 species and 178,663 sequences. We then searched GenBank and downloaded squamate UCE data from Leaché and Linkem (2015), Leaché et al. (2016), and Linkem et al. (2016). These three studies used a custom probe set to target 541 UCEs from the tetrapod 5k UCE locus set. These data encompassed 127 species (excluding subspecies) and 76,697 sequences. Overall, the sampled species represented 54 families, including most of the ~62 frequently recognized squamate families (e.g., Zheng and Wiens 2016). Missing families were within the well-established clades Gekkota ($n=2$ missing families), Amphisbaenia (4), and Serpentes (2).

Clearly, not all species had data for all loci, especially for the squamate data set. However, this is typical for UCE data sets, even those based on sequencing of whole genomes (e.g., our bird and tetrapod data sets here).

Data Processing

We used SuperCRUNCH (Portik and Wiens 2020) to process all UCE data separately for squamates, birds, and tetrapods. SuperCRUNCH is a bioinformatics toolkit for creating large phylogenetic data sets from GenBank data and/or local (i.e., newly generated) sequence data. The overall workflow involves parsing starting sequences to create locus-specific fasta files, filtering and selecting sequences, performing alignment, and conducting various postalignment tasks, such as relabeling, trimming, concatenation, and format conversion. To properly process local sequence data (i.e., data not downloaded directly from GenBank) SuperCRUNCH requires fasta description lines to contain a unique identifier, taxon name, and locus abbreviation/description (similar to NCBI GenBank format). We relabeled the sequence data obtained from the whole genomes and supplemental data packages to comply with these criteria. We created a general-use script (https://github.com/dportik/phyluce-genomes-to-supercrunch) to process the results of PHYLUCE for sequenced genomes. This script relabels the UCE sequences obtained from genomes to create an input fasta file compatible with SuperCRUNCH. For the bird and tetrapod data sets, we combined all the relabeled UCE sequences obtained from the genomes into a single fasta file, which contained a total of 509,667 sequences and 130 species. A single file was created because there was considerable sampling overlap between these two data sets, and because SuperCRUNCH can easily extract all relevant sequences for a user-defined set of species. For squamates, the relevant sequences obtained from all sources (genome and sequence-capture) were combined into a single fasta file, which contained a total of 338,942 sequences and 236 species. The two fasta files containing the complete UCE sequences are available from an Open Science Framework (OSF) project page created for this study: https://osf.io/qa9r8/.

SuperCRUNCH requires a list of taxa and locus search terms to assemble locus-specific fasta files. We obtained an initial taxon list from each of the two UCE sequence sets (birds/tetrapods, squamates) using SuperCRUNCH (*Fasta_Get_Taxa*). We subsequently pruned the list obtained from the bird/tetrapod sequence set to include 108 ingroup species and 22 outgroup species for birds, and 110 ingroup species for tetrapods. The tetrapod tree was rooted at the branch separating amphibians and amniotes, rather than using outgroups. For squamates, we limited each genus to one representative species, resulting in 119 ingroup and 4 outgroup species. We acknowledge that our phylogenetic results might be improved in some portions of the squamate tree by including multiple species per genus. However, our primary focus was comparing alignment and trimming methods. To search for UCE loci, we used the UCE 5k search terms file included with SuperCRUNCH. For our squamate data set, this file was modified to include the *Sceloporus occidentalis* genome coordinates used to relabel the UCE sequences deposited on GenBank (i.e., Leaché and Linkem 2015; Leaché et al. 2016; Linkem et al. 2016). The identity of these GenBank sequences was revealed through an initial BLAST search to the genome-extracted UCE data, which allowed us to match all 541 of the coordinate labels from *S. occidentalis* to particular UCE loci. The resulting taxon lists and locus search terms files were used to run *Parse_Loci* independently for squamates, birds, and tetrapods. This step generated a data set of 4997 UCE loci for squamates, 5040 loci for birds, and 5040 loci for tetrapods, with each locus containing at least two sequences.

Given that all UCE data were identified using PHYLUCE, we did not perform a sequence-similarity filtering step (which is otherwise standard for SuperCRUNCH analyses). However, in our squamate data set there were four instances in which a species included in our taxon list (*Gambelia wislizenii, Phrynosoma platyrhinos, Plestiodon fasciatus*, and *Uta stansburiana*) had been used in different sequence capture experiments (tetrapod 5k UCE set: Faircloth et al. 2012; 541 UCE set:

Leaché and Linkem 2015). This could result in multiple sequences available for a given species for a given UCE locus. We used *Filter_Seqs_and_Species* to select a single representative sequence per species per locus, taking the longest available sequence if multiple sequences were present. When filtering all sequences with the *Filter_Seqs_and_Species* module, we also enforced a 200-base pair minimum length to retain a sequence (following recommendations of Hosner et al. 2015). We assumed that shorter sequences would contain fewer flanking regions (and consequently fewer variable sites), which could be an additional source of gene-tree error. For the squamate, bird, and tetrapod data sets, we removed all loci containing fewer than 10 taxa. This number is also somewhat arbitrary, but it is important to note that 10 taxa is <10% taxon sampling here (and <4 would be uninformative). This filtering step reduced the final set to 4430 loci for squamates, 4992 loci for birds, and 5024 loci for tetrapods. Finally, for each data set we made the direction of sequences within each locus uniform using MAFFT (Katoh et al. 2002; Katoh and Standley 2013) in the *Adjust_Direction* module.

Multiple Sequence Alignment, Trimming, and Evaluation

We compared alignments using Clustal-O, MAFFT, and MUSCLE because these three methods are commonly used in published phylogenetic/phylogenomic bioinformatics pipelines (e.g., Pearse and Purvis 2013; Freyman 2015; Faircloth 2016; Antonelli et al. 2017; Andermann et al. 2018; Bennett et al. 2018; Smith and Walker 2019). Methods that coestimate alignments and trees can produce more accurate alignments than the three methods used here. These coestimation methods include SATé-II (Liu et al. 2012), PASTA (Mirarab et al. 2015), and UPP (Nguyen et al. 2015). However, these coestimation methods are not as frequently available as options in phylogenomic pipelines (see references above). Given that our main goal was to evaluate the performance of the most commonly used alignment methods for phylogenomics, we did not explore these other methods here.

We used the automatic option for selecting parameters and/or algorithms in both Clustal-O and MAFFT (–auto flag) and the default settings in Muscle (maxiters=16). Given the widespread use of MAFFT, we also chose to align sequences using the FFT-NS-i algorithm in MAFFT. The FFT-NS-i algorithm is an iterative refinement method that is slower than alternative progressive methods but is scalable and capable of producing more accurate alignments under certain conditions (Katoh et al. 2002; Katoh and Standley 2013). To distinguish between the two MAFFT analyses, we refer to these as MAFFT-auto and MAFFT-FNi. Alignments were constructed for all UCE loci with Clustal-O, MAFFT-auto, MAFFT-FNi, and Muscle, using the *Align* module of SuperCRUNCH.

We created three different trimming categories, including no trimming (untrimmed) and two trimming strategies implemented in trimAl (Capella-Gutiérrez et al. 2009). For the first strategy, we used a gap-threshold value of 0.2 to trim alignments, which removed columns containing gaps for more than 80% of the sequences present. This threshold value was set to target and remove poorly aligned regions in the extended ends of the sequences (Fig. 1b). The second strategy used the gappyout method, which calculates a minimum gap-score cut-off based on the input alignment characteristics and trims all alignment columns falling below the threshold value. This automated method adapts parameters for each input alignment, rather than applying the same fixed parameters to all alignments (like the gap-threshold method). The gappyout method was used to trim poorly aligned regions aggressively. For squamates in particular, we expected the gappyout method to trim the extended regions of the genome-based sequences to the same approximate length as the sequence capture sequences (Fig. 1c). We also analyzed untrimmed alignments, for a total of three trimming strategies.

The four alignment methods (Clustal-O, MAFFT-auto, MAFFT-FNi, Muscle) and three trimming categories (untrimmed, gap-threshold, gappyout) yielded 12 distinct alignment and trimming combinations. For each combination, we also constructed a concatenated alignment using the *Concatenation* module of SuperCRUNCH. The complete set of per-locus alignments and concatenated alignments for squamates, birds, and tetrapods are available on OSF: https://osf.io/qa9r8/. We used the *Alignment_Assessment* tool from Portik et al. (2016) to generate summary statistics for all individual and concatenated alignments. These statistics included alignment length, number of informative sites (defined here as sites containing at least two different nucleotides that are each present in at least two sequences, synonymous with parsimony-informative sites), and percent missing data (relative frequency of cells with missing data in the alignment or concatenated matrix). We give alignment lengths in base pairs (bp) but note that these lengths can also include inferred insertions/gap positions. To measure sequence-length heterogeneity within alignments we calculated the coefficient of variation (CV; the ratio of the standard deviation to the mean) from the set of aligned-sequence lengths (ASL), hereafter referred to as ASL-CV. The ASL-CV is a standardized measure for comparing length variation across alignments, with higher ASL-CV values indicating greater variability in sequence lengths. We created a new module in SuperCRUNCH (*Sequence_Length_Heterogeneity*) to calculate ASL-CV and several associated metrics from alignment files. We use these summary statistics as a way to objectively compare alignments characteristics across methods. We emphasize that there are not necessarily "better" or "worse" values with regards to length, informative, sites, missing data, or ASL-CV in this context.

We sought to determine if alignment lengths, number of informative sites, and percent missing data differed significantly across the four alignment methods within a

given trimming category. These data frequently deviated from a normal distribution (based on Shapiro–Wilk tests), and we therefore used nonparametric methods. We used the Kruskal–Wallis rank-sum test to evaluate potential differences across groups. When significant differences were detected between alignment methods, we performed pairwise comparisons using the Wilcoxon rank-sum test using a Bonferroni correction for multiple testing. All statistical tests were conducted in R v3.5.2 (R Core Team 2018).

Phylogenetic Analyses

We performed phylogenetic analyses using two standard approaches for phylogenomic data. First, we analyzed the concatenated alignment from each alignment and trimming combination using unpartitioned maximum likelihood analysis. We used RAxML v8.2 (Stamatakis 2014) to perform a single search for the best-scoring ML tree and conduct 100 rapid bootstrap analyses using the standard GTRCAT model. All concatenated analyses were run on the CIPRES Science Gateway (Miller et al. 2010). We did not partition the UCE data because vertebrate UCE loci are often not protein-coding, and it is therefore unclear what partitions would be appropriate for them, if any (but see Tagliacollo and Lanfear 2018). Furthermore, Roch et al. (2019) recently raised concerns about the potential statistical inconsistency of both fully partitioned and unpartitioned maximum likelihood for phylogenomic analyses, indicating that neither choice is necessarily more appropriate.

For the second phylogenetic approach, we conducted species-tree analyses using the gene-tree summary method, ASTRAL-III (Mirarab et al. 2014; Mirarab and Warnow 2015; Zhang et al. 2017). We first used RAxML v8.2 to construct gene trees for all UCE loci, using the GTRCAT model. This was repeated for each of the 12 alignment and trimming combinations. For each combination, the complete set of gene trees was used to infer a species tree using ASTRAL-III. Important properties of ASTRAL-III are that it: (i) employs a quartet-based approach that is consistent with the multispecies coalescent process, (ii) can resolve gene-tree discordance caused by incomplete lineage sorting, and (iii) allows for missing taxa across gene trees (Mirarab et al. 2014; Mirarab and Warnow 2015). Branch support was assessed using local posterior probabilities (LPP), which are computed from gene-tree quartet frequencies (Sayyari and Mirarab 2016). We acknowledge that other phylogenetic methods could also be used, but our main focus was on alignment methods, and we wished to limit the overall parameter space to explore.

As a proxy for gene-tree error, we estimated how similar gene trees from different alignment methods were to one another. We calculated pairwise normalized Robinson–Foulds (RF) distances between all gene trees estimated for the same locus. Within a trimming category, there were four gene trees per locus (resulting from Clustal-O, MAFFT-auto, MAFFT-FNi, and Muscle), which resulted in six pairwise tree comparisons. We used the Kruskal–Wallis rank-sum test to determine if the RF values of gene tree comparisons differed significantly from one another. We also examined variation in the phylogenies produced by each phylogenetic method. We did so by calculating the average normalized RF distance from all pairwise comparisons of the 12 trees produced by either RAxML or ASTRAL-III. Finally, we examined variation across phylogenetic methods by calculating RF distances between the trees produced by RAxML and ASTRAL-III for each alignment and trimming combination.

Evaluating Method Performance

To compare the accuracy of the trees from each alignment and trimming combination and phylogenetic method, we focused on the ability of each approach to recover and support well-established clades. These clades acted as a proxy for a "true" species tree, which is generally unknown for empirical data. Clades were chosen after taxa were sampled in each data set.

For squamates, we selected clades that are (i) recognized in traditional taxonomies, (ii) supported by morphological synapomorphies (e.g., Estes et al. 1988), and (iii) supported by recent molecular analyses (including concatenated likelihood and species-tree methods; Wiens et al. 2012; Pyron et al. 2013; Streicher et al. 2016; Streicher and Wiens 2016, 2017). Several snake taxa are traditionally recognized and appear in molecular phylogenies but were not used here, because their composition is very different relative to traditional taxonomies, and so their morphological support is therefore unclear (e.g., Boidae, Colubridae). The 35 clades included families, subfamilies, and some higher taxa. For birds and tetrapods, we used 21 and 30 clades (respectively) that are both recognized in traditional morphology-based taxonomy and supported in recent molecular phylogenies (e.g., Jarvis et al. 2014; Prum et al. 2015; Irisarri et al. 2017; Reddy et al. 2017). All major clades of birds and tetrapods were included, but species sampling was limited (as is typical in higher-level, phylogenomic analyses). A list of species and their clade assignments (given our taxon sampling) is provided for each of the three analyses in Supplementary File S1: Tables S3–S5 available on Dryad, along with further justification for the choice of clades in each group (Supplementary File S1 Text S1 available on Dryad).

To rapidly summarize sets of relationships within trees we developed a program called MonoPhylo. MonoPhylo assesses the status (monophyletic, paraphyletic, polyphyletic) of any number of user-defined groupings (genus, subfamily, family, etc.) for the tips present in a given tree. For each grouping, MonoPhylo outputs the number of taxa defining the group, the status of the group, a support value if it is monophyletic (for trees with support measures), and if it is not found to be monophyletic the number of interfering taxa and their corresponding tip labels. MonoPhylo is written in Python and

MANUSCRIPT CATEGORY: Systematic Biology

relies on the ETEv3 toolkit (Huerta-Cepas et al. 2016). It is open-source and freely available with detailed instructions at: https://github.com/dportik/MonoPhylo. We used MonoPhylo to summarize whether clades were recovered as monophyletic and if so, to obtain their corresponding support values. We used a nonparametric unpaired two-sample Wilcoxon test to determine if the mean number of clades recovered differed significantly between the ASTRAL-III and RAxML analyses.

We recognize that these clades are not known to the same degree that clades are known in simulations. Nevertheless, it is difficult to imagine scenarios that would cause both molecular and morphological data to frequently generate concordant yet misleading clades.

Subsampling Loci

One possible outcome of our initial study design was that the size of the full data sets (from 4430 to 5024 loci) would overwhelm any potential differences among alignment and trimming methods. We therefore investigated the effects of subsampling loci. We began this exploration by initially focusing on the squamate data set. We randomly sampled loci (without replacement) from the full set of 4430 loci to produce subsampled data sets containing 10% (400) and 1% (40) of the total available loci. For the species-tree method, we assembled 20 replicates of 400 and 40 randomly selected gene trees. For a given replicate, the same set of 400 or 40 loci was used across all alignment and trimming combinations. For the concatenated analyses, we generated 10 replicates of concatenated alignments that were composed of 400 or 40 randomly selected loci for each alignment and trimming combination. However, the set of loci selected for concatenation replicates were not necessarily identical to those selected for the species-tree analyses. We used a smaller number of replicates for the concatenated analyses given that these were much more computationally intensive, and because initial analyses showed little variation among replicates. These concatenated data sets were analyzed with RAxML as described above, but using 50 rapid bootstrap replicates. Initial results obtained from the squamate data set indicated that effects (i.e., differences from the full data sets) were mainly observed using 1% subsampling. We therefore repeated our subsampling procedure for the bird and tetrapod data sets, but only at the 1% level (which produced sets of 50 loci).

We used MonoPhylo to assess how many of the well-established higher taxa were recovered per analysis, and to obtain their support values. For each alignment and trimming combination, we obtained the average number of these clades recovered based on 20 (ASTRAL-III) or 10 replicates (RAxML). We also calculated an average support value for each clade and across all clades based on the replicates for a given alignment and trimming combination. Subsampling loci led to some variation in terminal taxa across replicates (given that not every species had data for every locus). If a higher taxon was represented by only a single species in a particular replicate, its monophyly and branch support were not testable and we excluded that replicate from the set of support values used to calculate the average support for the clade. However, if a higher taxon contained two or more sampled species in a replicate and it was not monophyletic in a given tree, it was assigned a support value of zero. We did this to penalize valid instances of nonmonophyly, rather than exclude the replicate from estimating the mean support for the clade.

We used the Kruskal–Wallis rank-sum test to determine if the mean number of clades recovered or average support values differed significantly across the four alignment methods within each trimming category. We conducted tests independently for the concatenated and species-tree methods. To compare differences in phylogenetic methods for a given gene sampling strategy (10% or 1%), we used the Kruskal–Wallis rank-sum test to determine if the mean number of clades recovered (based on all replicates from the 12 alignment and trimming categories) differed significantly between the concatenated and species-tree analyses. Finally, we sought to determine if the subsampled data sets (10% or 1%) resulted in lower clade recovery and/or support values relative to the full data sets. We used unpaired two-sample Wilcoxon tests to compare clade recovery, RAxML bootstrap support, and ASTRAL-III LPP of the 10% or 1% subsampled data sets to the full data sets for squamates, birds, and tetrapods.

Results

UCE Data

The final squamate data set contained 123 species (from 54 families), 4430 loci, and 202,570 total sequences. There was considerable variation in the number of loci across species, which was largely attributable to the method used to obtain the sequence data (Supplementary File S1, Table S2 available on Dryad). The squamate data set included 19 species with data from whole genomes (average number of loci = 3824; range among species = 3074–4382), 82 from the tetrapod 5k UCE probe set (1375; 49–2280), 18 from the custom 541 UCE probe set (457; 427–539), and four from both the tetrapod 5k and custom 541 UCE probe sets (2240; 2142–2335). Across the entire squamate data set, each species on average had data for 1647 loci (SD: ±1117 loci), and the average number of taxa per locus was 45 species (±22; range: 10–97). The species included are listed in Supplementary File S1, Table S3 and the number of loci for each species is given in Table S2.

The bird data set contained 66 species, 4992 loci, and 287,868 total sequences. Species had an average of 4428 loci (±770 loci), and the average number of taxa per locus was 57 species (±7; range: 10–65). The species included in this data set are listed in Supplementary File S1, Table S4 (along with the major clades that they belong

MANUSCRIPT CATEGORY: Systematic Biology

to), and the number of loci for each species is given in Supplementary File S1, Table S1.

The tetrapod data set contained 110 species, 5024 loci, and 418,715 total sequences. Species had an average of 3806 loci (±1146 loci), and the average number of taxa per locus was 83 species (±19; range: 10–109). The species included (and their major clades) are given in Supplementary File S1, Table S5, and the number of loci for each species is in Table S1.

Sequence Alignment and Trimming

Untrimmed alignments.—For each of the three groups, the untrimmed sets of per-locus alignments created from each alignment method (Clustal-O, MAFFT-auto, MAFFT-FNi, Muscle) differed significantly in average length, number of informative sites, and percent missing data (Figs. 2–4, Tables 1 and 2, Supplementary File S2: Tables S1–S30 available on Dryad). *Post hoc* pairwise comparisons using the Wilcoxon rank-sum test revealed significant differences between all four methods for each of these alignment characteristics (Supplementary File S2: squamates: Tables S2–S4; birds: Tables S12–S14: tetrapods: Tables S22–S24, with few exceptions (e.g., number of informative sites for some comparisons in birds and tetrapods).

Across all three data sets, the MAFFT-FNi strategy produced the longest average alignments (squamates: 1564 bp, birds: 1677 bp, tetrapods: 2037 bp), followed closely by MAFFT-auto and Muscle (Figs. 2a–4a, Table 1). By contrast, Clustal-O produced considerably shorter average alignments relative to the other methods (squamates: 1307 bp, birds: 1368 bp, tetrapods: 1409 bp). For squamates and birds, the average number of informative sites was highest in Clustal-O alignments, but similar among MAFFT-FNi, MAFFT-auto, and Muscle (Figs. 2b and 3b, Table 1). For tetrapods, all four alignment methods produced a similar average number of informative sites (Fig. 4b). Across all three data sets the mean percentage of missing data was lowest for Clustal-O, with higher and more similar values among MAFFT-auto, MAFFT-FNi, and Muscle (Figs. 2d–4d, Table 1).

The squamate data set contained a mix of data from published genomes and sequence-capture experiments, whereas the bird and tetrapod data sets were derived solely from published genomes. Consequently, there was considerably more sequence-length heterogeneity in the alignments of the squamate data set, relative to the tetrapod and bird data sets (Figs. 2c–4c, Table 1). For squamates, the average ASL-CV value was highest in MAFFT-auto alignments (0.58), followed by MAFFT-FNi (0.57), Clustal-O (0.54), and Muscle (0.53). For birds and tetrapods, the average ASL-CV values were substantially lower and uniform across alignment methods (0.04–0.05; Table 1).

The per-locus effects of different alignment methods were amplified in the resulting concatenated alignments (Table 2). For squamates, birds, and tetrapods, Clustal-O resulted in the shortest alignment (∼5.8 million bp, ∼6.8 million bp, and ∼7.1 million bp, respectively) and MAFFT-FNi resulted in the longest (∼6.9 million bp, ∼8.3 million bp, and ∼10.2 million bp). Differences in the number of informative sites were also large (Table 3). Clustal-O resulted in the highest number of informative sites in squamates (∼3.0 million sites) and birds (∼3.9 million sites), whereas MAFFT-FNi produced the greatest number of informative sites for tetrapods (∼5.1 million sites). Missing data were similar across the concatenated alignments in squamates, ranging from 83.1% to 85.9% (Table 2). For birds and tetrapods, the concatenated Clustal-O alignments had the least missing data (28.2% and 40.5%, respectively), and quantities were higher but similar among the other three alignment methods (39.2%–41.4% and 56.2%–58.8%).

Effects of trimming.—As expected, gap-threshold trimming removed fewer alignment columns than gappyout trimming (Figs. 2–4, Tables 1–3). Gap-threshold trimming targeted poorly aligned regions in the extended ends of the genome sequences, but still left considerable data in these extended ends (e.g., Fig. 1b). In contrast, the gappyout trimming was far more aggressive in removing alignment columns and tended to trim alignments to a core alignment block, thereby reducing missing data (e.g., Fig. 1c). The average number of bases trimmed using the gap-threshold method differed between squamates (range across alignment methods: 220–490 bp), birds (174–515 bp), and tetrapods (147–759 bp). Differences between data sets were also apparent using the gappyout method, which removed a greater average number of bases relative to gap-threshold trimming for squamates (range across alignment methods: 742–1017 bp), birds (292–599 bp), and tetrapods (265–988 bp). Across all three data sets, the gap-threshold and gappyout trimming removed the fewest bases from Clustal-O alignments and the greatest from MAFFT-FNi alignments (Table 3).

In general, gap-threshold and gappyout trimming produced similar effects across the three data sets: shortening alignment lengths, reducing the amount of missing data, and decreasing the number of informative sites (Figs. 2–4). However, the magnitude of these changes differed across squamates, birds, and tetrapods (Tables 1 and 2). For birds and tetrapods, trimming tended to reduce the initial differences in alignment lengths, missing data, and informative sites across the alignment methods (Tables 1 and 2). However, after trimming using either strategy, we still frequently observed significant differences in these variables across alignment methods (Supplementary File S2: squamates: Tables S5–S10; birds: Tables S15–S20; tetrapods: Tables S25–S30 available on Dryad). Thus, trimming failed to mitigate the initial relative differences produced by the different alignment methods. For squamates, trimming produced similar average alignment lengths and missing data values across alignment methods, which led to fewer significant differences in pairwise comparisons of these metrics as compared to untrimmed alignments (Supplementary File S2: untrimmed: Tables S2 and S4; trimmed: Tables S5,

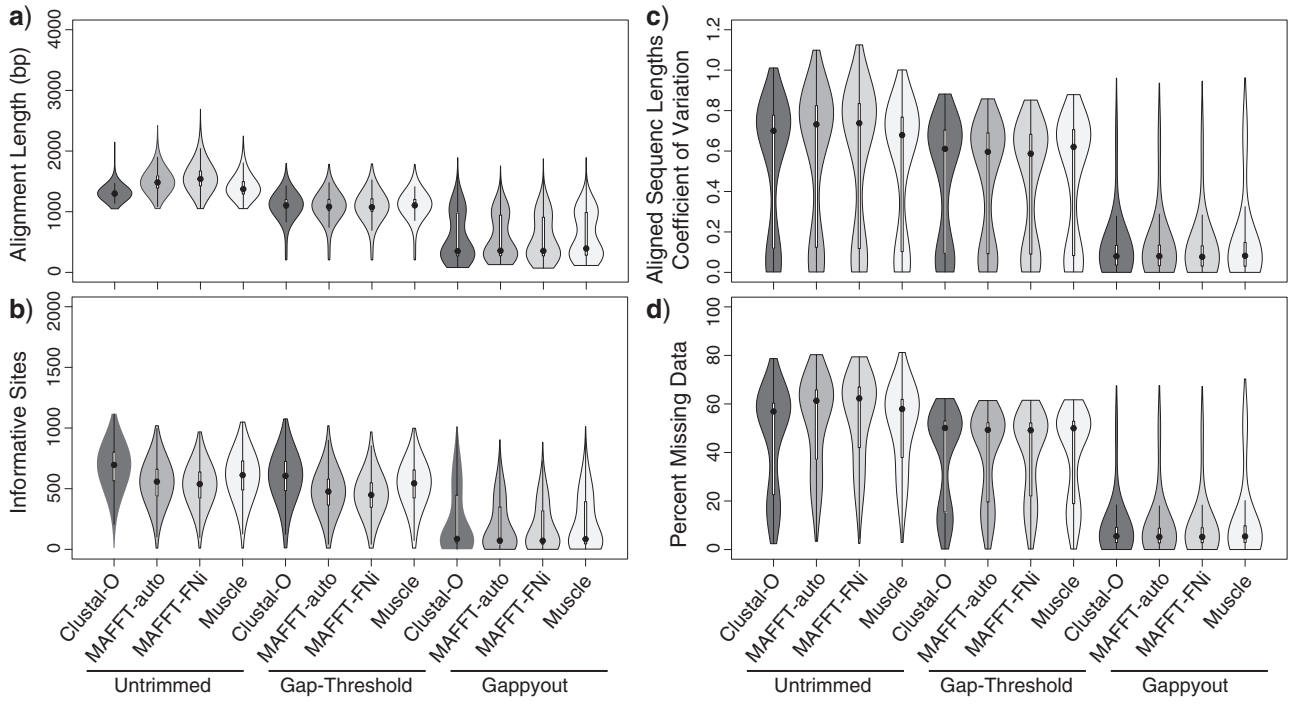

FIGURE 2. Violin plots for squamate alignments showing (a) alignment lengths, (b) number of informative sites, (c) aligned-sequence lengths coefficient-of-variation, and (d) percent missing data, for each of the 12 alignment and trimming combinations for the full data set. The width of each plot is equivalent to the frequency of different values among the 4430 alignments. The interiors of plots contain black dots representing median values, white bars representing interquartile values, and black lines representing the minimum and maximum values.

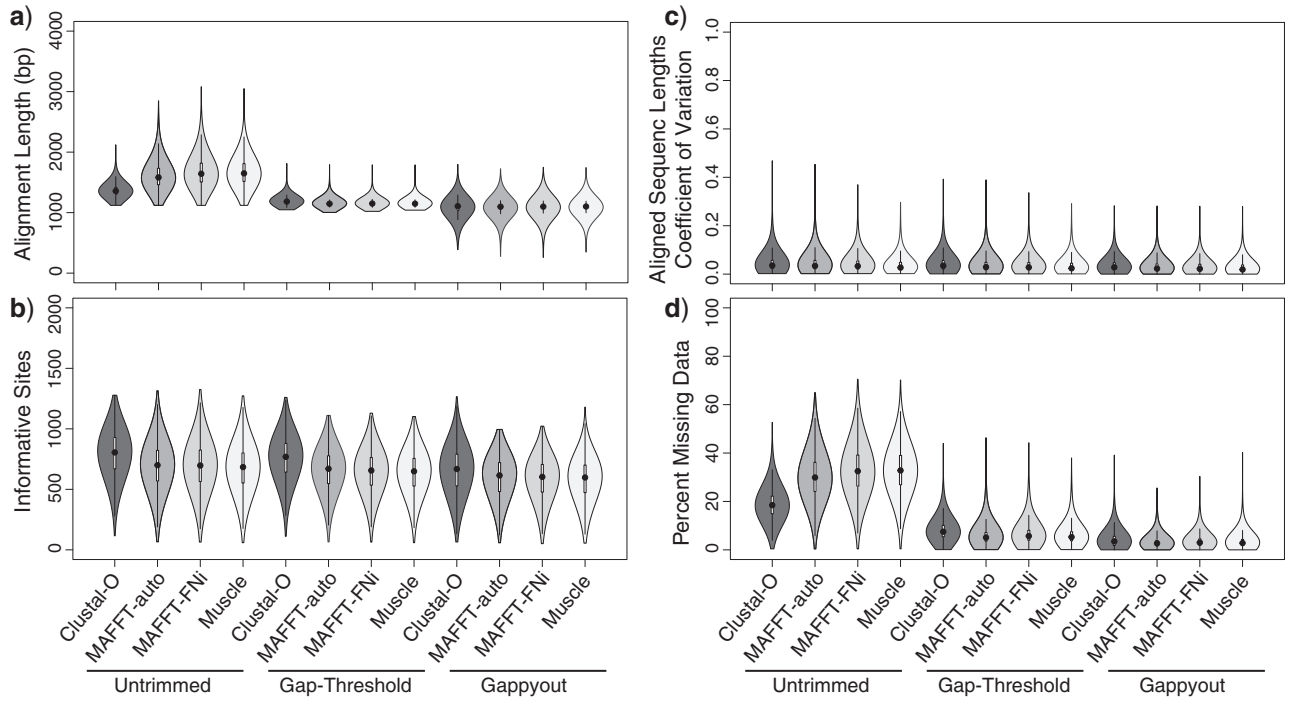

FIGURE 3. Violin plots for bird alignments showing a) alignment lengths, b) number of informative sites, c) aligned-sequence lengths coefficient-of-variation, and d) percent missing data, for each of the 12 alignment and trimming combinations for the full data set. The width of each plot is equivalent to the frequency of different values among the 4992 alignments. The interiors of plots contain black dots representing median values, white bars representing interquartile values, and black lines representing the minimum and maximum values.

TABLE 2. Summary statistics for the concatenated alignments produced from each of the 12 alignment and trimming combinations for squamates, birds, and tetrapods, including alignment length (bp), number of informative sites, and percent missing data.

Dataset	Trimming category	Alignment method	Alignment length	Informative sites	Missing data (%)
Squamates	Untrimmed	Clustal-O	5,789,745	2,997,360	83.1
		MAFFT-auto	6,645,866	2,430,395	85.3
		MAFFT-FNi	6,927,188	2,339,511	85.9
		Muscle	6,247,932	2,660,035	84.4
	Gap-threshold	Clustal-O	4,815,309	2,626,910	80.7
		MAFFT-auto	4,755,441	2,082,436	80.8
		MAFFT-FNi	4,753,604	1,979,769	80.9
		Muscle	4,853,985	2,339,403	80.8
	Gappyout	Clustal-O	2,500,803	1,059,972	75.4
		MAFFT-auto	2,471,075	839,346	74.9
		MAFFT-FNi	2,421,811	780,385	74.6
		Muscle	2,672,719	965,055	75.7
Birds	Untrimmed	Clustal-O	6,829,371	3,954,646	28.2
		MAFFT-auto	8,064,147	3,451,312	39.2
		MAFFT-FNi	8,375,194	3,457,428	41.4
		Muscle	8,351,737	3,369,760	41.3
	Gap-threshold	Clustal-O	5,957,920	3,765,617	18.6
		MAFFT-auto	5,780,854	3,276,350	16.7
		MAFFT-FNi	5,799,768	3,216,155	17.3
		Muscle	5,781,822	3,180,628	16.8
	Gappyout	Clustal-O	5,371,344	3,259,976	15.0
		MAFFT-auto	5,379,261	2,971,642	14.0
		MAFFT-FNi	5,384,533	2,910,950	14.3
		Muscle	5,406,916	2,895,705	14.2
Tetrapods	Untrimmed	Clustal-O	7,082,287	5,027,442	40.5
		MAFFT-auto	9,836,476	5,049,857	57.2
		MAFFT-FNi	10,233,946	5,160,357	58.8
		Muscle	9,612,816	5,083,168	56.2
	Gap-threshold	Clustal-O	6,341,770	4,809,782	34.2
		MAFFT-auto	6,317,019	4,464,483	35.3
		MAFFT-FNi	6,419,271	4,509,521	36.5
		Muscle	6,205,445	4,359,788	34.2
	Gappyout	Clustal-O	5,749,631	4,266,865	31.4
		MAFFT-auto	5,341,073	3,608,384	30.0
		MAFFT-FNi	5,269,194	3,503,478	30.5
		Muscle	5,389,100	3,634,475	29.4

S7, S8, S10). However, despite the similar alignment lengths, the number of informative sites remained significantly different between alignment methods after gap-threshold trimming, and between nearly all alignment methods after gappyout trimming (Tables 1 and 2; Supplementary File S2: untrimmed: Table S3; trimmed: Tables S6 and S9). In squamates, neither gap-threshold nor gappyout trimming removed the initial differences in the number of informative sites across alignment methods, but trimming reduced the initial differences in alignment lengths and missing data.

The effect of trimming on ASL-CV differed greatly between the squamate data set and the bird and tetrapod data sets (Figs. 2c–4c, Table 1). Both birds and tetrapods displayed low ASL-CV values across alignment methods in the untrimmed category (range: 0.4–0.5), indicating the starting sequences were generally uniform in length. Although both trimming methods removed up to several hundred base pairs per alignment, the impact on ASL-CV values was minimal (post-trimming range: 0.3–0.4; Table 1). Squamates displayed much higher ASL-CV values across alignment methods in the untrimmed category (range: 0.53–0.58). Values were reduced somewhat with gap-threshold trimming (0.45–0.46), and even more with gappyout trimming (0.11–0.16). Therefore, trimming was effective in reducing sequence length heterogeneity when it was present.

The per-locus patterns from trimming were amplified in the concatenated alignments (Table 2). These differences are illustrated with MAFFT-FNi and Clustal-O (which represent the extremes). For example, in squamates, gap-threshold trimming removed ~970,000

MANUSCRIPT CATEGORY: Systematic Biology

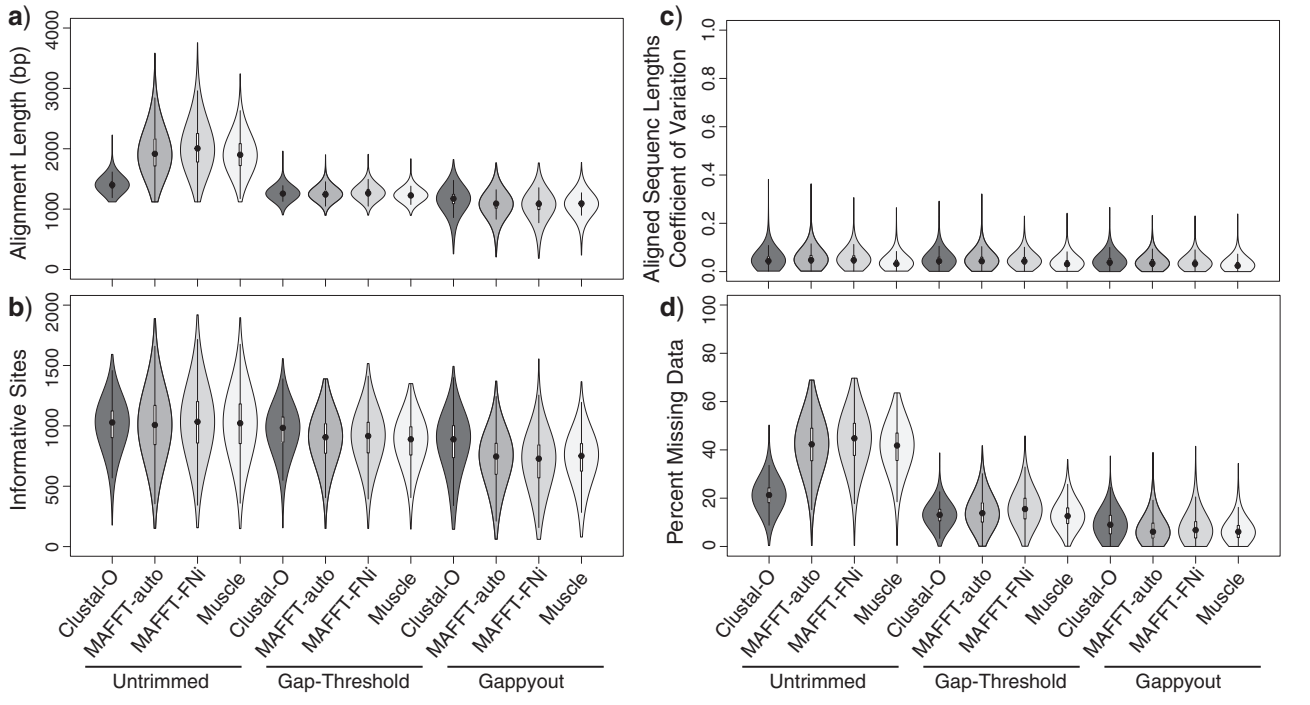

FIGURE 4. Violin plots for tetrapod alignments showing a) alignment lengths, b) number of informative sites, c) aligned-sequence lengths coefficient-of-variation, and d) percent missing data, for each of the 12 alignment and trimming combinations for the full data set. The width of each plot is equivalent to the frequency of different values among the 5024 alignments. The interiors of plots contain black dots representing median values, white bars representing interquartile values, and black lines representing the minimum and maximum values.

bp from the Clustal-O alignments (17% of the total bp) and >2,000,000 bp from MAFFT-FNi alignments (31%). Nevertheless, trimming resulted in similar concatenated alignment lengths for these methods (~4.8 and ~4.7 million bp; Table 2). The reduction in concatenated alignment lengths was even more dramatic after gappyout trimming. In squamates, the gappyout trimming removed ~3.2 million bp from Clustal-O alignments (55% of alignment columns) and ~4.5 million bp from the MAFFT-FNi alignments (65%), resulting in concatenated alignment lengths of ~2.5 and ~2.6 million bp (Table 2). Similar patterns are present in birds and tetrapods (Table 2). Yet, the relative differences in the number of informative sites persisted (Clustal-O > Muscle > MAFFT-auto > MAFFT-FNi; Table 3), mirroring the per-locus results.

Phylogenetic Analyses and Clade Support

Gene-tree comparisons.—We performed pairwise comparisons of gene trees from different alignment methods to measure overall gene-tree similarity (Supplementary File S3: Tables S1–S16 available on Dryad). Across all three trimming categories (untrimmed, gap-threshold, gappyout), gene-tree comparisons involving Clustal-O consistently resulted in significantly higher average normalized RF distances (squamates: 0.52–0.53 [Tables S1–S5]; birds: 0.57–0.58 [Tables S6–S10]; tetrapods: 0.50–0.54 [Tables S11–S15]) than those for other methods (squamates: 0.35–0.44; birds: 0.41–0.45; tetrapods: 0.29–0.43). Overall, trimming did not change the average gene-tree distance relationships between alignment methods, and gene-trees resulting from Clustal-O were consistently the most dissimilar.

Phylogenetic results.—Representative phylogenies for the squamate data set are shown in Figure 5 (RAxML) and Figure 6 (ASTRAL-III). Phylogenies for the bird and tetrapod data sets are provided in Supplementary Figures S1–S4 available on Dryad. The squamate phylogenies are based on MAFFT-FNi without trimming, whereas the bird and tetrapod phylogenies are based on MAFFT-auto alignments without trimming. These particular alignment and trimming methods performed as well as several other combinations, and the performance of all 12 combinations was similar for each data set (Supplementary File S5, Tables S1–S3 available on Dryad). For comparison, the complete set of 30 rooted trees for squamates, birds, and tetrapods is available on OSF: https://osf.io/qa9r8/.

Squamate phylogenies from the 12 concatenated analyses recovered an average of 34.3 of 35 well-established clades (proportion = 0.98), whereas phylogenies from ASTRAL-III recovered an average of 32.8 clades (0.94; Supplementary File S5; Table S1). For recovered clades, support values were consistently high. The average bootstrap score from the concatenated analyses was 99.9% (SD: ± 0.4; from Supplementary File S4; Tables S1–S3), and the average LPP from species-tree analyses was 0.999 (±0.002; from Supplementary File S4; Tables S4–S6). Amphisbaenia was not recovered as monophyletic

TABLE 3. Summary of the average number of base pairs and informative sites that were trimmed from the alignments of the squamate, bird, and tetrapod datasets for each alignment method, using the gap-threshold and gappyout trimming strategies.

			Base pairs trimmed		Informative sites removed	
Dataset	Trimming category	Alignment method	Average	SD	Average	SD
Squamates	Gap-threshold	Clustal-O	220	189	83	98
		MAFFT-auto	427	236	78	91
		MAFFT-FNi	490	257	81	90
		Muscle	314	190	72	97
	Gappyout	Clustal-O	742	374	437	283
		MAFFT-auto	942	385	359	230
		MAFFT-FNi	1,017	388	351	215
		Muscle	807	350	382	272
Birds	Gap-threshold	Clustal-O	174	71	38	23
		MAFFT-auto	457	193	35	28
		MAFFT-FNi	515	218	48	37
		Muscle	514	202	38	29
	Gappyout	Clustal-O	292	158	139	120
		MAFFT-auto	537	240	96	93
		MAFFT-FNi	599	266	109	99
		Muscle	589	236	95	87
Tetrapods	Gap-threshold	Clustal-O	147	60	43	22
		MAFFT-auto	700	287	116	81
		MAFFT-FNi	759	307	129	85
		Muscle	678	239	144	88
	Gappyout	Clustal-O	265	181	151	153
		MAFFT-auto	895	395	287	220
		MAFFT-FNi	988	435	329	246
		Muscle	840	302	288	178

in 8/12 concatenated analyses (Supplementary File S4; Tables S1–S3 available on Dryad). Amphisbaenia and Colubroidea were not recovered as monophyletic in any ASTRAL-III analyses, and Leiosauridae was not recovered in two ASTRAL-III analyses (Supplementary File S4; Tables S4–S6).

For birds, all 12 concatenated analyses recovered 19 of 21 well-established clades (proportion = 0.90), whereas phylogenies from ASTRAL-III recovered an average of 18.7 (0.89; Supplementary File S5: Table S1). Well-established clades that were recovered received 100% bootstrap support or an LPP of 1.0 (Supplementary File S4: Tables S19–S24). Two taxa were not supported in any analyses (Coraciiformes, Gruiformes), and Pelecaniformes was not recovered as monophyletic in 3 of 12 ASTRAL-III analyses (Supplementary File S4: Tables S19–S24).

For tetrapods, concatenated analyses recovered an average of 28.1 of 30 well-established clades (proportion = 0.94; from Supplementary File S4: Tables S31–S33), whereas ASTRAL-III recovered 27.9 (0.93; Supplementary File S4: Tables S34–S36). Recovered clades received 100% bootstrap support or an LPP of 1.0 (except Batrachia with ASTRAL-III; Supplementary File S4: Tables S31–S36). Archosauria was not recovered as monophyletic in any analyses. Lepidosauria was only monophyletic in four concatenated and four ASTRAL analyses, and Batrachia was monophyletic in nine concatenated and seven ASTRAL analyses (Supplementary File S4: Tables S31–S36).

For all three data sets, we found no significant differences in LPP or bootstrap support for well-established clades between alignment methods within each trimming category or across categories (Supplementary File S5: Table S3). However, the average proportion of clades recovered across all 12 alignment and trimming analyses was higher for concatenated analyses than species-tree analyses (see above), and this difference was significant for squamates ($P < 0.001$), but not birds ($P = 0.07$) or tetrapods ($P = 0.35$), which had fewer clades (Supplementary File S5: Table S4).

Beyond the well-established clades, there was variation in the overall topology within and between phylogenetic methods (Supplementary File S3: Table S16). Pairwise comparisons among the 12 trees from the RAxML analyses yielded average RF distances ($\pm$ SD) of 0.10 $\pm$ 0.05, 0.15 $\pm$ 0.12, and 0.13 $\pm$ 0.08 for squamates, birds, and tetrapods. ASTRAL-III analyses

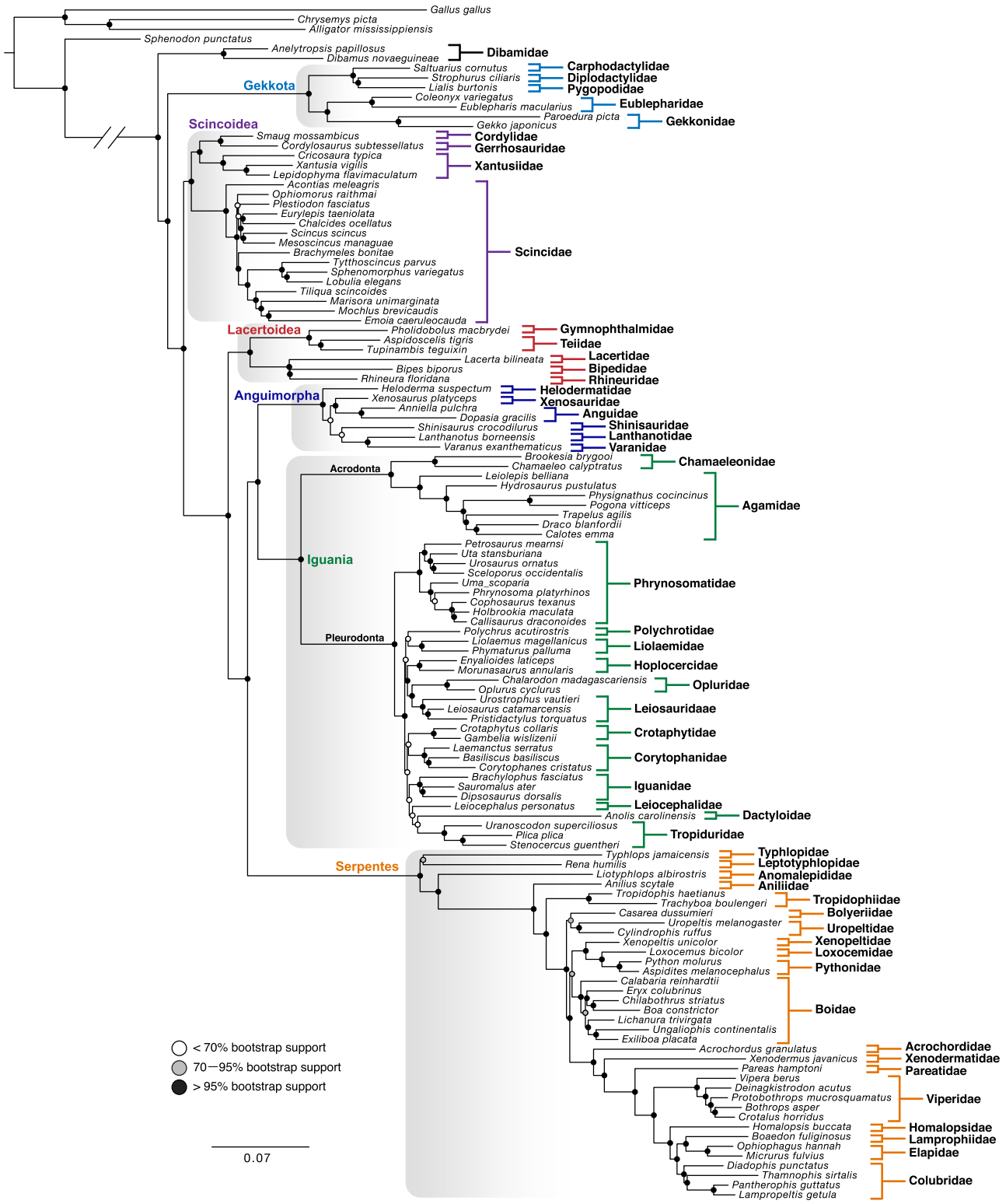

FIGURE 5. Phylogenetic estimate for squamate reptiles based on the concatenated maximum likelihood analysis of 4430 UCE loci using RAxML. The data set is based on the concatenated untrimmed MAFFT-FNi alignments (6,927,188 base pairs, 85.9% total missing data). Scale bar represents substitutions per site.

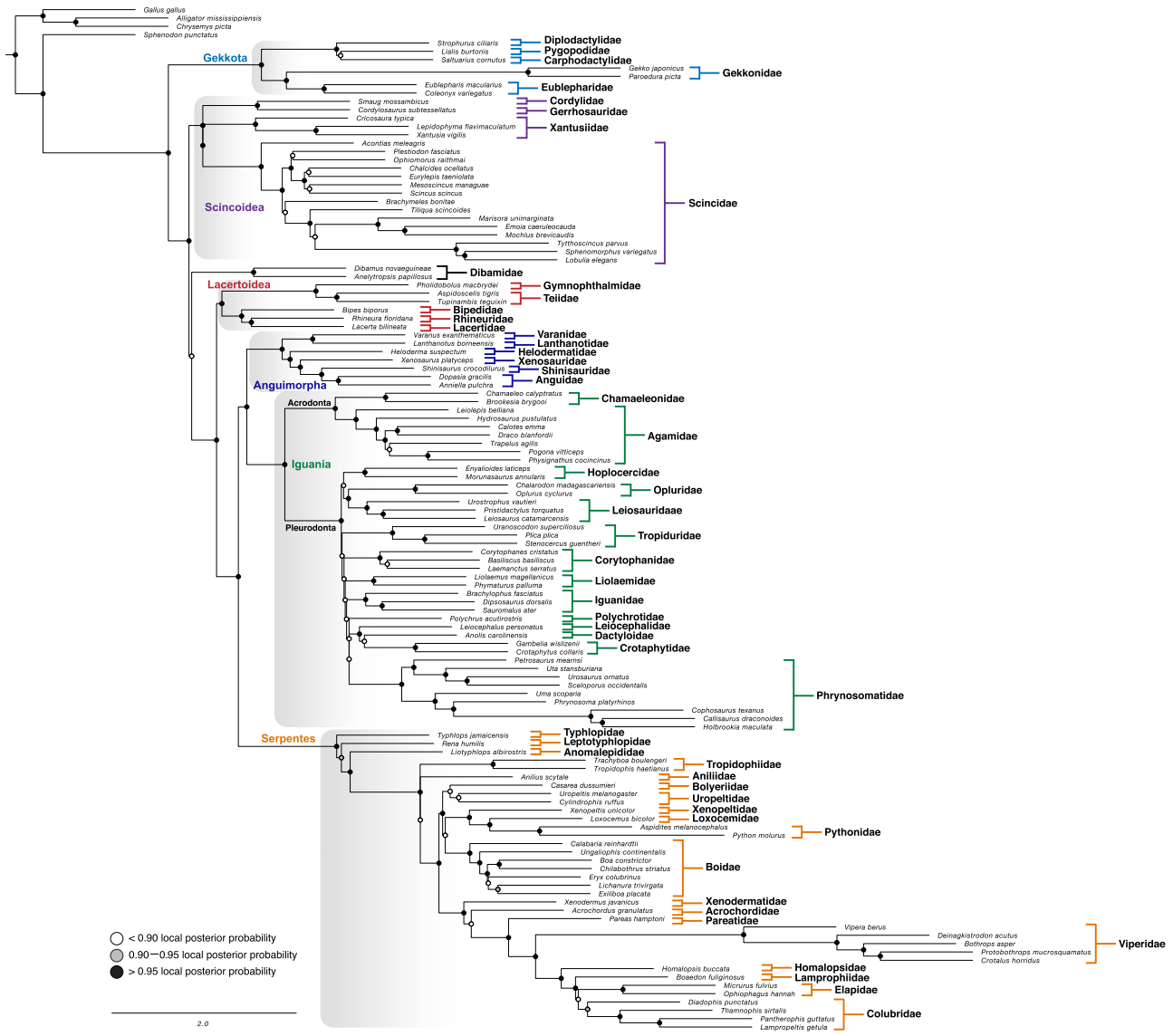

FIGURE 6. Phylogenetic estimate for squamate reptiles based on the species-tree analysis of 4430 UCE loci using ASTRAL-III. The analysis is based on gene trees generated from untrimmed MAFFT-FNi alignments. Scale bar represents coalescent units.

yielded distances of 0.14 ± 0.05, 0.23 ± 0.16, and 0.13 ± 0.09. A direct comparison of trees from each method for a given alignment and trimming combination produced greater average RF distances (squamates: 0.25 ± 0.02; birds: 0.20 ± 0.08; tetrapods: 0.14 ± 0.05). Thus, phylogenetic methods had a stronger impact on overall topology than alignment and trimming methods.

Subsampling Loci

We performed analyses with reduced sampling to determine if differences in topology among methods were masked by sampling many loci (Supplementary Files S4 and S5). We sampled 10% and 1% for squamates (400 and 40 loci), but only 1% for birds and tetrapods (50 loci). We present complete results for squamates, and brief summaries for birds and tetrapods.

For squamates, 10% subsampling (400-locus) resulted in a ∼5% reduction in mean bootstrap values and a decrease of ∼0.11 in mean LPP, relative to the full data set (Supplementary File S5: Tables S29 and S30). However, there were no significant differences in mean support values between alignment methods within a trimming category (Fig. 7a,c; Supplementary File S5: Table S9). For some trimming categories, we found significant differences in the proportion of well-established clades recovered across alignment methods (Fig. 7b,d; Supplementary File S5: Table S7). Clustal-O alignments resulted in significantly fewer clades recovered than MAFFT-auto (and sometimes MAFFT-FNi and Muscle), whereas other methods were not significantly different (Supplementary File S5: Tables S5 and S7). This effect

occurred in the untrimmed and gap-threshold trimmed alignments, but was somewhat reduced by gappyout trimming, which caused the other methods to recover fewer clades (Supplementary File S5: Table S5). Again, the concatenated analyses recovered a significantly higher proportion of well-established clades (average = 0.95; range = 0.91–0.97) than the species-tree method (average = 0.86; range = 0.83–0.89) across all alignment and trimming combinations ($P < 0.001$; Supplementary Files S5: Table S11).

For squamates, the 1% subsampling (40-locus) analyses produced even lower support values (Supplementary File S5: Table S13). We found significant differences in the support values from RAxML and ASTRAL-III analyses for untrimmed alignments (but not gap-threshold or gappyout trimmed alignments; Supplementary File S5; Table S14). Specifically, analyses using MAFFT alignments (FNi and auto) produced significantly higher support values than analyses of Clustal-O alignments for both phylogenetic methods (Fig. 8a,c; Supplementary File S5; Table S15 available on Dryad). We did not find significant differences in the proportion of well-established clades recovered across alignment methods for any trimming categories (Fig. 8b,d), with one exception (Supplementary File S5: Table S23). For ASTRAL-III analyses of the untrimmed alignments, Clustal-O recovered significantly fewer clades than MAFFT-auto (Supplementary File S5: Table S24). Concatenated analyses recovered a significantly higher proportion of clades (average = 0.69; range = 0.58–0.75) than species-tree analyses (average = 0.36; range = 0.34–0.40) across alignment and trimming combinations ($P < 0.001$; Supplementary File S5: Table S25).

For birds, 1% subsampling revealed significant differences in clade recovery using both phylogenetic methods for all alignment and trimming combinations (Supplementary File S5: Tables S12 and S14). Analyses with Clustal-O alignments recovered significantly fewer clades than MAFFT-FNi (and frequently MAFFT-auto and Muscle; Supplementary File S5: Tables S16–S20). However, there were no significant differences in support values (Supplementary File S5: Tables S13 and S23).

For tetrapods, 1% subsampling revealed significantly fewer clades recovered from ASTRAL-III analyses with Clustal-O versus those with MAFFT-auto and MAFFT-FNi (Supplementary File S5: Tables S14, S21, S22). We did not find significant differences in support values for clades from either phylogenetic method (Supplementary File S5: Tables S13 and S23).

As in squamates, concatenated analyses in birds and tetrapods recovered a significantly higher proportion of clades than the species-tree analyses across alignment and trimming combinations (birds: averages: 0.89 vs. 0.82; $P < 0.001$; tetrapods: 0.94 vs. 0.89; $P < 0.001$; Supplementary File S5: Table S26).

Comparisons of the full data sets to the subsampled data sets (10% and 1%) revealed significant differences in clade recovery and support values (Supplementary File S5: Tables S26–S32). For squamates, the 10% and 1% data sets recovered significantly fewer clades from both phylogenetic methods (full: 0.95; 400 loci: 0.70; 40 loci: 0.53; Supplementary File S5: Tables S27 and S28) and significantly lower mean support values from RAxML (full: 98.1%; 400 loci: 93.7%; 40 loci: 59.5%; Supplementary File S5: Tables S29 and S31) and ASTRAL-III (full: 0.94; 400 loci: 0.83; 40 loci: 0.31; Supplementary File S5: Tables S30 and S32). For birds, the 1% data sets recovered a significantly lower proportion of clades than the full data set, but there were no differences in average support values (Supplementary File S5: Tables S28 and S31). For tetrapods, there were no significant differences in the proportion of clades recovered in the full versus 1% data sets, but the 1% data sets had significantly lower average support values (for both ASTRAL-III and RAxML; Supplementary File S5: Tables S28 and S31).

Discussion

In this study, we address whether different alignment and trimming methods impact phylogenomic analyses. We found significant differences in the data sets generated by different alignment and trimming methods, including differences in length and the number of informative sites. However, our results suggest that different alignment and trimming methods need not strongly impact phylogenomic results, in terms of topologies and clade support. Nevertheless, we do provide some observations that should be relevant to method choice. Specifically, with fewer genes sampled (10% and 1% of ~5000 loci), we found that MAFFT and Muscle performed better than Clustal-O and that aggressive trimming (gappyout) sometimes performed significantly worse than other methods (in terms of recovering and strongly supporting well-established clades). Intriguingly, we found much stronger impacts of phylogenetic methods, with concatenated RAxML analyses performing better than the species-tree method used here (ASTRAL-III) when fewer genes were sampled. Below, we emphasize several caveats about our conclusions. We then address the implications of our results for alignment method choice, sequence-length heterogeneity, data set size, phylogenetic methods, and squamate phylogeny.

Potential Caveats

The most important caveat about our conclusions is that they are based only on vertebrate UCE data. Therefore, it is crucial to consider whether our results will apply to other data sets or not. First, alignment effects might be stronger at deeper phylogenetic scales, with sequences that are more divergent. The oldest group considered here was ~350 Myr old (tetrapods). However, we note that across the Tree of Life, many more extant clades are younger rather than older. We did not include species-level data sets because we would expect

MANUSCRIPT CATEGORY: Systematic Biology

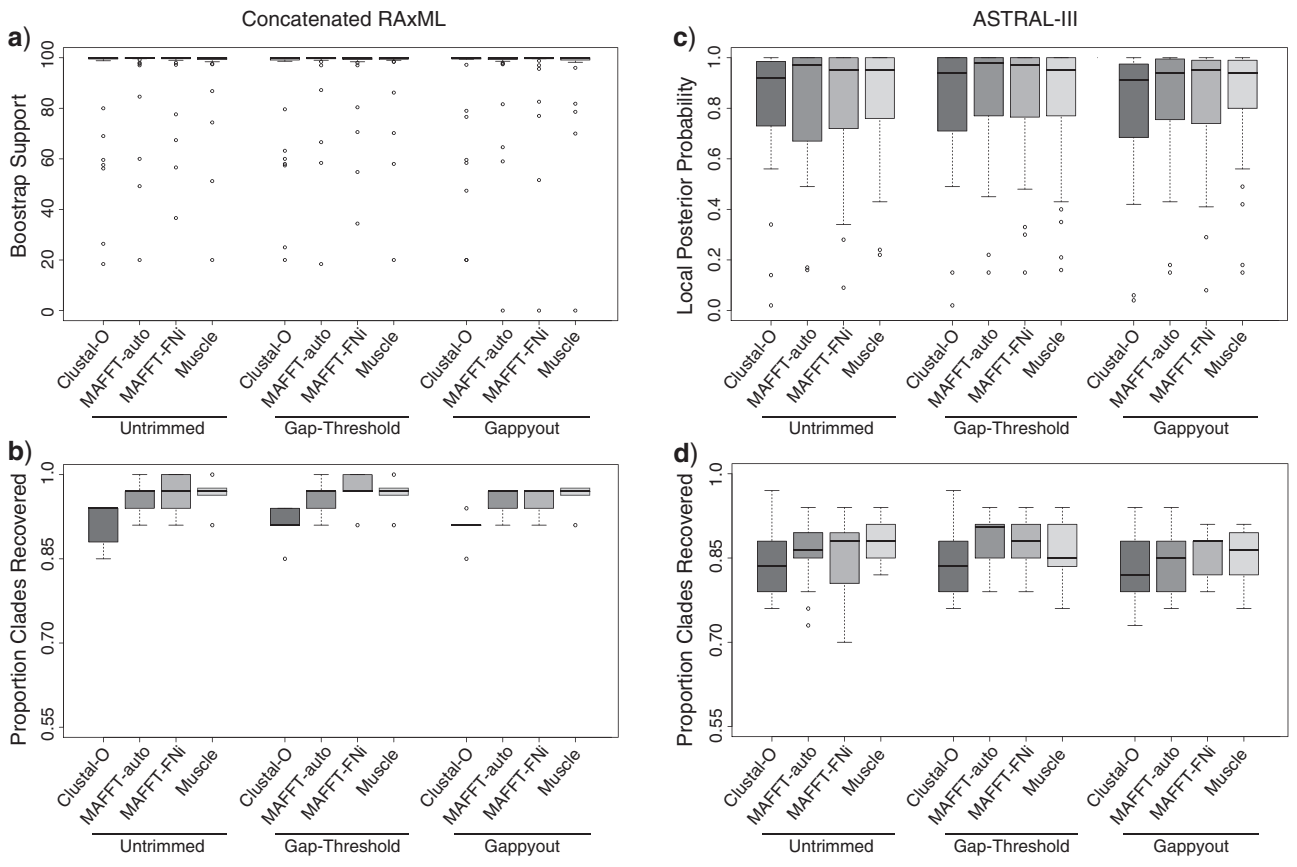

FIGURE 7. Mean support values and proportion of clades recovered for the squamate 10% (400-locus) data sets for the 35 well-established clades. Results are based on concatenated maximum likelihood (RAxML) and a species-tree method (ASTRAL-III). For RAxML, a) depicts the mean bootstrap support across the 35 well-established clades for each replicate (n= 10 replicates) and b) depicts the proportion of the 35 clades recovered for each replicate. For ASTRAL-III, c) depicts the mean LPP support across the 35 well-established clades for each replicate (n= 20 replicates) and d) depicts the proportion of the 35 clades recovered for each replicate. Boxplots show the median value (black line), interquartile range (box), values up to 1.5 times the interquartile range (whiskers), and outliers (dots).

to see even smaller impacts of different alignment and trimming methods on topologies at this shallow scale.

Second, our data set consists of UCE data, and other results are possible for other kinds of molecular data. For example, studies have demonstrated clear differences in the performance of alignment methods for RNA sequences (Liu et al. 2012; Mirarab et al. 2015; Nguyen et al. 2015). Yet, phylogenomic data sets include many genes (by definition). Therefore, even if ribosomal genes were included (and were more strongly influenced by different alignment methods), their overall impacts should be mitigated by other genes. We observed substantial differences in alignments produced from the same UCE sequences, indicating that the hypervariable flanking sequences of UCE loci may be challenging for alignment methods. Hutter et al. (2019) found that prior to trimming, UCE alignments display qualities more similar to those from introns than exons, but that this also depended on phylogenetic scale (see also Chan et al. 2020). Based on those findings, our results may be informative for introns, particularly at deeper phylogenetic scales. In contrast, other types of phylogenomic data (such as exons) may show fewer effects of different alignment and trimming methods than these UCE data. We also note that UCE data sets in other groups (like arthropods) may be dominated by exons (e.g., Bossert and Danforth 2018; Hedin et al. 2019), and so may also show limited impacts of alignment and trimming methods.

Third, there were considerable missing data in these UCE data sets (up to 86% overall; Table 2). However, it is not clear how this would bias or affect our inferences about the impact of different alignment methods.

Fourth, there might also be other factors that we have not considered that might cause other data sets to yield different results. Importantly, the analyses that we did here can be easily repeated in other clades and with other types of sequence data (e.g., using the same options in SuperCRUNCH to streamline data processing, alignment, and trimming).

Another important caveat is that our results may only apply to the particular methods that we looked at, and

Copyedited by: YS MANUSCRIPT CATEGORY: Systematic Biology

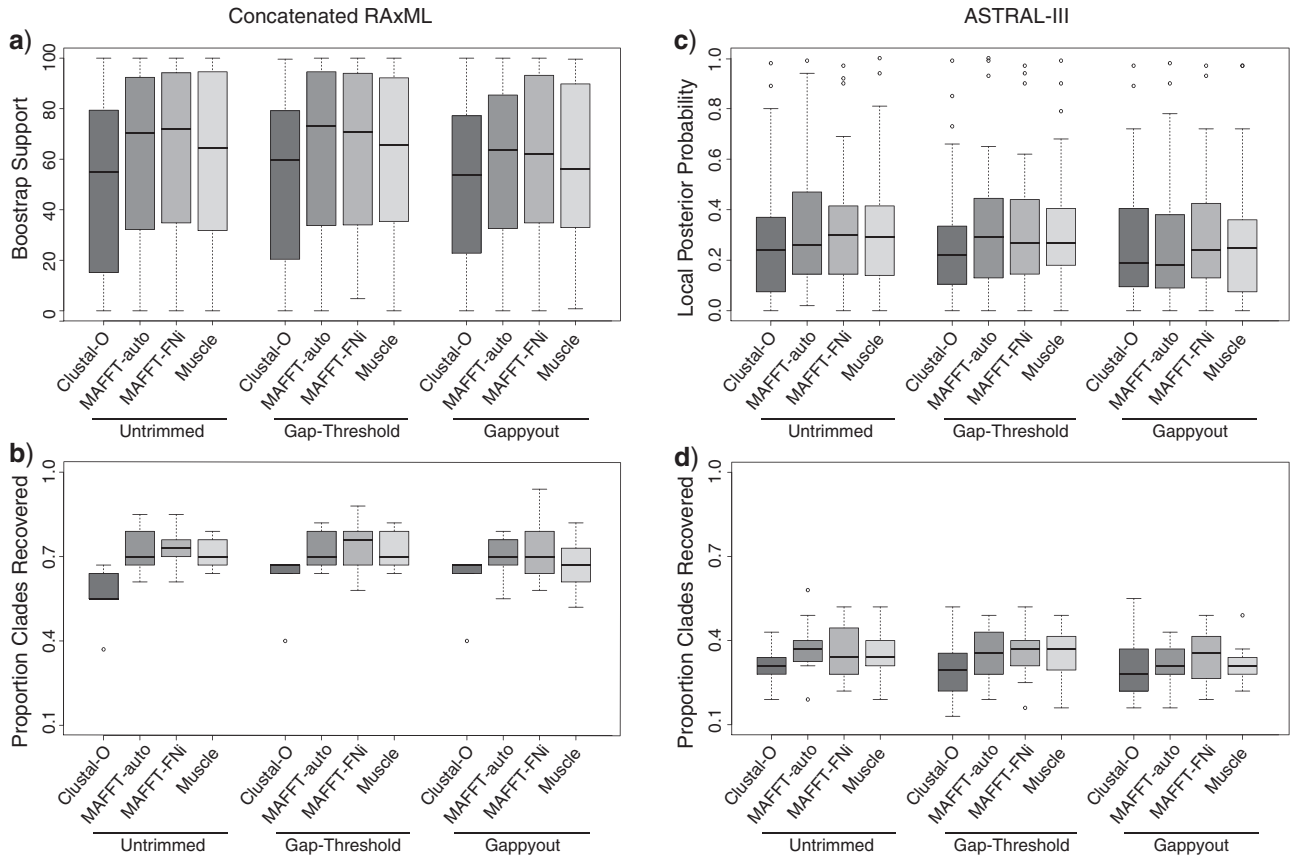

FIGURE 8. Mean support values and proportion of clades recovered for the squamate 1% (40-locus) data sets for the 35 well-established clades. Results are based on concatenated maximum likelihood (RAxML) and a species-tree method (ASTRAL-III). For RAxML, a) depicts the mean bootstrap support across the 35 well-established clades for each replicate ($n = 10$ replicates) and b) depicts the proportion of the 35 clades recovered for each replicate. For ASTRAL-III, c) depicts the mean LPP support across the 35 well-established clades for each replicate ($n = 20$ replicates) and d) depicts the proportion of the 35 clades recovered for each replicate. Boxplots show the median value (black line), interquartile range (box), values up to 1.5 times the interquartile range (whiskers), and outliers (dots).

other methods might give different results. For example, several methods such as SATé-II (Liu et al. 2012), PASTA (Mirarab et al. 2015), and UPP (Nguyen et al. 2015) have been shown to produce more accurate alignments than the methods used here. However, these methods were primarily designed to produce ultralarge alignments (>1000 sequences per gene region) and are not widely included in phylogenomic packages. We also limited our exploration of trimming to gap-rich sites, because our primary focus was on the effects of reducing sequence-length heterogeneity and missing data. Trimming based on variable sites, or using other popular trimming methods (e.g., Gblocks; Talavera and Castresana 2007), might produce different results than those observed here (see Ranwez and Chantret 2020). For example, the default settings of Gblocks cause it to aggressively remove gap-rich sites and nonconserved sites simultaneously. Although trimming nonconserved sites can potentially be mitigated (by tuning four parameters), strict gap removal only includes two options: (i) eliminate all columns containing any gaps or (ii) eliminate all columns containing gaps in >50% of sequences. In this study, we focused on the effect of trimming poorly aligned flanking regions, which resulted from a combination of true alignment gaps and gap sites resulting from missing data. Given this focus, we found the gap removal options of Gblocks were too coarse, and consequently we did not use them here. However, we acknowledge Gblocks offers a variety of useful options for eliminating highly variable sites (particularly in columns with low missing data), which we did not explore in our analyses.

We also acknowledge that the clades used here to evaluate method performance are not truly known. Nevertheless, the most important result here is that there was generally little difference in the trees from different alignment and trimming methods, regardless of whether these trees are right or wrong. Furthermore, we do not know of any realistic scenarios by which so many clades would be supported by both molecular and morphological data and would still be incorrect.

However, these clades might not be a random sample of all clades throughout the tree. Specifically, we expect well-established clades to be associated with longer branches, as these are the clades on which most genes agree (e.g., Wiens et al. 2008, 2012). On the other hand, dismissing these results based on the idea that all of these clades are "easy" to reconstruct is not accurate either. All methods had difficulty recovering one or more of the well-established clades in each of the three data sets, even when thousands of loci were sampled (e.g., Amphisbaenia, Coraciiformes, Archosauria). Moreover, when we reduced the number of loci sampled, methods sometimes had difficulty in recovering even 50% of these clades. (e.g., Fig. 8b, d).

Recommendations for Alignment Methods in Phylogenomics

We found that different alignment methods (Clustal-O, MAFFT, Muscle) estimated alignments that differed significantly in lengths and number of informative sites (Figs. 2–4, Tables 1 and 2). Multiplied across loci, these different methods generated concatenated alignments that differed by up to 3.1 million base pairs and 650,000 informative sites (Table 2). Despite these differences, we did not find any significant differences in clade recovery or support values across alignment methods using our full squamate, bird, and tetrapod data sets (for the well-established clades). However, with reduced gene sampling the Clustal-O alignments recovered significantly fewer established clades than other methods (Fig. 8; Supplementary File S5). Clustal-O produced the shortest alignments with the highest number of informative sites (Tables 1 and 2), and the most dissimilar gene trees relative to other methods (Supplementary File S3). The higher number of informative sites likely resulted from poorly aligned flanking regions, resulting in higher gene tree error. Given our observations, we do not recommend Clustal-O for UCE data.

Overall, we found similar results using Muscle and MAFFT (auto and FFT-NS-i), suggesting that both are good options for UCE data. One benefit of using MAFFT over Muscle is the automatic selection of the alignment algorithm based on the input alignment characteristics. During analyses, we observed that the MAFFT-auto option generally selected the L-INS-i algorithm. This algorithm is particularly well-suited to loci with one main alignable domain surrounded by flanking sequences, and with <200 taxa (Katoh et al. 2005; Katoh and Standley 2013). Based on this (and our results), we recommend MAFFT-auto for UCE data.

Recommendations for Trimming Methods

We examined the impact of sequence-length heterogeneity on phylogenomics by lightly trimming (gap-threshold) and aggressively trimming (gappyout) our alignments. Our squamate data set contained a mix of short (sequence-capture) and long (genome-extracted) UCE sequences, whereas our bird and tetrapod data sets contained primarily long (genome-extracted) UCE sequences. The squamate data set therefore had greater sequence-length heterogeneity. For birds and tetrapods, ASL-CV values were already low and trimming did not further reduce heterogeneity (Figs. 3c and 4c). For squamates, light and aggressive trimming reduced ASL-CV values (Fig. 2c, Table 1).

Across the three data sets, trimming did not increase clade recovery or support values for well-established clades for any alignment method (Supplementary File S5). In contrast, aggressive trimming decreased clade recovery for species-tree analyses when gene sampling was reduced (10% and 1% of loci), particularly for squamates. Thus, our results mirror the single-locus trimming effects found by Tan et al. (2015), but at the phylogenomic scale.

Overall, the type of sequence-length heterogeneity present in the squamate UCE data set (e.g., driven by longer genome-extracted sequences) did not appear to be problematic for our analyses (but see Hosner et al. 2016 for a different example with UCEs). Under this type of scenario, we do not recommend aggressive trimming to eliminate sequence-length heterogeneity, because it had greater potential to negatively impact analyses.

We found light trimming (e.g., gap-threshold, 16–30% of total alignment columns) was useful for eliminating poorly aligned flanking regions (e.g., the change from Fig. 1a to b) without negative downstream effects. Our best phylogenetic results for all three data sets were obtained from untrimmed and lightly trimmed alignments, and we recommend both options.

Many phylogenomic workflows employ a trimming routine. The custom trimming routine available in PHYLUCE is rather aggressive in removing alignment columns and is most comparable to gappy-out trimming in our study (Supplementary File S2, Table S31). Our results suggest trimming with this method may not be advantageous, at least under the default settings. We also suggest that the ASL-CV index introduced here might be useful for summarizing sequence-length heterogeneity in future studies.

Data Set Size in Phylogenomic Studies: Inadequate versus Adequate versus Overkill

We found that different alignment and trimming methods had little impact on trees from our full data sets, but decreasing the number of loci did. Specifically, the proportion of well-established clades recovered and/or their mean support values dropped significantly when we only included 10% (∼500) or 1% (∼50) of the total loci (∼5000). These results confirm that it is worthwhile to obtain data from thousands of loci, rather than dozens or hundreds. However, ∼5000 loci were still not enough to strongly resolve all relationships within squamates, birds, and tetrapods (even when only considering well-established clades). We also note

that for squamates, UCE loci were seemingly not as informative as similar numbers of nuclear protein-coding loci, since analyses of 44 nuclear protein-coding loci (Wiens et al. 2012) recovered stronger support for most clades than 40 subsampled UCE loci, including many of the well-established clades considered here.

Phylogenomic Methods: Concatenation versus Species-Tree Analyses

One particularly important and unexpected aspect of our results is that the concatenated analyses (RAxML) recovered a higher proportion of well-established clades than the species-tree method used (ASTRAL-III), particularly with fewer loci. There is a large literature suggesting that species-tree methods should be more accurate than concatenated analyses, especially when incomplete lineage sorting is high (Liu et al. 2010; Liu and Yu 2011; Mirarab et al. 2014; Mirarab and Warnow 2015; Vachaspati and Warnow 2015). However, simulation studies have also revealed that concatenated analyses can be more accurate when incomplete lineage sorting is low (Leaché and Rannala 2010; Bayzid and Warnow 2013; Patel et al. 2013; Bayzid et al. 2015; Chou et al. 2015; Mirarab et al. 2016). It is possible that the levels of incomplete lineage sorting associated with the well-established clades are sufficiently low to drive the observed differences in method performance. Another potential explanation is that UCE data have properties that differ from the data simulated in the studies cited above. For example, Mirarab et al. (2014) found that concatenated analyses might be more accurate than species-tree methods when gene trees each have relatively poor phylogenetic signal. Our results suggest that concatenated analyses may outperform species-tree analyses most strongly when fewer loci are sampled (and holding constant the branch lengths and phylogenetic signal of genes). We have observed similar patterns in other empirical analyses of UCE data that compared the ability of these methods to recover well-established clades (e.g., Streicher et al. 2016, 2018). However, we note that we have not tested all species-tree and concatenated methods. Overall, we simply caution that species-tree methods should not be assumed to perform better than concatenated methods in phylogenomic analyses, especially for UCE data sets.

Implications for Squamate Phylogeny

Here, we present possibly the most extensive phylogenomic analysis of higher-level squamate phylogeny to date (our data sets for birds and tetrapods are not so exceptional; Jarvis et al. 2014; Irisarri et al. 2017). For example, some previous studies included more taxa but far fewer loci (e.g., 161 taxa, 44 loci; Wiens et al. 2012) whereas others had similar numbers of loci but far fewer taxa (4178 loci, 32 taxa; Streicher and Wiens 2017). Here, we simultaneously analyze a relatively large number of loci and taxa (123 species, up to 4430 loci per species). Our concatenated analyses recovered the highest proportion of the 35 well-established clades, and we focus on those results here (Fig. 5). Overall, our results are largely congruent with previous higher-level analyses but provide strong support for some previously controversial relationships.

First, we strongly support dibamids as the sister group to all other squamates. This has precedents in some previous studies (e.g., Townsend et al. 2004; Pyron et al. 2013; Tonini et al. 2016), but others found only weak support (e.g., Zheng and Wiens 2016; Streicher and Wiens 2017) or conflicting relationships (Wiens et al. 2012; Reeder et al. 2015). Interestingly, our analyses using ASTRAL-III place dibamids in an unusual position (relatively distant from the root) that we have not seen reported in any earlier studies (Fig. 6). This seems problematic.

Second, we strongly support snakes as the sister group to a clade including Iguania and Anguimorpha, in both concatenated and species-tree analyses (Figs. 5 and 6). The placement of snakes within Toxicofera has been controversial or weakly supported in previous studies with fewer loci (e.g., Vidal and Hedges 2005; Pyron et al. 2013; Zheng and Wiens 2016).

Our results for pleurodont iguanians are generally weakly supported (as in most previous studies). However, we do find strong support for placing Phrynosomatidae as the sister taxon to other members of this large clade (see also Townsend et al. 2011; Streicher et al. 2016).

Finally, our results help resolve the controversial placement of iguanians (e.g., Losos et al. 2012) and show that they are not at the base of squamate phylogeny. Overall, we provide strong support for many higher-level squamate relationships based on extensive sampling of genes and taxa.

Data Accessibility

We developed a publicly available project page using the Open Science Framework (OSF) that contains the complete set of data and instructions required to replicate all of our analyses, available at: https://osf.io/qa9r8/. This material includes the starting UCE sequence sets for SuperCRUNCH, inputs and outputs for key steps conducted in SuperCRUNCH, and the final per-locus and concatenated alignments for each data set (squamates, birds, and tetrapods). We provide all necessary inputs for phylogenetic analyses, the results of all phylogenetic analyses, and summaries of all phylogenetic results using MonoPhylo. We also provide the data sets resulting from gene subsampling, along with results from all subsequent analyses. The sequence length heterogeneity calculator (for ASL-CV and other metrics) has been made available as a module of SuperCRUNCH (Sequence_Length_Heterogeneity.py), and is freely available at: https://github.com/dportik/SuperCRUNCH. MonoPhylo is open-source and

freely available at https://github.com/dportik/MonoPhylo.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.p8cz8w9mh.

Funding

This work was supported by U.S. National Science Foundation (DEB 1655690).

Acknowledgments

We thank Jeff Streicher for assistance with the squamate UCE data sets. We thank Brant Faircloth, Richard Glor, and anonymous reviewers for helpful comments on the manuscript.

References

Andermann T., Cano A., Zizka A., Bacon C., Antonelli A. 2018. SECAPR—a bioinformatics pipeline for the rapid and user-friendly processing of targeted enriched Illumina sequences, from raw reads to alignments. PeerJ 6:e5175.

Andermann T., Jiménez M.F.T., Matos-Maraví P., Batista R., Blanco-Pastor J.L., Gustafsson A.L.S., Kistler L., Liberal I.M., Oxelman B., Bacon C.D., Antonelli A. 2020. A guide to carrying out a phylogenomic target sequence capture project. Front. Genet. 10:1407.

Antonelli A., Hettling H., Condamine F.L., Vos K., Nilsson R.H., Sanderson M.J., Sauquet H., Scharn R., Silvestro D., Töpel M., Bacon C.D., Oxelman B., Vos R.A. 2017. Toward a self-updating platform for estimating rates of speciation and migration, ages, and relationships of taxa. Syst. Biol. 66:152–166.

Bayzid M.S., Warnow T. 2013. Naive binning improves phylogenomic analyses. Bioinformatics 29:2277–2284.

Bayzid M.S., Mirarab S., Boussau B., Warnow T. 2015. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. PLoS One 10:30129183.

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.

Bennett D.J., Hettling H., Silvestro D., Zizka A., Bacon C.D., Faurby S., Vos R.A., Antonelli A. 2018. phylotaR: an automated pipeline for retrieving orthologous DNA sequences from GenBank in R. Life 8:20.

Bi K., Vanderpool D., Singhal S., Linderoth T., Moritz C., Good J.M. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics 13:403.

Bossert S., Danforth B.N. 2018. On the universality of target-enrichment baits for phylogenomic research. Methods Ecol. Evol. 9:1453–1460.

Capella-Gutiérrez S., Silla-Martínez JM., Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.

Chan K.O., Hutter C.R., Wood P.L. Jr., Grismer L.L., Brown R.F. 2020. Larger, unfiltered datasets are more effective at resolving phylogenetic conflict: introns, exons, and UCEs resolve ambiguities in Golden-backed frogs (Anura: Ranidae; genus Hylarana). Mol. Phylog enet. Evol. 151:106899.

Chatzou M., Magis C., Chang J.M., Kemena C., Bussotti G., Erb I., Notredame C. 2016. Multiple sequence alignment modeling: methods and applications. Brief. Bioinform. 17:1009–1023.

Chou J., Gupta A., Yaduvanshi S., Davidson R., Nute M., Mirarab S., Warnow T. 2015. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. BMC Genomics 16:S2.

Dress A.W.M., Flamm C., Fritzsch G., Gruñewald S., Kruspe M., Prohaska S.J., Stadler P.F. 2008. Noisy: identification of problematic columns in multiple sequence alignments. Algorithm Mol. Biol. 3:7.

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

Edgar R.C., Batzoglou S. 2006. Multiple sequence alignment. Curr. Opin. Struct. Biol. 16:368–373.

Estes R, de Queiroz K., Gauthier J. 1988. Phylogenetic relationships within Squamata. In: Estes R. and Pregill G., editors. Phylogenetic relationships of the lizard families. Palo Alto, CA: Stanford University Press. p. 119–281.

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. 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics 32:786-788.

Freyman W.A. 2015. SUMAC: constructing phylogenetic supermatrices and assessing partially decisive taxon coverage. Evol. Bioinformatics 11:263–266.

Harris R.S. 2007. Improved pairwise alignment of genomic DNA [Ph.D. Thesis]. The Pennsylvania State University.

Hedin M., Derkarabetian S., Alfaro A., Ramírez M.J., Bond, J.E. 2019. Phylogenomic analysis and revised classification of atypoid mygalomorph spiders (Araneae, Mygalomorphae), with notes on arachnid ultraconserved element loci. PeerJ 7:e6864.

Hosner P.A., Faircloth B.C., Glenn T.C., Braun E.L., Kimball R.T. 2016. Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes). Mol. Biol. Evol. 33:1110–1125.

Huerta-Cepas J., Serra F., Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. 33:1635–1638.

Hutter C.R., Cobb K.A., Portik D.M., Travers S., Wood P.L. Jr., Brown R.M. 2019. FrogCap: A modular sequence capture probe set for phylogenomics and population genetics for all frogs, assessed across multiple phylogenetic scales. bioRxiv 825307.

Irisarri I., Baurain D., Brinkmann H., Delsuc F., Sire J.Y., Kupfer A., Petersen J., Jarek M., Meyer A., Vences M., Philippe H. 2017. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat. Ecol. Evol. 1:1370-1378.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J.W., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldon T., Capella-Gutierrez S., HuertaCepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X.J., Dixon A., Li S.B., Li N., Huang Y.H., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M.V., AlfaroNunez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z.J., Zeng Y.L., Liu S.P., Li Z.Y., Liu B.H., Wu K., Xiao J., Yinqi X., Zheng Q.M., Zhang Y., Yang H.M., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jonsson K.A., Johnson W., Koepfli K.P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alstrom P., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G.J 2014. Whole genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:722–780.

Katoh K., Kuma K., Toh H., Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.

Katoh K., Misawa K., Kuma K., Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.

Kemena C., Notredame C. 2009. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. Bioinformatics 25:2455–2465.

Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D. 2002. The human genome browser at UCSC. Genome Res. 12:996–1006.

Leaché A.D., Linkem C.W. 2015. Phylogenomics of horned lizards (Genus: *Phryonosoma*) using targeted sequence capture data. Copeia 103:586–594.

Leaché A.D., Rannala B. 2010. The accuracy of species tree estimation under simulation: a comparison of methods. Syst. Biol. 60:126–137.

Leaché A.D., Banbury B.L., Linkem C.W., Nieto-Montes de Oca, A. 2016. Phylogenomics of a rapid radiation: is chromosomal evolution linked to increased diversification in North American spiny lizards (Genus *Sceloporus*)? BMC Evol. Biol. 16:63.

Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61:727–744.

Linkem C.W., Minin V.N., Leaché A.D. 2016. Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (Squamata: Scincidae). Syst. Biol. 65:465–477.

Liu K., Warnow T.J., Holder M.T., Nelesen S.M., Yu J., Stamatakis A.P., Linder C.R. 2012. SATé-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. Syst. Biol. 61:90–106.

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. 10:1–18.

Longo S.J., Faircloth B.C., Meyer A., Westneast M.W., Alfaro M.E., Wainwright P.C. 2017. Phylogenomic analysis of a rapid radiation of misfit fishes (Syngnathiformes) using ultraconserved elements. Mol. Phylogenet. Evol. 113:33–48.

Losos J.B., Hillis D.M., Greene H.W. 2012. Who speaks with a forked tongue? Science 338:1428–1429.

Mirarab S., Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31:i44–i52.

Mirarab 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.

Mirarab S., Nguyen N., Guo S., Wang L.-S., Kim J., Warnow T. 2015. PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. J. Comput. Biol. 22:377–386.

Mirarab S., Bayzid M.S., Warnow T. 2016. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. 65:366–380.

Molloy E.K., Warnow T. 2018. To included or not to include: the impact of gene filtering on species tree estimation methods. Syst. Biol. 67:285–303.

Nguyen N., Mirarab S., Kumar K., Warnow T. 2015. Ultra-large alignments using phylogeny-aware profiles. Genome Biol. 16:124.

Nute M., Warnow T. 2016. Scaling statistical multiple sequence alignment to large datasets. BMC Genomics 17(Suppl 10):764.

Nute M., Chou J., Molloy E.K., Warnow T. 2018. The performance of coalescent-based species tree estimation methods under models of missing data. BMC Genomics 19(Suppl 5):286.

Ogden T.H., Rosenberg M.S. 2006. Multiple sequence alignment accuracy and phylogenetic inference. Syst. Biol. 55:314–328.

Patel S., Kimball R., Braun E. 2013. Error in phylogenetic estimation for bushes in the tree of life. J. Phylogenet. Evol. Biol. 1:110.

Pearse W.D., Purvis A. 2013. phyloGenerator: an automated phylogeny generation tool for ecologists. Methods Ecol. Evol. 4:692–698.

Portik D.M., Wiens J.J. 2020. SuperCRUNCH: a toolkit for creating and manipulating supermatrices and other large phylogenetic datasets. Methods Ecol. Evol. 11:763–772.

Portik D.M., Smith L.L., Bi K. 2016. An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). Mol. Ecol. Resour. 16:1069–1083.

Prum R.O., Berv J.S., Dornburg A., Field D.J, Townsend J.P, Lemmon E.M, Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526:569–573.

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.

R Core Team. 2018. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Ranwez V., Chantret N. 2020. Strengths and limits of multiple sequence alignment and filtering methods. In: Scornavacca C., Delsuc F., and Galtier N., editors. Phylogenetics in the genomic era. No commercial publisher (open access book). p. 2.2.1–36.

Reddy S., Kimball R.T., Pandey A., Hosner P.A., Braun M.J., Han K.-L., Harshman J., Hackett S.J., Huddleston C.J., Kingston S., Marks B.D., Miglia K.J., Moore W.S., Sheldon F.H., Witt C.C., Yuri T., Braun E.L. 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. Syst. Biol. 66:857–879.

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.

Roch S., Nute M., Warnow T. 2019. Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods. Syst. Biol. 68:281–297.

Sayyari E., Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33:1654–1668.

Schott R.K., Panesar B., Card D.C., Preston M., Castoe T.A., Chang B.S.W. 2017. Targeted capture of complete coding regions across divergent species. Genome Biol. Evol. 9:398–414.

Sievers F., Wilm A., Dineen D., Gibson T.J., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., Thompson J.D., Higgins D.G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7:539.

Smith S.A., Walker J.F. 2019. PyPHLAWD: a python tool for phylogenetic dataset construction. Methods Ecol. Evol. 10:104–108.

Smythe A.B., Sanderson M.J., Nadler S.A. 2006. Nematode small subunit phylogeny correlates with alignment parameters. Syst. Biol. 55:972–992.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Streicher J.W., Wiens J.J. 2016. Phylogenomic analyses reveal novel relationships among snake families. Mol. Phylogenet. Evol. 100:160–169.

Streicher J.W., Wiens J.J. 2017. Phylogenomic analyses of more than 4,000 nuclear loci resolve the origin of snakes among lizard families. Biol. Lett. 13:20170393.

Streicher J.W., Schulte J.A., Wiens J.J. 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. Syst. Biol. 65:128–145.

Streicher J.W., Miller E.C, Guerrero P.C., Correa C., Ortiz J.C., Crawford A.J., Pie M.R., Wiens J.J. 2018. Evaluating methods for phylogenomic analyses, and a new phylogeny for a major frog clade (Hyloidea) based on 2,214 loci. Mol. Phylogenet. Evol. 119:128–143.

Tagliacollo V.A., Lanfear R. 2018. Estimating improved partitioning schemes for ultraconserved elements. Mol. Biol. Evol. 35:1798–1811.

Talavera G., Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56:564–577.

Tan G., Muffato M., Ledergerber C., Herrero J., Goldman N., Gil M., Dessimoz C. 2015. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. Syst. Biol. 64:778–791.

Thompson J.D., Linard B., Lecompte O., Poch O. 2011. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PLoS One 6:e18093.

Tonini J.F.R., Beard K.H., Ferreira R.B., Jetz W., Pyron R.A. 2016. Fully-sampled phylogenies of squamates reveal evolutionary patterns in threat status. Biol. Conserv. 204:23–31.

Townsend T., Larson A., Louis E.J., Macey J.R. 2004. Molecular phylogenetics of Squamata: the position of snakes, amphisbaenians, and dibamids, and the root of the squamate tree. Syst. Biol. 53:735–757.

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.

Vachaspati P., Warnow T. 2015. ASTRID: accurate species trees from internode distances. BMC Genomics 16:1–13.

Vidal N., Hedges S.B. 2005. The phylogeny of squamate reptiles (lizards, snakes, and amphisbaenians) inferred from nine nuclear protein coding genes. C. R. Biol. 328:1000–1008.

White, N.D., Braun, M.J. 2019. Extracting phylogenetic signal from phylogenomic data: higher-level relationships of the nightbirds (Strisores). Mol. Phylogenet. Evol. 141:106611.

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

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.

Wu M., Chatterji S., Eisen J.A. 2012. Accounting for alignment uncertainty in phylogenomics. PLoS One 7:e30288.

Xi Z., Liu L., Davis C.C. 2016. The impact of missing data on species tree estimation. Mol. Biol. Evol. 33:838–860.

Zhang C., Sayyari E., Mirarab S. 2017. ASTRAL-III: increased scalability and impacts of contracting low support branches. In: Meidanis J., Nakhleh L., editors. Comparative genomics. RECOMB-CG 2017. Lecture notes in computer science, vol. 10562. Cham: Springer. p. 53-75.

Zheng Y., Wiens J.J. 2016. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4,162 species. Mol. Phylogenet. Evol. 94:537–547.