PHYLOGENOMICS OF RAPID AVIAN RADIATIONS

By

Carl H. Oliveros

Submitted to the graduate degree program in Ecology and Evolutionary Biology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

Chairperson: Robert G. Moyle

_____

Rafe M. Brown

_____

John K. Kelly

_____

Xingong Li

_____

A. Townsend Peterson

Date Defended: 17 July 2015

The Dissertation Committee for Carl H. Oliveros

certifies that this is the approved version of the following dissertation:

Phylogenomics of Rapid Avian Radiations

_____

Chairperson: Robert G. Moyle

Date approved: 20 July 2015

ABSTRACT

I use data from sequence capture of ultraconserved elements to resolve three rapid radiations in the avian tree of life and in the process gain insights on applying analytical strategies with gene tree-based coalescent methods (GCM). In Chapter 1, I explore analytical strategies that can be employed with GCMs to increase phylogenetic resolution and minimize highly supported conflicting results, including subsampling taxa to increase the number of gene trees analyzed, trimming sequences to eliminate sequence length heterogeneity, and filtering loci based on information content. These strategies are used to reconstruct a highly resolved and consistent phylogenetic hypothesis for the relatively young avian family, Zosteropidae. I show how conflicting results from different GCMs can arise from biases introduced by sequence length heterogeneity and uninformative loci that can lead to strongly supported incorrect estimates of phylogeny. In Chapter 2, I examine higher-level relationships in the enigmatic core Corvoidea group of Oscine passerines. A highly resolved phylogeny of core Corvoidea is recovered, with a majority of nodes receiving high support from both ML and coalescent analyses. I show that short sequence lengths do not bias species tree estimates of GCMs if informative sites are present in these sequences. In contrast, some samples that have longer sequence lengths compared to most taxa but shorter sequence lengths compared to taxa in its clade can also bias species tree estimates of GCMs. In Chapter 3, I develop a hypothesis on the origins of the trogons (Trogonidae) based on a robust dated phylogeny estimated from thousands of genome-wide loci. I recover the first well-supported hypothesis of relationships among trogon genera. This topology, combined with the trogon fossil record, geologic, and climatic data, suggests an Old World origin in the Late Oligocene/Early Miocene for the crown group. In this

chapter, I show that in some datasets in which loci have high information content, exclusion of less informative loci in analysis can lead to lower bootstrap support of species tree estimates of GCMs.

ACKNOWLEDGEMENTS

I am grateful to the following people and their institutions for providing tissue and toe pad loans: Nate Rice, Academy of Natural Sciences of Drexel University; Paul Sweet, Tom Trombone, and Peter Capainolo, American Museum of Natural History; Herman Mays, Cincinnati Museum Center; Ben Marks, Field Museum of Natural History; Donna Dittmann, Louisiana State University Museum of Natural Science; Helen James and Chris Milensky, National Museum of Natural History; Mark Robbins, University of Kansas Natural History Museum; Chris Witt and Andrew Johnson, University of New Mexico Museum of Southwestern Biology; Sharon Birks, University of Washington Burke Museum of Natural History and Culture; Ron Johnstone, Western Australia Museum; Kristof Zyskowski, Yale Peabody Museum.

Brant Faircloth, Noor White, Brian Smith, and Michael Harvey gave advice on UCE laboratory protocols; Brant Faircloth offered guidance on UCE data processing.

I thank Mike Andersen and Rob Moyle for providing edits and suggestions on Chapters 1 and 2. Rob Moyle, Mike Andersen, Pete Hosner, Fred Sheldon, and Joel Cracraft provided comments and edits on an earlier version of Chapter 3.

I am grateful to the following funding sources for supporting my research: National Science Foundation (NSF) Doctoral Dissertation Improvement Grant, American Museum of Natural History Frank M. Chapman Fund, KU Biodiversity Institute Panorama Grants, KU Graduate Student Research Fund. The KU Graduate Fellowships, KU EEB Summer Fellowships, and NSF grants to Rob Moyle and Rafe Brown also supported me as a Graduate Research Assistant.

My graduate career was enriched by unforgettable (and some forgettable) experiences in the field and I thank Rob Moyle, Town Peterson, and Mark Robbins for giving me the opportunity to work in places like Cambodia, the Democratic Republic of Congo, Indonesia, Palau, and the Philippines. I also express my deep appreciation for collaborators and field workers in those countries, who are too many to name, for all the memories.

My graduate career and life in Lawrence would not have been as enjoyable without friends in Lawrence including EEB graduate students and faculty, their significant others and family, and the Filipino jayhawks. I especially thank members of the Moyle lab—Michael Andersen, Luke Campillo, Pete Hosner, Muhammad Janra, Robin Jones, Luke Klicka, and Joe Manthey—for being sources of advice, support, fun, and merriment.

Members of my Ph.D. committee, Rafe Brown, John Kelly, Xingong Li, Rob Moyle, and Town Peterson, have been very supportive throughout all these years. Mark Holder, whom I recently had to replace in my committee with John Kelly because of a scheduling conflict, has been very helpful as well. I am deeply grateful to my main advisor, Rob Moyle, who has been a terrific mentor and friend.

My family in the Philippines has been a great source of energy, encouragement, and love even if they are on the other side of the globe. Most especially, I am privileged to have the love and support of my wife, Cynthia.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

INTRODUCTION

High-throughput sequencing is revolutionizing the field of systematics. Techniques such as whole-genome sequencing and re-sequencing, reduced representation strategies, sequence capture, and RNA sequencing are now employed to examine population structure, patterns of diversification, and phylogenetic relationships in birds (Toews et al., In review). The advances in sequence data acquisition has brought an unprecedented amount of data to ornithologists, exemplified in a recent study of higher-level phylogenetic relationships of birds based on 48 complete genomes (Jarvis et al. 2014). Sequence capture techniques (Bi et al. 2012; Faircloth et al. 2012; Lemmon et al. 2012) have been favored in deep level phylogenetic studies involving several taxa because they provide a large number of orthologous sequences for comparisons.

The massive amount of sequence data from new sequencing techniques also poses challenges for systematic ornithologists. First, bioinformatics skills are required to handle and examine the data. Second, phylogenetic inference methods that process data in this scale are still in their infancy. For example, a debate on the use of traditional concatenation approaches vs. gene tree-based coalescent methods (GCMs) in species tree inference is ongoing (Song et al. 2012; Gatesy and Springer 2014; Springer and Gatesy 2014; Zhong et al. 2014; Liu et al. 2015; Tonini et al. 2015). At the same time, the large number of loci in phylogenomic datasets provide opportunities to resolve nodes in the Tree of Life that are otherwise difficult to untangle with small datasets, especially rapid radiations, which cause high levels of incomplete lineage sorting (Whitfield and Lockhart 2007). A high number of loci allows us to better sample gene tree distributions generated by the species trees.

In this dissertation, I use data from sequence capture of ultraconserved elements to resolve three rapid radiations in the avian tree of life and in the process gain insights on analytical approaches and strategies for dealing with phylogenomic datasets.

In Chapter 1, I explore analytical strategies that can be employed with GCMs to increase phylogenetic resolution and minimize highly supported conflicting results. I show how conflicting results from different GCMs can arise from biases introduced by sequence length heterogeneity and uninformative loci that can lead to strongly supported incorrect estimates of phylogeny. Strategies for eliminating these biases and increasing resolution of species tree estimates of GCMs are proposed and tested, including subsampling taxa to increase the number of gene trees analyzed, trimming sequences to eliminate sequence length heterogeneity, and filtering loci based on information content. These strategies are used to reconstruct phylogenetic relationships in the relatively young avian family, Zosteropidae, using four different GCMs from sequence capture data. The resulting estimates are highly resolved, consistent with each other, and comparable to the maximum likelihood estimate of the concatenated data. The analytical strategies learned in this Chapter are applied and explored further in the subsequent chapters.

In Chapter 2, I examine higher-level relationships in the enigmatic core Corvoidea group of Oscine passerines, a clade comprising 773 species that occur worldwide. Phylogenetic relationships among the 29 families in this large clade have been largely unresolved owing to a combination of inadequate character and taxon sampling in previous studies and short time intervals between lineage splitting events. I use sequence data from thousands of ultraconserved element loci obtained from 86 species to estimate higher level phylogenetic relationships in the group using maximum likelihood (ML) and coalescent methods. A highly resolved phylogeny of core Corvoidea is recovered, with a majority of nodes receiving high support from both ML and

coalescent analyses. Four families restricted to or thought to have originated from the Australia-New Guinea region and New Zealand form the earliest branching lineages in the group. The other families are divided into three major clades each consisting of 8–9 families. In this chapter, I show that short sequence lengths do not bias species tree estimates of GCMs if informative sites are present in these sequences. In contrast, some samples that have longer sequence lengths compared to most taxa but shorter sequence lengths compared to taxa in its clade can also bias species tree estimates of GCMs.

In Chapter 3, I develop a hypothesis on the origins of the trogons (Trogonidae) based on a robust dated phylogeny estimated from thousands of genome-wide loci. Phylogenetic reconstruction in the trogons has been problematic in previous studies for two reasons: (1) successive short internodes at the base of this radiation cause high gene tree discordance, and (2) a long branch leading to the family makes root placement problematic. I recover the first well-supported hypothesis of relationships among trogon genera. Trogons comprise three clades, each confined to one of three biogeographic regions: Africa, Asia, and the Neotropics, with the African clade sister to the rest of trogons. This topology, combined with the trogon fossil record, geologic, and climatic data, suggests an Old World origin for the crown group. Continental connections during the warm Late Oligocene/Early Miocene facilitated dispersion between Eurasia, North America, and Africa, and subsequent global cooling plausibly caused divergence between main trogon lineages. In this chapter, I show that in some datasets in which loci have high information content, exclusion of less informative loci can lead to lower bootstrap support of species tree estimates of GCMs.

CHAPTER 1

Strategies for Consistent Species Tree Inference from Sequence Capture Data: a Case Study with the White-eyes (Aves: Zosteropidae)

ABSTRACT

Gene tree-based coalescent methods (GCM) for estimating species trees from phylogenomic datasets have been criticized, among others, for producing conflicting results when applied to the same data and yielding lower resolution compared to results from maximum likelihood inference on concatenated data. In this paper we show how conflicting results from different GCMs can arise from biases introduced by sequence length heterogeneity and uninformative loci that can lead to strongly supported incorrect estimates of phylogeny. Strategies for eliminating these biases and increasing resolution of species tree estimates of GCMs are proposed and tested, including subsampling taxa to increase the number of gene trees analyzed, trimming sequences to eliminate sequence length heterogeneity, and filtering loci based on information content. These strategies are used to reconstruct phylogenetic relationships in the avian family Zosteropidae using four different GCMs from sequence capture of ultraconserved elements. The resulting estimates are highly resolved, consistent with each other, and comparable to the maximum likelihood estimate of the concatenated data.

INTRODUCTION

Collection of DNA sequence data from hundreds to thousands of genome-wide loci is now in wide practice for phylogenetic studies. Sequence capture techniques (Bi et al. 2012; Faircloth et al. 2012; Lemmon et al. 2012) provide large numbers of orthologous sequences that have been used to resolve deep-level phylogenetic relationships in various groups of organisms (Crawford et al. 2012; McCormack et al. 2012, 2013; Faircloth et al. 2013; Brandley et al. 2015). Two major approaches are currently employed to estimate phylogenies with datasets of this scale. In the traditional approach, all loci are concatenated into a large supermatrix on which maximum likelihood (ML) or Bayesian tree inference is performed. This approach is statistically inconsistent and yield highly supported incorrect topologies if species trees are in the anomaly zone (Kubatko and Degnan 2007; Roch and Steel 2015), but it is otherwise expected to perform well if species tree branch lengths are long enough such that individual gene trees are congruent (Kubatko and Degnan 2007; Mirarab et al. 2014b; Liu et al. 2015). A second approach incorporates gene tree discordance in species tree inference. An entire class of methods that use information from inferred gene trees to estimate the species tree has gained wide use owing to their theoretical strengths and computational tractability. These methods follow a two-step process. First, gene trees are estimated for each locus (usually using a maximum likelihood approach), then the species tree is estimated based on information from these gene trees. The weaknesses of these gene tree-based methods include: (1) different gene tree-based methods produce highly supported conflicting results from the same data (Gatesy and Springer 2014; Springer and Gatesy 2014), (2) branch support values are lower compared to those obtained from concatenation techniques (Edwards 2009; Liu et al. 2009b), (3) some uncontroversial clades are not recovered by these methods, whereas concatenation methods recover them with high support

(Gatesy and Springer 2014), (4) species tree estimates are less accurate compared to those obtained by concatenation when the inference is based on gene trees estimated from sequence data (Gatesy and Springer 2014), and (5) potential recombination within individual loci violates one of the main assumptions of the multispecies coalescent model (Gatesy and Springer 2014).

In this paper, we propose and test analytical strategies to address the first three weaknesses of gene tree-based methods in an empirical context. Specifically, we demonstrate that analyzing subsamples of taxa, trimming sequence data from some loci, and filtering loci based on information content can increase resolution and minimize highly supported conflicting results in gene tree-based species tree methods. We illustrate these techniques using sequence capture of ultraconserved elements (UCEs) to estimate the phylogeny of the diverse avian family Zosteropidae, which contains one of the most rapid radiations known among vertebrates (Moyle et al. 2009).

Gene Tree-Based Coalescent Methods

Several gene tree-based techniques have been developed to estimate species trees: STEM (Kubatko et al. 2009), STAR (Liu et al. 2009b), STEAC (Liu et al. 2009b), MP-EST (Liu et al. 2010), GLASS (Mossel and Roch 2010), NJ-ST (Liu and Yu 2011), STELLS (Wu 2012), MulRF (Chaudhary et al. 2014), and ASTRAL (Mirarab et al. 2014c). In this paper, we compare the performance of four methods, STAR, STEAC, MP-EST, and ASTRAL, which are not truly coalescent based but are statistically consistent under the multispecies coalescent model (Liu et al. 2009b, 2010; Mirarab et al. 2014c). We henceforth refer to these methods as gene tree-based coalescent methods, or GCMs (Liu et al. 2015).

These four GCMs take different approaches to species tree estimation. STAR and STEAC are most similar in that both use distance-based tree reconstruction based on coalescent

information from gene trees; the former uses a matrix of average ranks of coalescences between taxa across all gene trees, whereas the latter uses a matrix of average coalescence times. In contrast, MP-EST maximizes a pseudo-likelihood function that is based on the frequency of rooted triples in gene trees. Lastly, ASTRAL finds the species tree that maximizes the number of compatible quartets in the gene trees. The first three methods require rooted gene trees as input, whereas the last accepts unrooted gene trees. In addition, STEAC is the sole method that requires branch length information in the gene trees.

Simulation and empirical studies provide insights into the performance of GCMs. For instance, these methods can be implemented with some species missing for some loci, but including such loci is not advisable for several reasons. First, missing species in some loci can introduce bias to species estimation results in some GCMs. A large number of missing species in some loci can bias the ranks in that particular locus in a STAR analysis (Zhong et al. 2014). The results of MP-EST may not be biased by missing species as long as missing species occur randomly across loci (Zhong et al. 2014), but this may not be the case with empirical data (see below). ASTRAL can also conceivably be biased by non-random missing lineages in cases where quartets that represent the correct topology are underrepresented in gene trees due to missing lineages from these quartets. Second, a large proportion of missing species can lead to low bootstrap support in MP-EST analysis, but this could be remedied by sampling a large number of loci (Liu et al. 2010; Zhong et al. 2014). Third, the effects of missing species in some loci have not been thoroughly studied for the GCMs used in this study, but see Hovmöller et al. (2013) for results on STEM. Fourth, because STAR, STEAC, and MP-EST require rooted gene trees, these require the same outgroup species to be present in all loci. Lastly, in order for results

of all four GCMs to be comparable in this study, species tree inferences should be based on the same data, requiring the use of datasets with no missing species in each locus.

Using a large number of loci benefits GCM analyses because they are statistically consistent; i.e., the probability of the estimated species tree being congruent to the true species tree increases to 1.0 as the number of gene trees analyzed increases (Liu et al. 2009b, 2010; Mirarab et al. 2014c). Apart from increasing accuracy, analyzing a large number of gene trees has also been shown to increase nodal support in species tree estimation (Song et al. 2012). Therefore, maximizing the number of loci should be a goal of data collection and assembly.

Poor phylogenetic signal in individual loci, which results in high gene tree estimation error, has been shown to result in poor accuracy of GCMs (Bayzid and Warnow 2013; Gatesy and Springer 2014) or low bootstrap support values for the estimated species trees (Zhong et al. 2013; Mirarab et al. 2014a; Liu et al. 2015). One approach proposed to overcome this issue is to combine alignments from multiple loci to increase phylogenetic signal in gene tree estimation (Bayzid and Warnow 2013; Mirarab et al. 2014a); but these "binning" techniques have been criticized as possibly biasing gene tree distributions and resulting in incorrect species tree topologies (Liu et al. 2015). Another strategy is to exclude weakly supported gene trees (e.g., omitting loci with < 50% average bootstrap support) from species tree analysis, which has been shown to improve bootstrap support values in MP-EST analyses (Zhong et al. 2013; Liu et al. 2015). Yet another solution is to augment phylogenetic signal by increasing phylogenomic data sets (i.e., increasing the number of loci or alignment lengths; Liu et al. 2015).

Missing Data in Sequence Capture Data Sets

It is important to understand the nature of missing data in empirical datasets. Sequence capture techniques result in missing data in phylogenomic datasets on two levels. First, for each

individual sample, not all loci targeted are enriched and some samples can have a significantly lower number of loci enriched compared with others (Fig. 1.1a). These samples are not missing data randomly across loci, thus they are expected to lack sequence data in a given locus more than other samples. Additionally, enrichment success rates are often significantly less than 100%, in our case ~80%. Combined with the stochastic nature of locus enrichment, this means that if 4000 out of 5000 loci are enriched for each of two samples, the number of enriched loci common to both samples will be < 4000. This number will be $0.80 \times 4000 = 3200$ loci, if enrichment was completely random. As the number of samples increases, therefore, the number of enriched loci in common to all samples decreases to a small fraction of loci enriched per individual (Fig. 1.1b). If GCM analyses are performed on complete matrices consisting of loci for which all samples have sequence data, only a small fraction of the data available is used. However, if datasets with fewer taxa are assembled for GCM analyses, the number of loci common to all samples can be significantly higher (Fig. 1.1b).

Secondly, sequence capture methods result in alignments with heterogeneous sequence lengths (i.e., some samples have missing nucleotides at their flanks caused by varying lengths and alignment position of assembled contigs). Retaining flanking sites with missing data in each alignment has the advantage of retaining informative sites for samples that have data at these sites, especially in some sequence capture techniques that yield loci in which informative sites are located towards the flanks (Lemmon et al. 2012; McCormack et al. 2012). However, missing data in alignment flanks can mislead gene tree estimation if the short sequences are missing a sufficient number of informative sites (Fig. 1.1c). Some samples have consistently short sequences across many loci because of the suboptimal quality of DNA extracts (e.g., DNA extracted from museum skins or from poorly preserved tissue). For these samples, if their

placement in gene trees is misestimated in a significant number of loci, then their placement in species tree analyses based on gene trees is expected to be misestimated as well.

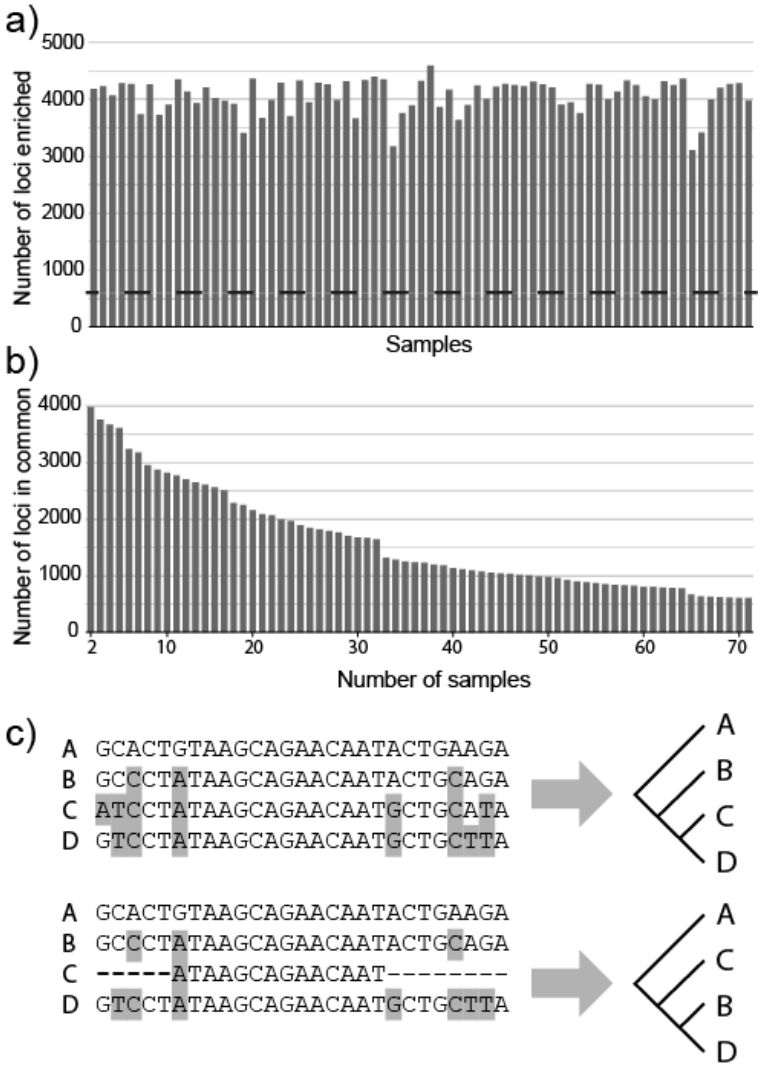


Figure 1.1. Missing data in sequence capture. a) Number of ultraconserved element (UCE) loci obtained for 71 samples of Zosteropidae using a sequence capture technique, with a few samples yielding substantially fewer loci. Dashed line indicates number of loci common to all samples. b) The number of loci common to a subset of samples decreases as the size of the subset increases. The order samples were added to the subset follows that in (a) from left to right. c) Missing data at flanks of alignment can mislead gene tree estimation. Top panel shows alignment with full data and correct gene tree estimate. Informative sites occur towards the flanks of the alignment as in UCE loci obtained from sequence capture. Bottom panel shows same alignment with one sequence truncated causing it to lose informative sites and resulting in the incorrect gene tree.

The Family Zosteropidae

The avian family Zosteropidae (white-eyes) consists of 121 recognized species in 12 genera (Dickinson and Christidis 2014; Alstrom et al. 2015). Most species are endemic to islands or archipelagos of Southeast Asia and Oceania, but the entire distribution spans a vast area in the Old World, from the eastern Atlantic to the western Pacific. A series of molecular studies clarified the species composition of the family (Cibois 2003; Gelang et al. 2009; Moyle et al. 2009, 2012; Alstrom et al. 2015) and it is now understood that members of *Yuhina*, which occur in mainland Asia and Borneo, form the earliest branches in Zosteropidae (Fig. 1.2). The oldest divergence in the family was estimated to have occurred in the late Miocene (Moyle et al. 2009). Next to branch off are "old" zosteropids of the genera *Zosterornis, Cleptornis, Dasycrotapha, Sterrhoptilus,* and *Lophozosterops*, a group whose distribution is concentrated in insular Southeast Asia. Finally, the largest clade in the family comprises the rapid, widespread, and species-rich radiation of white-eyes that consists of the genus *Zosterops*, but also includes the monotypic genus *Chlorocharis*. This enormous radiation involves ~80 species that began diversifying only very recently in the early Pleistocene, yielding one of the highest speciation rates known among terrestrial vertebrates (Moyle et al. 2009; Jetz et al. 2012). Phylogenetic relationships in the family remain largely unresolved owing to short intervals between speciation events in multiple sections of the phylogeny, especially within the rapid radiation of *Zosterops*. A high level of gene tree discordance is therefore expected in this radiation. Thus, the family provides an ideal system to evaluate GCM analytical strategies with empirical data.

METHODS

Sampling

We sampled 70 individuals from 65 of 121 recognized species belonging to 9 of 12 genera in the family Zosteropidae (Dickinson and Christidis 2014; Alstrom et al. 2015): one individual from 61 species and 2–3 individuals from 4 species (Table 1.1). Three samples of *Z. lateralis* and two samples of *Z. luteus* were included because of paraphyly in their mitochondrial ND2 gene trees (Nyári and Joseph 2013). Two subspecies of *Z. palpebrosus* were sampled because we had access to material from two very disparate geographic regions (the Lesser Sundas and Vietnam). Two individuals of *Z. kulambangrae* were used as control samples that should be sister to each other in analyses. The monotypic genera *Tephrozosterops* and *Apalopteron,* and the Caroline Islands-endemic genus *Rukia* were not sampled. The genus *Madanga,* which has long been classified in Zosteropidae, was recently found to belong to the family Motacillidae (Alstrom et al. 2015). A sample of *Timalia pileata* (Timaliidae), a member of the sister clade to Zosteropidae (Moyle et al. 2012), was used as an outgroup.

Table 1.1. Sampling information

Taxon	Accession Number	Locality	Number of UCE loci enriched	Mean contig length	Mean coverage
Chlorocharis emiliae	KU 17802	Borneo, Malaysia	3981	765.6	26.3
Cleptornis marchei	KU 22576	Saipan, Micronesia	4280	895.2	39.6
Dasycrotapha platen	KU 19056	Mindanao, Philippines	4132	701.6	33.0
Dasycrotapha pygmaea	AMNH 708397	Samar, Philippines	3999	306.9	54.6
Lophozosterops goodfellowi	KU 28427	Mindanao, Philippines	4270	847.2	37.6
Lophozosterops squamiceps	AMNH DOT12549	Sulawesi, Indonesia	4205	934.9	34.3
Lophozosterops squamifrons	LSUMNS B51197	Borneo	3110	700.4	28.3
Lophozosterops superciliaris	WAM 23291	Flores, Indonesia	3997	834.0	33.1
Lophozosterops wallacei	WAM 22903	Sumba, Indonesia	4186	930.4	34.5

Megazosterops palauensis	KU 23671	Palau	3417	662.7	17.2
Sterrhoptilus capitalis	KU 28326	Mindanao, Philippines	4247	929.7	38.1
Sterrhoptilus dennistouni	KU 20225	Luzon, Philippines	4317	929.4	43.1
Sterrhoptilus nigrocapitatus	KU 18034	Luzon, Philippines	4253	900.4	38.2
Timalia pileata	KU 23375	Vietnam	4268		
Yuhina brunneiceps	AMNH DOT5230	Taiwan	3904	772.3	29.1
Yuhina castaniceps	KU 13784	China	4208	892.9	34.9
Yuhina diademata	KU 11118	China	4263	776.4	38.9
Yuhina everetti	KU 17756	Borneo, Malaysia	4312	890.6	39.1
Yuhina flavicollis	KU 15170	Myanmar	4227	782.0	36.1
Yuhina gularis	KU 15173	Myanmar	4253	887.1	37.7
Yuhina nigrimenta	KU 9997	China	4271	958.0	35.0
Yuhina occipitalis	KU 15177	Myanmar	4225	825.5	33.4
Zosterops atricapilla	KU 17735	Borneo, Malaysia	3901	641.6	26.8
Zosterops atrifrons	AMNH DOT12620	Sulawesi, Indonesia	4243	866.5	39.1
Zosterops capensis	FMNH 390165	South Africa	4230	929.2	34.8
Zosterops chloris	AMNH DOT12558	Sulawesi, Indonesia	3636	700.9	23.4
Zosterops cinereus	KU 23678	Palau	3756	745.1	21.9
Zosterops citrinella	WAM 23542	Roti, Indonesia	4172	749.8	33.8
Zosterops conspicillatus	KU 22583	Saipan, Micronesia	3864	655.3	24.0
Zosterops erythropleurus	KU 28088	Vietnam	4593	847.8	107.9
Zosterops everetti	KU 13949	Camiguin Sur, Philippines	4326	908.8	40.1
Zosterops explorator	KU 24401	Fiji	3893	803.8	29.0
Zosterops flavifrons	LSUMNS B45805	Vanuatu	3171	707.9	30.8
Zosterops fuscicapilla	NMNH 2003-062	Louisiade Island, PNG	4350	1040.5	56.7
Zosterops griseotinctus	NMNH 2003-067	Louisiade Island, PNG	4401	991.9	47.4
Zosterops hypoxanthus	KU 27718	New Ireland, PNG	4338	937.1	42.4
Zosterops japonicus	KU 28142	Vietnam	3668		
Zosterops kulambangrae 1	UWBM 76278	Kohingo, Solomon Islands	4314	967.1	45.2
Zosterops kulambangrae 2	KU 15931	Kohingo, Solomon Islands	3935	689.3	26.4
Zosterops lacertosus	KU 19413	Santa Cruz, Solomon Islands	3759	697.7	23.2
Zosterops lateralis chloronotus	KU 6094	SW Australia	4265	918.0	40.9
Zosterops lateralis flaviceps	KU 22568	Fiji	3988	777.1	26.1
Zosterops lateralis lateralis	KU 14863	New Zealand	3949	476.8	34.0

Zosterops luteirostris	AMNH DOT113	Ghizo, Solomon Islands	4288	902.3	46.9
Zosterops luteus balstoni	KU 8904	Western Australia	4332	970.3	47.9
Zosterops luteus luteus	KU 22720	Northern Territory, Australia	3704	686.4	23.4
Zosterops maderaspatanus	FMNH 345980	Madagascar	4287	927.5	39.3
Zosterops melanocephalus	AMNH 461540	Cameroon	4012	300.5	76.1
Zosterops metcalfii	UWBM 63177	Choiseul, Solomon Islands	3988	785.0	29.1
Zosterops meyeni	KU 17853	Batan, Philippines	3671	664.7	21.2
Zosterops montanus	KU 20891	Negros, Philippines	4364	870.6	43.0
Zosterops murphyi	AMNH DOT193	Kolombangara, Solomon Islands	3407	627.6	20.7
Zosterops nigrorum	KU 12519	Camiguin Norte, Philippines	3920	721.5	22.5
Zosterops novaeguineae	KU 12068	Papua New Guinea	3973	798.7	23.9
Zosterops oleaginous	LSUMNS B48626	Yap, Micronesia	4367	414.3	93.4
Zosterops palpebrosus siamensis	KU 23522	Vietnam	4022	771.8	25.9
Zosterops palpebrosus unicus	WAM 23218	Flores, Indonesia	4210	886.7	38.8
Zosterops rendovae	UWBM 76356	Tetepare, Solomon Islands	4269	965.6	38.7
Zosterops rennellianus	UWBM 69808	Rennel, Solomon Islands	4132	834.3	30.7
Zosterops santaecrucis	KU 19412	Santa Cruz, Solomon Islands	4355	940.0	43.8
Zosterops semperi	KU 23684	Palau	3906	778.3	25.1
Zosterops senegalensis	KU 19969	Sierra Leone	3724	687.0	23.4
Zosterops splendidus	AMNH DOT171	Rannonga, Solomon Islands	4264	903.9	35.6
Zosterops stresemanni	KU 19428	Malaita, Solomon Islands	3738	701.0	23.4
Zosterops superciliosus	UWBM 58818	Rennel, Solomon Islands	3945	724.3	25.8
Zosterops ugiensis	KU 12803	Makira, Solomon Islands	4285	912.7	38.2
Zosterops vellalavella	AMNH DOT166	Vella Lavella, Solomon Islands	4066	911.7	35.3
Zosterornis hypogrammicus	CMC 37765	Palawan, Philippines	4005	903.3	31.2
Zosterornis latistriatus	CMC 34221	Panay, Philippines	4057	922.5	34.6
Zosterornis nigrorum	KU 27209	Negros, Philippines	4329	898.5	44.4
Zosterornis whiteheadi	KU 18001	Luzon, Philippines	4256	919.9	36.9

Note: Abbreviations: AMNH (American Museum of Natural History), CMC (Cincinnati Museum Center), FMNH (Field Museum of Natural History), KU (University of Kansas Natural History Museum), LSUMNS (Louisiana State University Museum of Natural Science), NMNH (National Museum of Natural History), UWBM (University of Washington Burke Museum), WAM (Western Australia Museum).

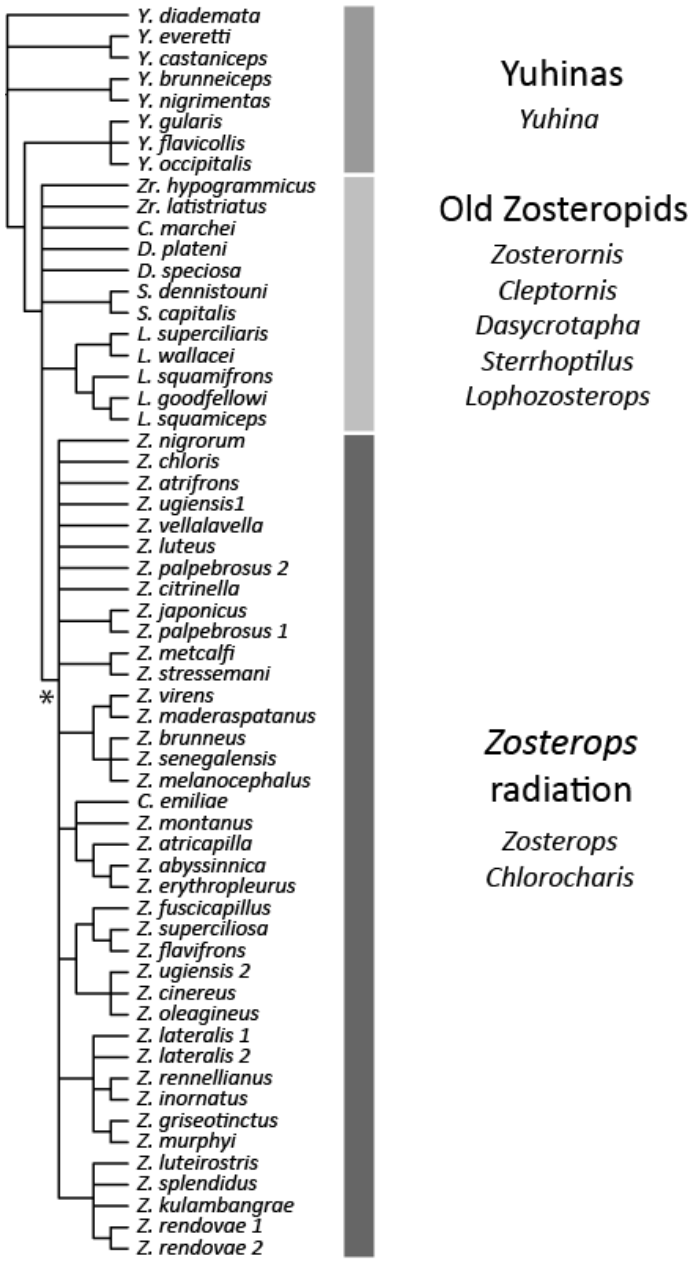


Figure 1.2. Estimate of phylogenetic relationships in Zosteropidae from the study of Moyle et al. (2009). Only well-supported nodes are shown. (*) *Zosterops* radiation estimated to have initiated diversification 1.65 Ma.

Laboratory Techniques

We extracted and purified DNA from fresh muscle or liver tissue or toe pad clips from museum specimens using the Qiagen DNeasy Blood and Tissue Kit following the manufacturer's

protocol. DNA extracts were quantified using a Qubit 2.0 Fluorometer, and 500 ng of DNA of each sample was sheared in 50 µl volume using a Covaris S220 sonicator at 175 W peak incident power, 2% duty factor, and 200 cycles per burst for 45 seconds. We performed end repair, A-tailing, adapter ligation, and amplification on sheared DNA using Kapa Biosystems Library Prep kits following the procedure of Faircloth et al. (2012) and described in detail at http://ultraconserved.org. We deviated from the above protocol in three ways: we used ¼ volume of reagents called for in the library prep kit (N. White, in litt.), we ligated universal iTru stubs instead of sample-specific adapters to allow for dual indexing, and we added a second AMPure XP bead clean up at 1.0x volume after stub ligation. Dual-indexed iTru adapters were added to DNA fragments through a 17-cycle PCR using NEB Phusion High-Fidelity PCR Master Mix. iTru stubs and adapters were developed and designed by T. Glenn et al. (unpublished data).

Libraries were quantified using a Qubit 2.0 Fluorometer and subsequently combined in pools of 8 equimolar samples for enrichment. We performed sequence capture and post-enrichment amplification following standard protocols (Faircloth et al. 2012) using the Mycroarray MYbaits kit for Tetrapods UCE 5K version 1, which targets 5,060 UCE loci. Briefly, biotinylated RNA probes were hybridized with pooled libraries for 24 hrs. Targeted DNA fragments were then recovered with the use of streptavidin-coated beads. These fragments were then amplified in a 17-cycle PCR amplification step using NEB Phusion High-Fidelity PCR Master Mix. After post-enrichment amplification, libraries were quantified using an Illumina Eco qPCR System, then sequenced in a high output, paired-end run of 100 cycles on an Illumina HiSeq 2500 System at the University of Kansas Genome Sequencing Core. Including samples for other projects, 96 individuals were multiplexed in a single Illumina lane.

Data Assembly

Raw reads were de-multiplexed using CASAVA ver. 1.8.2. Low-quality bases and adapter sequences were trimmed from reads using illumiprocessor ver. 1 (https://github.com/faircloth-lab/illumiprocessor). Subsequent data processing was performed using the python package phyluce (Faircloth 2014) and outlined below. Cleaned reads were assembled into contigs using the program Trinity (Grabherr et al. 2011). Contigs matching UCE loci were extracted for each taxon. An incomplete dataset containing UCE loci that were present in at least 75% of all 71 taxa was assembled for maximum likelihood (ML) analysis. For species tree analyses, complete datasets consisting of loci common to all taxa being analyzed were assembled (see below). For each dataset, each locus was aligned using MAFFT (Katoh and Standley 2013), allowing missing nucleotides at the flanks of the alignment only if at least 65% of taxa contained data, which is the default option in phyluce. These alignments were trimmed using Gblocks (Castresana 2000) with default parameters except for the minimum number of sequences for a flank position in Gblocks, which we set at 65% of taxa. The alignments were formatted to phylip files for phylogenetic analysis.

Phylogenetic Analyses

We performed maximum likelihood (ML) inference on the concatenated loci of the incomplete dataset using RAxML ver. 8.1.3 (Stamatakis 2014) assuming a general time reversible model of rate substitution and gamma-distributed rates among sites. Node support was evaluated using 500 rapid bootstraps.

GCM Analysis Strategies.—Three general strategies were employed to increase resolution and minimize highly supported conflicting results in GCM analyses. First, datasets with subsets of taxa were compiled to increase the number of loci available for GCM analysis.

Subsampling taxa has been shown to yield consistent results with GCM analysis (Song et al. 2012). Second, we used only the most informative loci to increase bootstrap support values in GCM species tree estimates. Lastly, we trimmed alignments in order to eliminate biases introduced by short sequence lengths of some samples. An outline of the series of analyses performed is presented in Figure 1.3 and described below.

An initial GCM analysis was performed with all 71 samples (Iteration 0). We then removed the 4 samples with the lowest average contig length (*Dasycrotapha pygmaea, Zosterops oleginea, Z. lateralis lateralis,* and *Z. melanocephalus*) to assess their potential to bias results and performed another set of species tree analyses with the remaining 67 samples (Iteration 1). From the results of Iteration 1, we looked for well-supported clades or groups across the four summary methods that contained many unresolved nodes and performed another set of species tree analyses on these groups. A well-supported clade containing all 43 samples from the genera *Chlorocharis* and *Zosterops* was recovered in Iteration 1, so the 67 samples were divided into two groups, the *Chlorocharis* + *Zosterops* clade (Iteration 2b) and all other Zosteropid genera (Iteration 2a). Each group was analyzed with a few samples from the other group either as outgroup or representatives for clades. The results from Iteration 2a yielded highly resolved trees and thus this group was not analyzed further. The analyses from Iteration 2b yielded a substantial number of polytomies, so the *Chlorocharis* + *Zosterops* clade was further subdivided in two and analyzed separately (Iterations 3a and 3b).

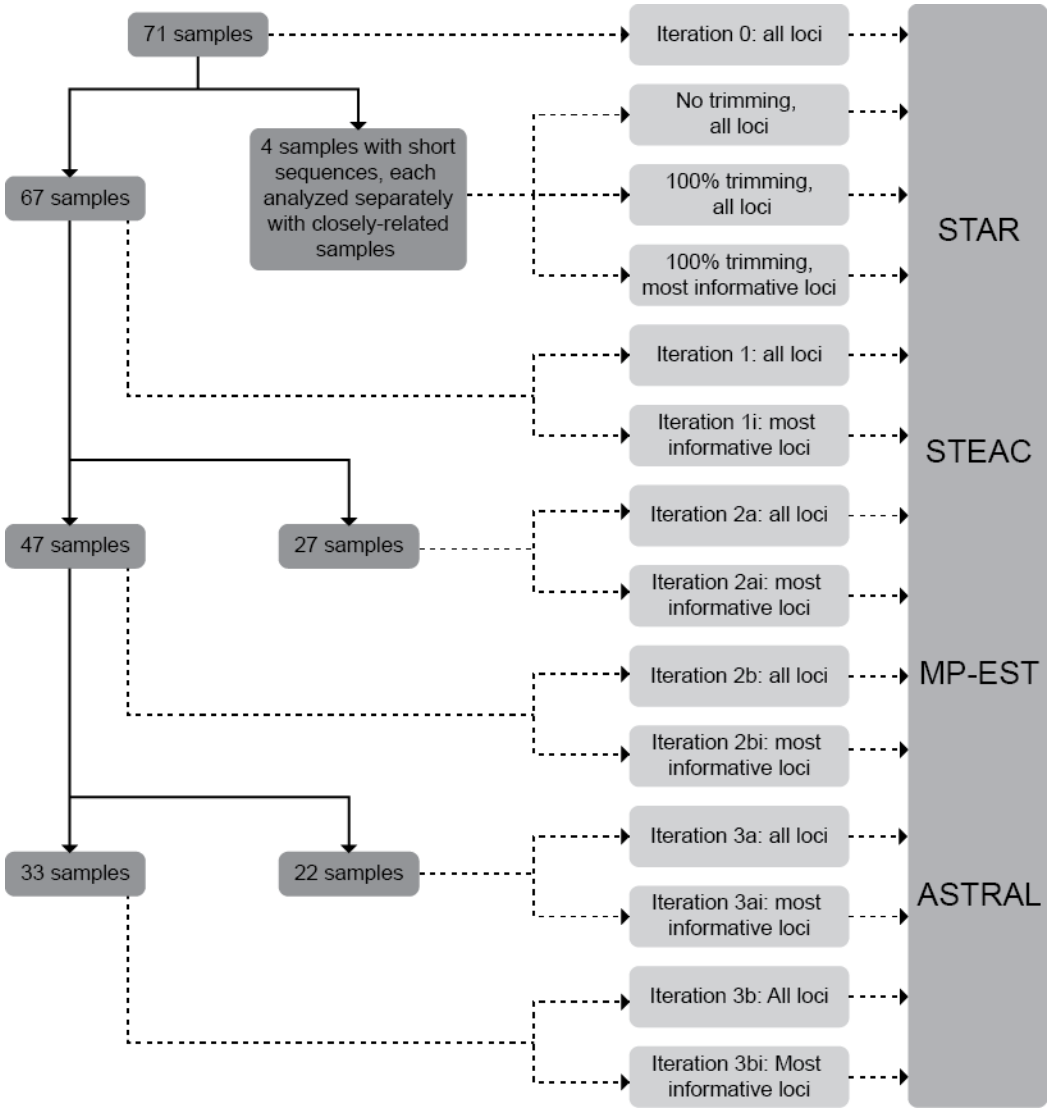


Figure 1.3. Analysis approach with gene tree-based coalescent methods (GCM). An initial analysis with all samples was performed after which four samples with short sequence lengths were analyzed separately. The remaining 67 samples were subsampled five times. Trimming was applied to datasets that contained the samples with short sequence lengths. Datasets were analyzed with all loci common to all samples present in the dataset, and with only the most informative loci. Dark gray boxes denote subsets of samples, light gray boxes indicate datasets and their treatments, if any. All datasets were analyzed using four GCMs.

In order to determine the effect of removing uninformative loci on species tree analysis, we repeated all analyses in Iterations 1–3 but discarded uninformative loci. Liu et al. (2015) suggested excluding weak loci (e.g., < 50% average bootstrap support) from species tree analysis. The maximum average bootstrap support was ~50% in some of our datasets, thus we

discarded loci whose number of parsimony-informative sites or average bootstrap support was below their respective means for all the loci in the dataset, in effect, using only the "most informative" loci.

The phylogenetic placement of the four samples that had the shortest average contig lengths was estimated by performing separate species tree analyses on smaller datasets consisting of the sample of interest along with other samples close to its expected position. This expected position was based on results of a previous study (Moyle et al. 2009), the species tree estimates in Iteration 0, and from the ML-estimated tree. A series of analyses was performed on these smaller datasets. First, species tree analyses were carried out using all loci common to the taxa in the dataset. Second, each locus was trimmed using Gblocks (Castresana 2000) so that no missing nucleotides were permitted at the flanks of each alignment (subsequently referred to as "100% trimming"). Lastly, uninformative loci were discarded similarly as described above. Species trees were estimated on the trimmed dataset before and after uninformative loci were discarded. Because the placement of *Z. melanocephalus* was still uncertain after performing the steps above, the dataset for this species was alternatively trimmed using a custom python script by removing from each alignment only flanking sites where the sequence of *Z. melanocephalus* had missing data. This alterative trimming truncated each locus to the alignment length of the *Z. melanocephalus* sequence but allowed other samples to have missing data at their flanks.

The final species tree estimate for each GCM was summarized by including only well-supported nodes (bootstrap proportion > 70%) that appear in at least one of the results of Iterations 1–3 that used only the most informative loci. The four samples with short sequence lengths were placed based on results of the trimmed datasets after uninformative loci were discarded.

For GCM analysis, gene tree inference and bootstrapping were performed with RAxML ver. 8.1.3 (Stamatakis 2014) using the python package phyluce (Faircloth 2014). We modified the phyluce scripts to implement multi-locus bootstrapping (i.e., sampling with replacement of loci and sites, Seo 2008) and generated 500 multi-locus bootstrap replicate sets of gene trees for each dataset. On each replicate set of gene trees, we ran four GCMs: STAR and STEAC as implemented in the R package phybase ver. 1.3 (Liu et al. 2009b), MP-EST ver. 1.4 (Liu et al. 2010), and ASTRAL ver. 4.7.7 (Mirarab et al. 2014c). All programs were run with the default options. Multi-locus bootstrapping may underestimate support for some correct nodes in point estimates of GCMs (Mirarab et al. 2014b). Because we were interested only on nodes with high support, for each method species trees inferred from the bootstrap replicates were summarized with 70% consensus trees using the sumtrees.py program in Dendropy (Sukumaran and Holder 2010). Command line, python, and R scripts used to process the data and run species tree analyses are available at https://github.com/carloliveros/uce-scripts.

RESULTS

We enriched an average of 4069.8 UCE loci per sample with an average contig length of 789.9 bp and average coverage of 35.7 X (Table 1.1). The incomplete dataset used for ML analysis included data from 3934 loci with a mean locus length of 765.9 bp. Nucleotide data were present in 88.7% of the data matrix. Alignment statistics for the various GCM analyses are presented in Table 1.2. Only 605 UCE loci were common to all 71 samples but in a subset of 33 samples, this more than doubled to 1297 loci and in subsets of 10–12 samples, this number increased 5-fold to ~3000 loci. After applying our filter for informative loci, datasets retained 32–45% of their loci in Iterations 1–3 but only 14–26% of their loci in analyses placing samples

with short sequence lengths. Average locus lengths ranged from 873–1024 bp and the percentage of missing nucleotides in alignments were within 5–10% with no trimming applied. Trimming missing data in alignment flanks considerably reduced the average locus length to 35–51% of their original length as well the average number of parsimony-informative sites to 12–31% of their original number.

Table 1.2. Dataset characteristics used in GCM analyses

Dataset Description	Number of Taxa	Number of Loci	Mean locus length (bp)	% Missing Nucleotides	Average Number of Parsimony-Informative Sites/Locus	Average Bootstrap Support/ Locus
Iteration 0, all loci	71	605	890.7	8%	42.7	26.2%
Iteration 1, all loci	67	654	902.9	6%	42.8	28.1%
Iteration 1i, most informative loci	67	273	986.4	5%	64.2	35.1%
Iteration 2a, all loci	47	1042	873.4	6%	16.8	21.1%
Iteration 2ai, most informative loci	47	377	964.2	5%	27.9	26.7%
Iteration 2b, all loci	27	1703	942.5	5%	35.0	45.7%
Iteration 2bi, most informative loci	27	762	1023.7	5%	53.7	56.6%
Iteration 3a, all loci	22	2086	867.0	7%	12.0	30.2%
Iteration 3ai, most informative loci	22	699	953.8	6%	21.4	39.3%
Iteration 3b, all loci	33	1297	884.8	7%	9.0	20.6%
Iteration 3bi, most informative loci	33	419	968.6	6%	16.1	26.1%
Placing *Zosterops melanocephalus*						
All loci	11	2941	895.3	10%	6.5	38.1%
100% trimming, all loci	11	2941	316.8	0%	0.8	16.0%
100% trimming, most informative loci	11	564	337.8	0%	3.1	33.2%
Trimming to *Z. melanocephalus* sequence, all loci	11	2941	327.6	0%	0.9	18.2%
Trimming to *Z. melanocephalus* sequence, most informative loci	11	487	341.6	0%	3.4	37.1%
Placing *Dasycrotapha pygmaea*						
All loci	12	3083	919.4	9%	13.7	49.3%
100% trimming, all loci	12	3083	325.7	0%	2.1	22.6%
100% trimming, most informative loci	12	797	350.2	0%	5.4	42.2%
Placing *Z. oleagineus*						
All loci	10	2986	898.8	9%	4.3	40.0%
100% trimming, all loci	10	2986	434.1	0%	1.2	22.1%
100% trimming, most informative loci	10	429	489.8	0%	4.9	46.9%

Placing *Z. lateralis lateralis*						
All loci	13	2742	964.2	8%	5.5	35.0%
100% trimming, all loci	13	2741	487.3	0%	1.7	17.9%
100% trimming, most informative loci	13	578	592.8	0%	5.3	35.8%

Maximum Likelihood

ML analysis yielded a well-supported phylogeny of Zosteropidae with 65 out of 70 (93%) internal nodes garnering >70% bootstrap support (Fig. 1.4). Of the 65 well-supported nodes, 57 had 100% bootstrap support. We recovered seven major clades that formed a completely asymmetric topology with 100% support in all nodes (Fig. 1.4). Each clade is described in increasing ladderized order. The first clade, which was sister to the rest of Zosteropidae, contained a single species, *Yuhina diademata*. The second major clade included two other species of *Yuhina*: *Y. castaniceps* and *Y. everetti*. The third major clade comprised five other species of *Yuhina*, which was sister to a clade with all other Zosteropid genera. Within this third clade, a subclade consisting of *Y. nigrimenta* and *Y. brunneiceps* was recovered sister to a subclade containing *Y. flavicollis*, *Y. occipitalis*, and *Y. gularis*. The fourth major clade in Zosteropidae included the genera *Cleptornis*, *Dasycrotapha*, and *Sterrhoptilus*. The monotypic genus *Cleptornis* is endemic to Saipan, Northern Mariana Islands, whereas *Dasycrotapha* and *Sterrhoptilus*, which were recovered as reciprocally monophyletic sister genera, are Philippine endemics. Another genus endemic to the Philippines, *Zosterornis*, comprised the fifth major clade in the family. The sixth major clade included the monotypic genus *Megazosterops*, which is endemic to Palau, embedded within the genus *Lophozosterops*, whose members are distributed in the Philippines, Wallacea, and the Greater Sundas. This *Megazosterops* + *Lophozosterops* clade was recovered sister to the last and largest major clade, the *Zosterops* radiation.

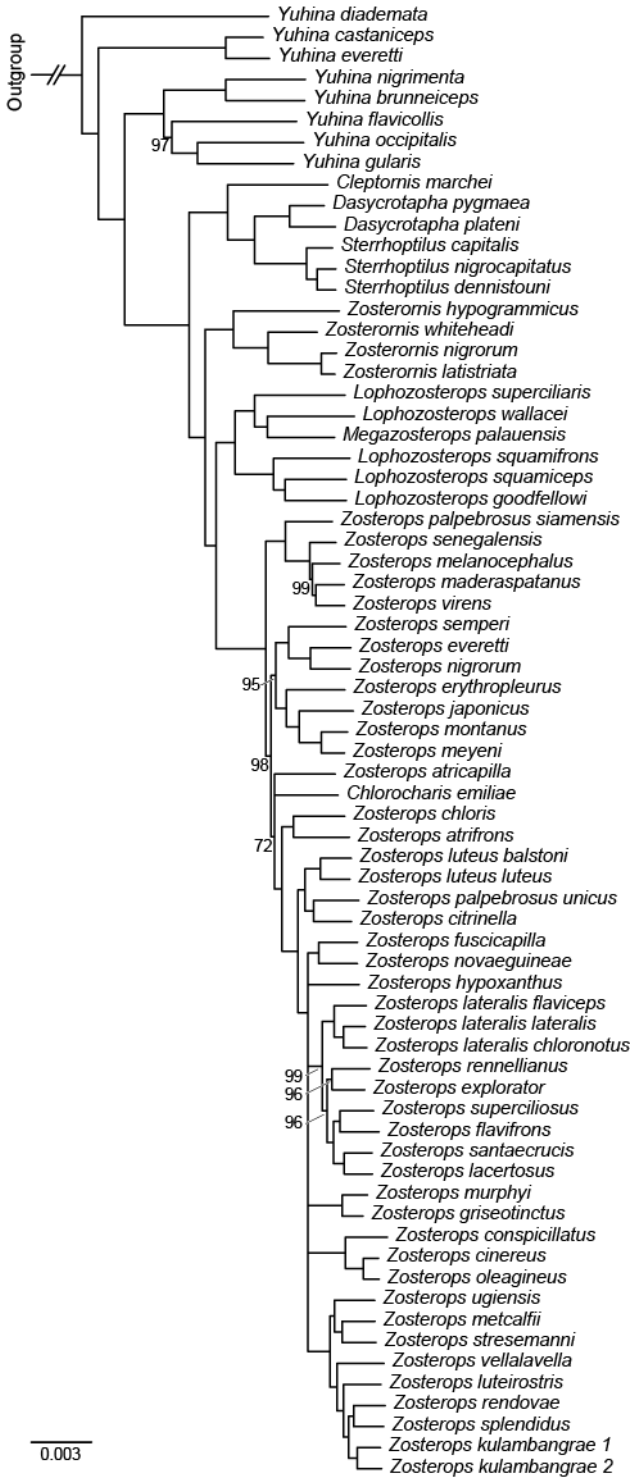


Figure 1.4. Maximum likelihood estimate of phylogenetic relationships in Zosteropidae based on 3934 concatenated UCE loci from 1000 bootstrap replicates. Numbers below branches indicate bootstrap support (BS) values. Nodes with < 70% BS are collapsed whereas nodes with no labels indicate 100% BS.

The *Zosterops* radiation included only one other genus, the monotypic *Chlorocharis*. Within this radiation, a clade consisting of five taxa from continental Asia, Africa, and Madagascar (*Z. palpebrosus siamensis, Z. senegalensis, Z. melanocephalus, Z. maderaspatanus,* and *Z. virens*) was recovered sister to a large clade with all other taxa in the radiation. This large clade was in turn subdivided into two subclades. One subclade contained seven species (*Z. semperi, Z. everetti, Z. nigrorum, Z. erythropleurus, Z. japonicus, Z. montanus*, and *Z. meyeni*) that occur in East Asia, continental Southeast Asia, the Philippines, and Palau. The other subclade was marginally well-supported (BS = 72%) and included three lineages in a polytomy: the Sumatra-Borneo endemic *Z. atricapilla*, the Bornean endemic *Chlorocharis emiliae*, and a large "Pacific" clade of 31 taxa distributed throughout Wallacea, the Australia-New Guinea region, and the Southwest Pacific. Within this Pacific clade, a pair of Wallacean species (*Z. chloris* and *Z. atrifrons*) was sister to a clade containing all other members of the Pacific clade, which in turn consisted of two subclades. One subclade included three species from Northern Australia (*Z. luteus*) and the Lesser Sundas (*Z. palpebrosus unicus* and *Z. citrinella*) and the other subclade contained six lineages from the Australia-New Guinea region and the Southwest Pacific that formed a polytomy. The six lineages consist of: (a) a pair of species from New Guinea and the adjacent Louisiade archipelago (*Z. novaeguineae* and *Z. fuscicapilla*), (b) *Z. hypoxanthus* from the Bismarck archipelago, (c) seven species that occur in the Australia-New Guinea region and the Southwest Pacific (*Z. lateralis, Z. rennellianus, Z. explorator, Z. superciliosus, Z. flavifrons, Z. santaecrucis,* and *Z. lacertosus*), (d) a pair of species from the New Georgia and Louisiade archipelagos (*Z. murphyi* and *Z. griseotinctus*), (e) three species from the West Pacific Islands (*Z. conspicillatus, Z. cinereus,* and *Z. oleagineus*), and (f) eight

species endemic to the Solomon Islands (*Z. ugiensis, Z. metcalfi, Z. stresemanni, Z. vellalavella, Z. luteirostris, Z. rendovae, Z. splendidus,* and *Z. kulambangrae*).

ML analysis recovered three species that were sampled with more than one individual as monophyletic: *Z. luteus*, *Z. lateralis,* and *Z. kulambangrae*. However, two subspecies of *Z. palpebrosus* were placed in different clades. The subspecies *siamensis,* represented by a sample from Vietnam, was found sister to a clade of African and Malagasy species, whereas the subspecies *unicus*, represented by a sample from the Lesser Sundas, was recovered sister to another Lesser Sundas endemic, *Z. citrinella*.

Gene-tree-based Coalescent Methods

Iteration 0.—Results from analyzing all 605 UCE loci common to all 71 samples exemplify some of the criticisms leveled against GCMs: poor resolution and highly supported conflicting results. Analyses in Iteration 0 yielded species tree estimates with varying levels of resolution across the four GCMs (Figs. 1.5, 1.6a). MP-EST produced the least resolved tree with only 49% of nodes being well-supported; whereas STEAC resulted in a species tree with the highest resolution with 89% of nodes being well-supported. The resolution of the STAR and ASTRAL species trees were comparable (76% and 79% of nodes being well-supported, respectively). As expected, samples with consistently short sequences across loci (*Dasycrotapha pygmaea, Z. melanocephalus, Z. oleagineus*, and *Z. lateralis lateralis*) were placed by the different GCMs in different clades, in many cases with high support (Fig. 1.5). For instance, STAR placed *D. pygmaea* sister to the *Zosterops* radiation with high support. STEAC and ASTRAL, on the other hand, both recovered it sister to the clade with the genera *Cleptornis, Sterrhoptilus,* and the other *Dasycrotapha* species, both with high support. MP-EST, however, was uncertain of its placement among old zosteropid lineages. Another well-supported

relationship that conflicted between GCMs was the placement of *Y. diademata*. STAR's placement of this species as sister to the pair consisting of *Y. castaniceps* and *Y. everetti* conflicted with ASTRAL's placement of this species as sister to all other members of the family. The last well-supported conflicting result between GCMs involved the paraphyly of *Z. kulambangrae* in STAR, MP-EST, and ASTRAL, whereas STEAC recovered the species monophyletic. With the exception of the placement of these six taxa, no other well-supported conflicts were recovered in the estimated species trees of the four GCMs in Iteration 0.

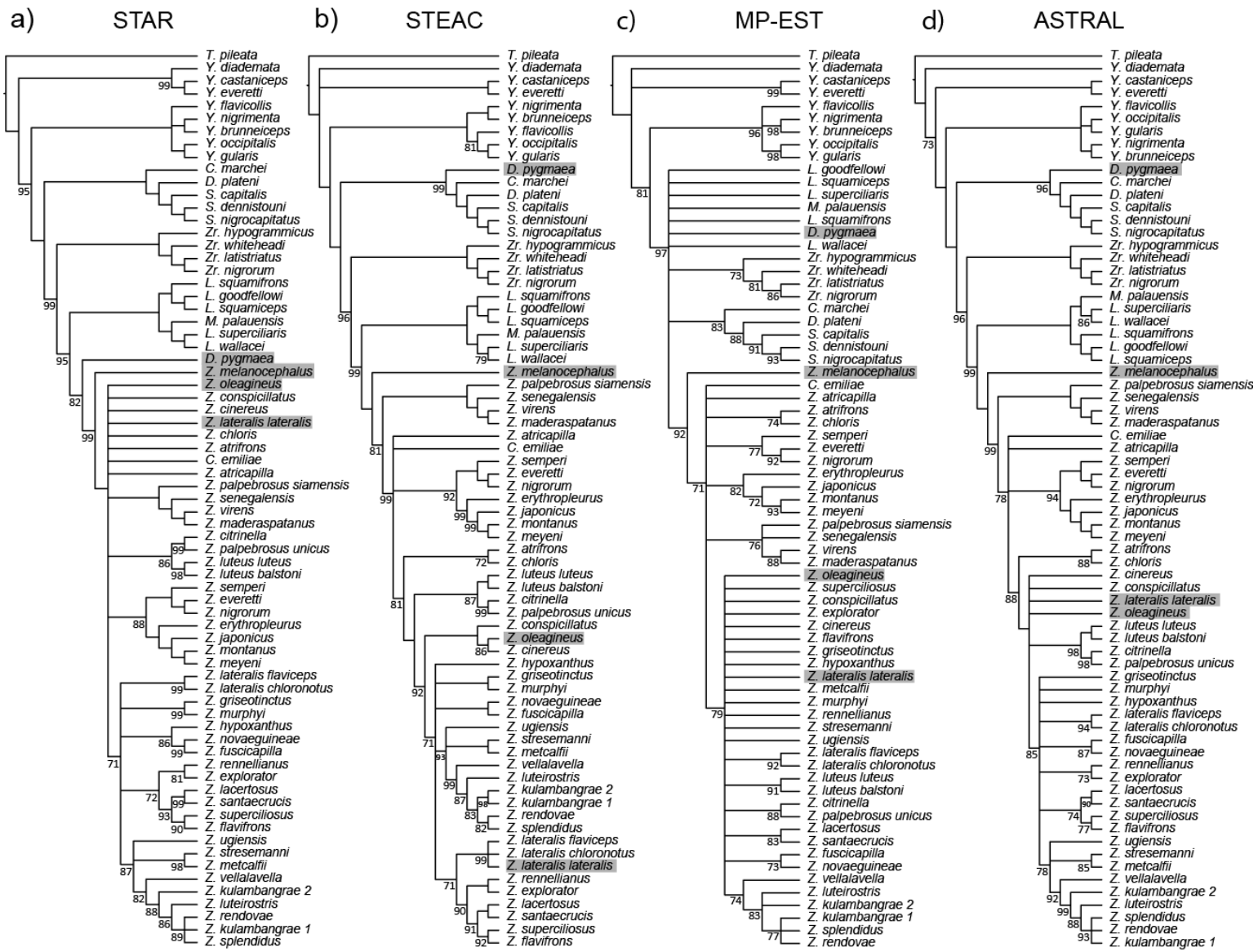

a) STAR
T. pileata
Y. diademata
Y. castaniceps
Y. everetti
Y. flavicollis
Y. nigrimenta
Y. brunneiceps
Y. occipitalis
Y. gularis
C. marchei
D. plateni
S. capitalis
S. dennistouni
S. nigrocapitatus
Zr. hypogrammicus
Zr. whiteheadi
Zr. latistriatus
Zr. nigrorum
L. squamifrons
L. goodfellowi
L. squamiceps
M. palauensis
L. superciliaris
L. wallacei
D. pygmaea
Z. melanocephalus
Z. oleagineus
Z. conspicillatus
Z. cinereus
Z. lateralis lateralis
Z. chloris
Z. atrifrons
C. emiliae
Z. atricapilla
Z. palpebrosus siamensis
Z. senegalensis
Z. virens
Z. maderaspatanus
Z. citrinella
Z. palpebrosus unicus
Z. luteus luteus
Z. luteus balstoni
Z. semperi
Z. everetti
Z. nigrorum
Z. erythropleurus
Z. japonicus
Z. montanus
Z. meyeni
Z. lateralis flaviceps
Z. lateralis chloronotus
Z. griseotinctus
Z. murphyi
Z. hypoxanthus
Z. novaeguineae
Z. fuscicapilla
Z. rennellianus
Z. explorator
Z. lacertosus
Z. santaecrucis
Z. superciliosus
Z. flavifrons
Z. ugiensis
Z. stresemanni
Z. metcalfii
Z. vellalavella
Z. kulambangrae 2
Z. luteirostris
Z. rendovae
Z. kulambangrae 1
Z. splendidus
b) STEAC
T. pileata
Y. diademata
Y. castaniceps
Y. everetti
Y. nigrimenta
Y. brunneiceps
Y. flavicollis
Y. occipitalis
Y. gularis
D. pygmaea
C. marchei
D. plateni
S. capitalis
S. dennistouni
S. nigrocapitatus
Zr. hypogrammicus
Zr. whiteheadi
Zr. latistriatus
Zr. nigrorum
L. squamifrons
L. goodfellowi
L. squamiceps
M. palauensis
L. superciliaris
L. wallacei
Z. melanocephalus
Z. palpebrosus siamensis
Z. senegalensis
Z. virens
Z. maderaspatanus
Z. atricapilla
C. emiliae
Z. semperi
Z. everetti
Z. nigrorum
Z. erythropleurus
Z. japonicus
Z. montanus
Z. meyeni
Z. atrifrons
Z. chloris
Z. luteus luteus
Z. luteus balstoni
Z. citrinella
Z. palpebrosus unicus
Z. conspicillatus
Z. oleagineus
Z. cinereus
Z. hypoxanthus
Z. griseotinctus
Z. murphyi
Z. novaeguineae
Z. fuscicapilla
Z. ugiensis
Z. stresemanni
Z. metcalfii
Z. vellalavella
Z. luteirostris
Z. kulambangrae 2
Z. kulambangrae 1
Z. rendovae
Z. splendidus
Z. lateralis flaviceps
Z. lateralis chloronotus
Z. lateralis lateralis
Z. rennellianus
Z. explorator
Z. lacertosus
Z. santaecrucis
Z. superciliosus
Z. flavifrons
c) MP-EST
T. pileata
Y. diademata
Y. castaniceps
Y. everetti
Y. flavicollis
Y. nigrimenta
Y. brunneiceps
Y. occipitalis
Y. gularis
L. goodfellowi
L. squamiceps
L. superciliaris
M. palauensis
L. squamifrons
D. pygmaea
L. wallacei
Zr. hypogrammicus
Zr. whiteheadi
Zr. latistriatus
Zr. nigrorum
C. marchei
D. plateni
S. capitalis
S. dennistouni
S. nigrocapitatus
Z. melanocephalus
C. emiliae
Z. atricapilla
Z. atrifrons
Z. chloris
Z. semperi
Z. everetti
Z. nigrorum
Z. erythropleurus
Z. japonicus
Z. montanus
Z. meyeni
Z. palpebrosus siamensis
Z. senegalensis
Z. virens
Z. maderaspatanus
Z. oleagineus
Z. superciliosus
Z. conspicillatus
Z. explorator
Z. cinereus
Z. flavifrons
Z. griseotinctus
Z. hypoxanthus
Z. lateralis lateralis
Z. metcalfii
Z. murphyi
Z. rennellianus
Z. stresemanni
Z. ugiensis
Z. lateralis flaviceps
Z. lateralis chloronotus
Z. luteus luteus
Z. luteus balstoni
Z. citrinella
Z. palpebrosus unicus
Z. lacertosus
Z. santaecrucis
Z. fuscicapilla
Z. novaeguineae
Z. vellalavella
Z. luteirostris
Z. kulambangrae 2
Z. kulambangrae 1
Z. splendidus
Z. rendovae
d) ASTRAL
T. pileata
Y. diademata
Y. castaniceps
Y. everetti
Y. flavicollis
Y. occipitalis
Y. gularis
Y. nigrimenta
Y. brunneiceps
D. pygmaea
C. marchei
D. plateni
S. capitalis
S. dennistouni
S. nigrocapitatus
Zr. hypogrammicus
Zr. whiteheadi
Zr. latistriatus
Zr. nigrorum
M. palauensis
L. superciliaris
L. wallacei
L. squamifrons
L. goodfellowi
L. squamiceps
Z. melanocephalus
Z. palpebrosus siamensis
Z. senegalensis
Z. virens
Z. maderaspatanus
C. emiliae
Z. atricapilla
Z. semperi
Z. everetti
Z. nigrorum
Z. erythropleurus
Z. japonicus
Z. montanus
Z. meyeni
Z. atrifrons
Z. chloris
Z. cinereus
Z. conspicillatus
Z. lateralis lateralis
Z. oleagineus
Z. luteus luteus
Z. luteus balstoni
Z. citrinella
Z. palpebrosus unicus
Z. griseotinctus
Z. murphyi
Z. hypoxanthus
Z. lateralis flaviceps
Z. lateralis chloronotus
Z. fuscicapilla
Z. novaeguineae
Z. rennellianus
Z. explorator
Z. lacertosus
Z. santaecrucis
Z. superciliosus
Z. flavifrons
Z. ugiensis
Z. stresemanni
Z. metcalfii
Z. vellalavella
Z. kulambangrae 2
Z. luteirostris
Z. splendidus
Z. rendovae
Z. kulambangrae 1

Figure 1.5. Species tree estimates obtained from analysis of 605 UCE loci common to all 71 samples (Iteration 0) using (a) STAR, (b) STEAC, (c) MP-EST, and (d) ASTRAL. Samples with short sequences are highlighted in gray. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed whereas nodes with no labels indicate 100% BS.

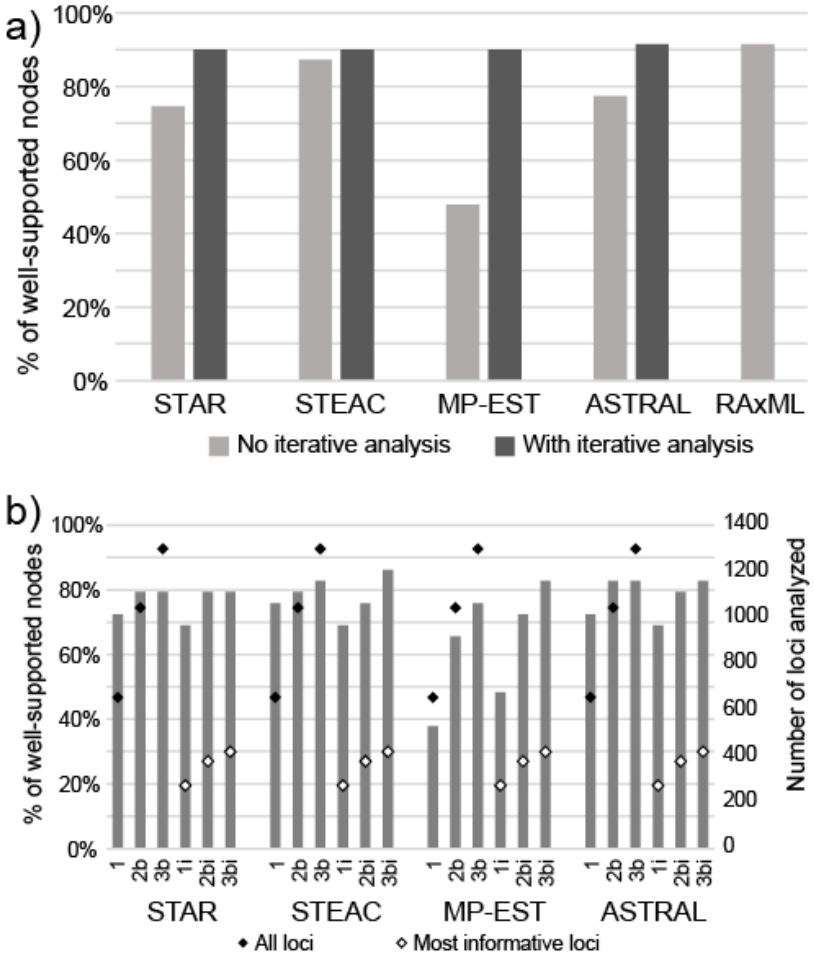


Figure 1.6. a) Resolution in final species tree estimates of Zosteropidae (Fig. 1.7) obtained using four GCMs and ML. b) Resolution in phylogenetic reconstruction of the Pacific clade using four GCMs across six iterations. Bars indicate percentage of well-supported nodes (BS ≥ 70%). Diamonds indicate number of loci analyzed.

Iterations 1, 2, and 3.—Full results of GCM analyses in Iterations 1–3 are provided in the Supplementary Material (App. 1.1–1.8). Figure 1.6b illustrates the effect of the number and informativeness of loci in the resolution of a clade of 29 taxa (30 samples) from the Australia-New Guinea region and Southwest Pacific, which were analyzed in Iterations 1, 2b, 3b, 1i, 2bi, and 3bi. Across all GCMs, the increase in the number of loci from Iteration 1 to 2b resulted in an increase in the proportion of well-supported nodes. The most notable improvement was observed in MP-EST analysis wherein the proportion of well-supported nodes jumped from 38% in Iteration 1 to 66% in Iteration 2b (Figs. 1.6b, 1.7). The further increase in the number of loci from Iteration 2b to 3b brought further increases in the proportion of well-supported nodes in STEAC and MP-EST, but not in STAR and ASTRAL. Analyzing a much lower number but more informative loci (Iterations 1i, 2bi, and 3bi) yielded species trees with slightly less or comparable resolution to those obtained using all loci for STAR, STEAC, and ASTRAL; but a significant improvement was observed for MP-EST (Figs. 1.6b, 1.7). STEAC produced the most highly resolved species tree in 5 out of the 6 iterations presented in this section (Figs. 1.6b, S2, S6).

Across all GCMs and all Iterations 1–3, only three well-supported conflicts were recovered. The first involves the placement of *Y. diademata* as sister to *Y. everetti* and *Y. castaniceps* by STAR in Iterations 1 and 1i, as it was similarly placed in Iteration 0. When it was placed with high support, *Y. diademata* came out sister to the rest of the family in analyses using the other three GCMs (App. 1.1–1.8). The unique placement of *Y. diademata* with STAR appears to be an artefact of the method (see below) and was therefore discounted. The second conflict concerns the position of *M. palauensis*. When its position was estimated with high support, this species was recovered sister to the pair of *L. wallacei* and *L. superciliaris* in all

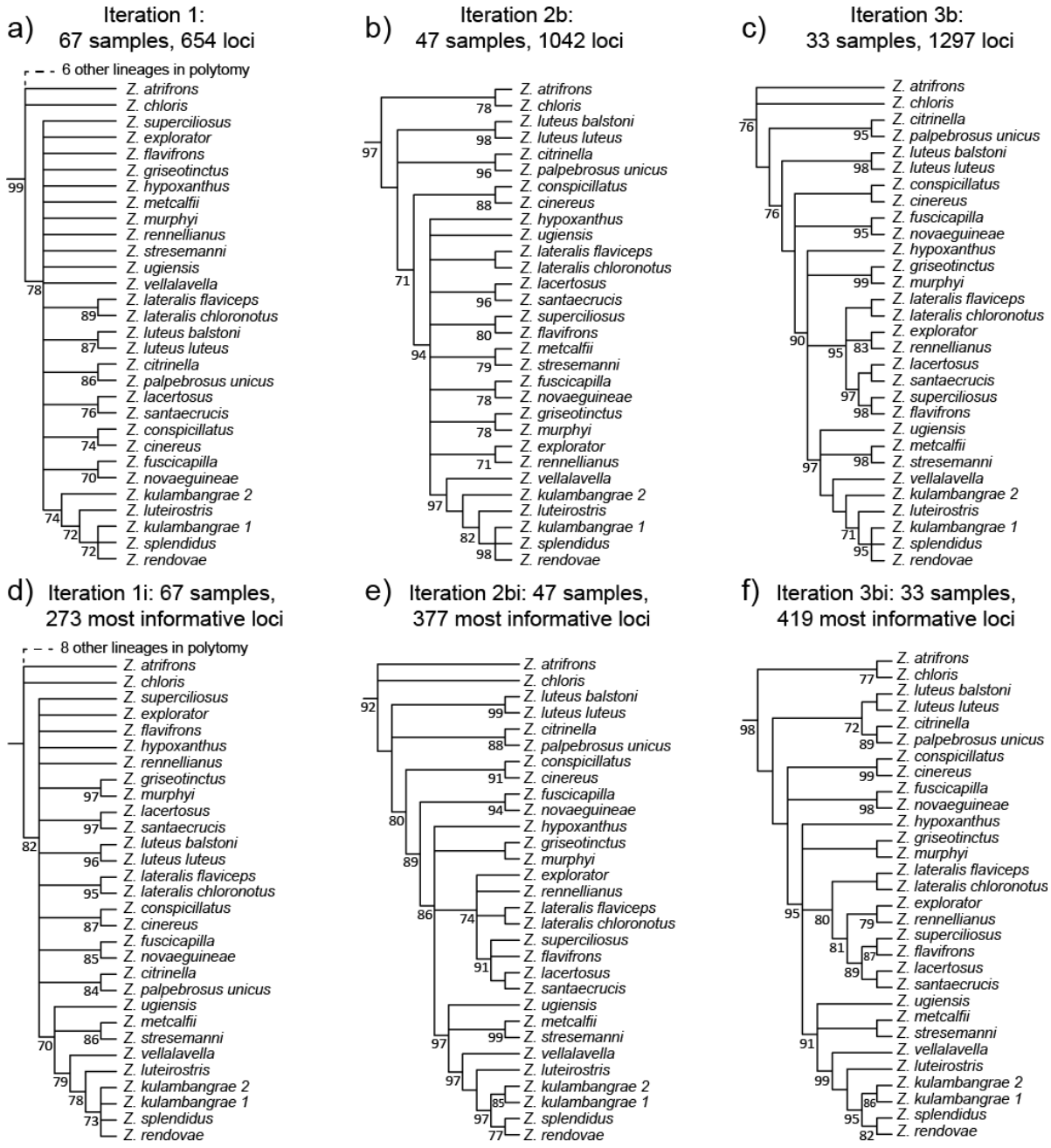


Figure 1.7. Excerpts of phylogenies containing the Pacific clade estimated with MP-EST across six iterations. a) Iteration 1: 67 taxa, 654 loci. b) Iteration 2b: 47 taxa, 1042 loci. c) Iteration 3b: 33 taxa, 1297 loci. d) Iteration 1i: 67 taxa, 273 loci. e) Iteration 2bi: 47 taxa, 377 loci. f) Iteration 3bi: 33 taxa, 419 loci. Full trees are shown in Appendices 1.3 and 1.7. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed whereas nodes with no labels indicate 100% BS.

GCM analyses, with the exception of STEAC analysis in Iterations 1i and 2ai and ASTRAL analysis in Iteration 1i, when it was recovered sister to *L. wallacei* (App. 1.1–1.8). Further examination of sequence lengths in the data suggest that this conflict may have arisen from bias introduced by short sequence lengths of *M. palauensis*. Analyses that focused on the *Megazosterops* + *Lophozosterops* clade and entailed 100% trimming confirmed this bias and *M. palauensis* was recovered sister to *L. wallacei* across all GCMs in this round of analyses (App. 1.13). The third conflict involves the relationships of the two samples of *Z. kulambangrae*. In GCM analyses using all loci (Iterations 1, 2b, 3b) and with the exception of STEAC, *Z. kulambangrae* was found to be paraphyletic (Figs. 1.7a–c, App. 1.1, 1.3–1.4). However, STEAC, as well as all the three other GCMs when only the most informative loci are analyzed (Iterations 1i, 2bi, 3bi), recovered a monophyletic *Z. kulambangrae* (Figs. 1.7d–f, App. 1.2, 1.5–1.8). STEAC uses branch length information from gene trees and is thus resistant to biases introduced by uninformative loci. The paraphyly of *Z. kulambangrae* obtained in the other GCMs appears to be a bias caused by the presence of uninformative loci and was thus ignored.

Placement of Samples with Short Sequences.—The strategy of performing GCM analysis separately for the four samples with short sequences was successful in estimating their phylogenetic position with high support, with the exception of *Z. melanocephalus*. Full results are available in the Supplementary Material (App. 1.9–1.12) but only results of MP-EST are discussed in detail (Fig. 1.8). In analyzing all 3083 loci common to 12 taxa, *D. pygmaea* was placed sister to a clade that includes *D. plateni* and the genus *Sterrhoptilus* with high support (Fig. 1.8a). However, with 100% trimming of the 3083 loci, *D. pygmaea* was recovered sister to *D. plateni* with high support, the expected result because these two Philippine endemics are quite similar and have often been considered a single species. Using the 797 most informative 100%-

trimmed loci yielded the same species tree topology as that obtained using all 3083 trimmed loci. The result obtained from analysis of all 3083 without trimming is likely due to short sequence length bias as discussed above so the result of this analysis was discounted. Analyses using the three sets of loci all yielded a fully resolved species tree.

The phylogenetic position of *Z. oleagineus*, an endemic to the Caroline Islands in the West Pacific, was consistent across analyses, whether all loci were used, all loci were trimmed, or only the most informative trimmed loci were analyzed (Fig. 1.8b). The species was recovered sister to the Palau endemic, *Z. finschii*; *Z. conspicillatus*, a Mariana Island endemic, was found sister to the clade containing *Z. oleagineus* and *Z. finschii*. The species tree resulting from analyzing all 2986 loci had the highest resolution (89% of nodes well-supported), followed by that from using only 429 most informative trimmed loci (78%), and lastly, that from using 2986 trimmed loci (67%).

The subspecies *Z. lateralis lateralis* was placed in a polytomy with the other two *Z. lateralis* samples when all 2742 loci were analyzed (Fig. 1.8c). However, it was recovered sister to *Z. lateralis chloronotus* both when all 2741 trimmed loci were analyzed and when only 578 most informative trimmed loci were used. The resolution of the species tree using all 2741 trimmed loci was higher (67% of nodes well-supported) than those obtained from the analysis of all 2742 loci or 578 of the most informative trimmed loci, which had the same level of resolution (58%).

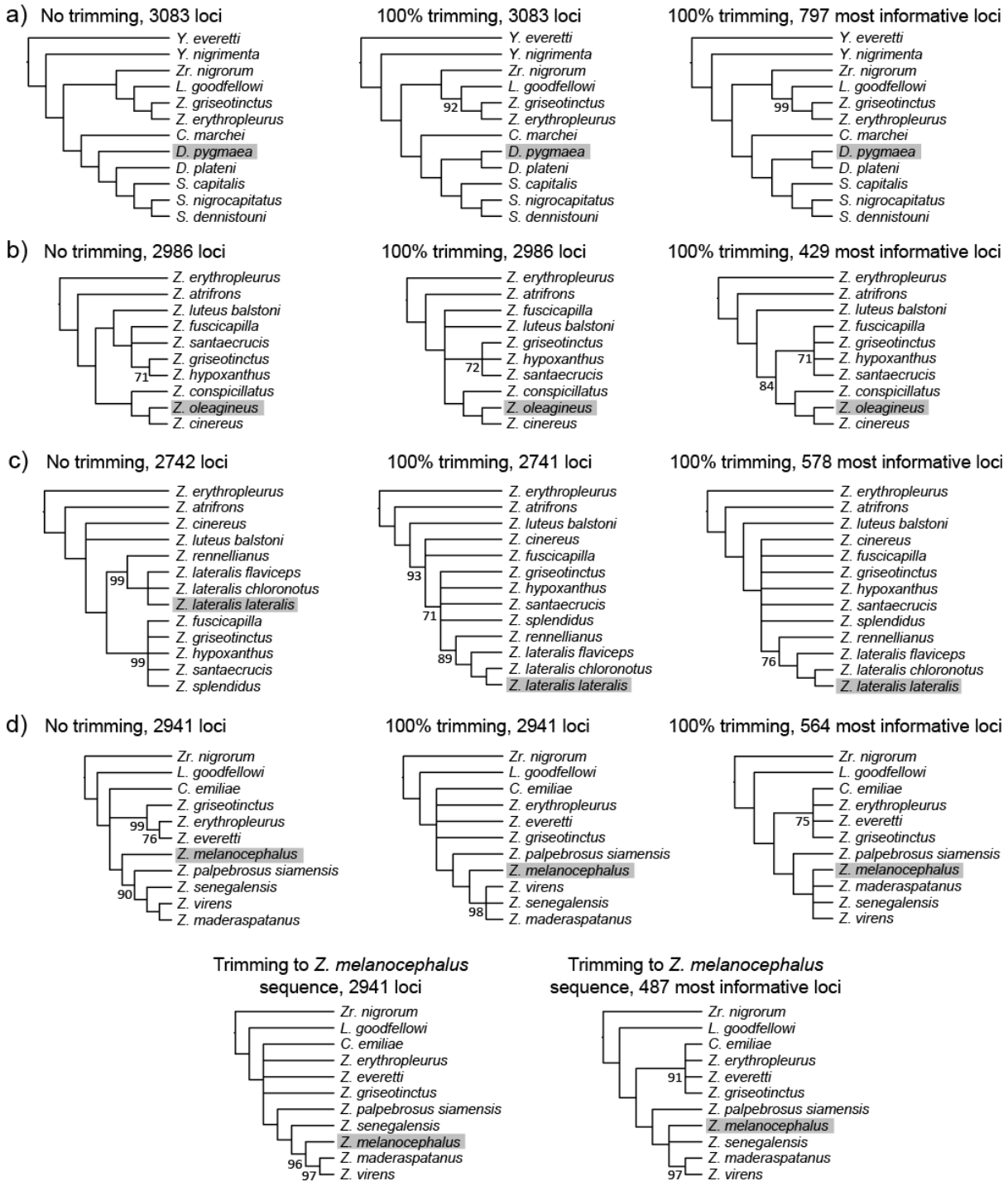


Figure 1.8. Phylogenetic position of samples with short sequences (highlighted in gray) as estimated with MP-EST. a) *Dasycrotapha pygmaea*. b) *Zosterops oleagineus.* c) *Zosterops lateralis lateralis.* d) *Zosterops melanocephalus*. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed whereas nodes with no labels indicate 100% BS.

Analysis of all 2941 loci placed the Cameroon endemic *Z. melanocephalus* sister to a clade that includes *Z. palpebrosus siamensis, Z. senegalensis, Z. virens,* and *Z. madaraspatanus* with high support (Fig. 1.8d). Analysis of the same loci when trimmed 100% recovered *Z. melanocephalus* sister to a clade with *Z. virens, Z. senegalensis,* and *Z. maderaspatanus*; but when only 564 of the most informative 100%-trimmed loci were used, *Z. melanocephalus* was placed in a four-way polytomy with these three species. *Z. virens* and *Z. senegalensis* are African endemics, whereas *Z. maderaspatanus* occurs only on Madagascar. Results from 100% trimming confirm that the result obtained by using all 2941 untrimmed loci was erroneous, caused by short sequence length bias, and should therefore be discounted. Trimming the 2941 loci to the sequence length of *Z. melanocephalus* allowed a gain of 243 (10%) parsimony-informative sites across all loci (vs. 100% trimming) and resulted in the placement of the species as sister to the pair of *Z. maderaspatanus* and *Z. virens*. However, using only 487 of the most informative sequence-length-trimmed loci recovered *Z. melanocephalus* in a three-way polytomy with *Z. senegalensis* and the sister pair of *Z. madaraspatanus* + *Z. virens*. Results of sequence trimming also suggest that either of the results found with 2941 100%-trimmed loci or 2941 sequence-length-trimmed loci could have been biased by uninformative loci, as observed in the placement of *Z. kulambangrae* in Iterations 1, 2b, and 3b above. Therefore, placing *Z. melanocephalus* in a three-way polytomy such as that recovered using 487 of the most informative sequence-length-trimmed loci would be the best estimate of its position. Resolution of the species tree was highest when all 2941 untrimmed loci was used (90%) followed by the two analyses that used sequence-length-trimmed loci, both of which had 70% well-supported nodes; resolution was lowest with the two analyses that used 100% trimming, both of which had

60% well-supported nodes. STAR and ASTRAL recovered the same position for *Z. melanocephalus* as MP-EST but STEAC found it sister to *Z. senegalensis* (App. 1.12).

Highly supported conflicting placement of taxa with short sequences were also observed for STAR and ASTRAL, but not STEAC (App. 1.9–1.12). However, these conflicts were caused by biases introduced by either short sequence length or uninformative loci, and thus, the biased topologies were ignored. The phylogenetic position of taxa with short sequences were taken from results of GCM analysis using the most informative trimmed loci.

GCM Species Trees.—Combining results from analyses using only the most informative loci (i.e., Iterations 1i, 2ai, 2bi, 3ai, and 3bi) and from the analyses placing samples with short sequences, well-supported estimates of species trees showing relationships among all 71 samples were produced for each of the four GCMs (Fig. 1.9). The resolution of these species trees were comparable to those obtained from ML analysis on the concatenated data (Fig. 1.6b). The four GCM species trees were compatible with each other with the exception of one well-supported conflict: STAR's placement of *Y. diademata* as described above.

Comparing the GCM species trees with the ML species tree, only two well-supported conflicts were found. The first involves STAR's erroneous placement of *Y. diademata* as described above. In the other conflict, STEAC recovered *Z. melanocephalus*, one of the samples with short sequences, sister to *Z. senegalensis*; whereas ML placed it sister to the pair of *Z. virens* and *Z. maderaspatanus*. The other GCMs placed this species in a three-way polytomy with *Z. senegalensis* and the pair of *Z. virens* and *Z. maderaspatanus*. Outside this clade, the species tree topologies from ML and GCM analyses were compatible.

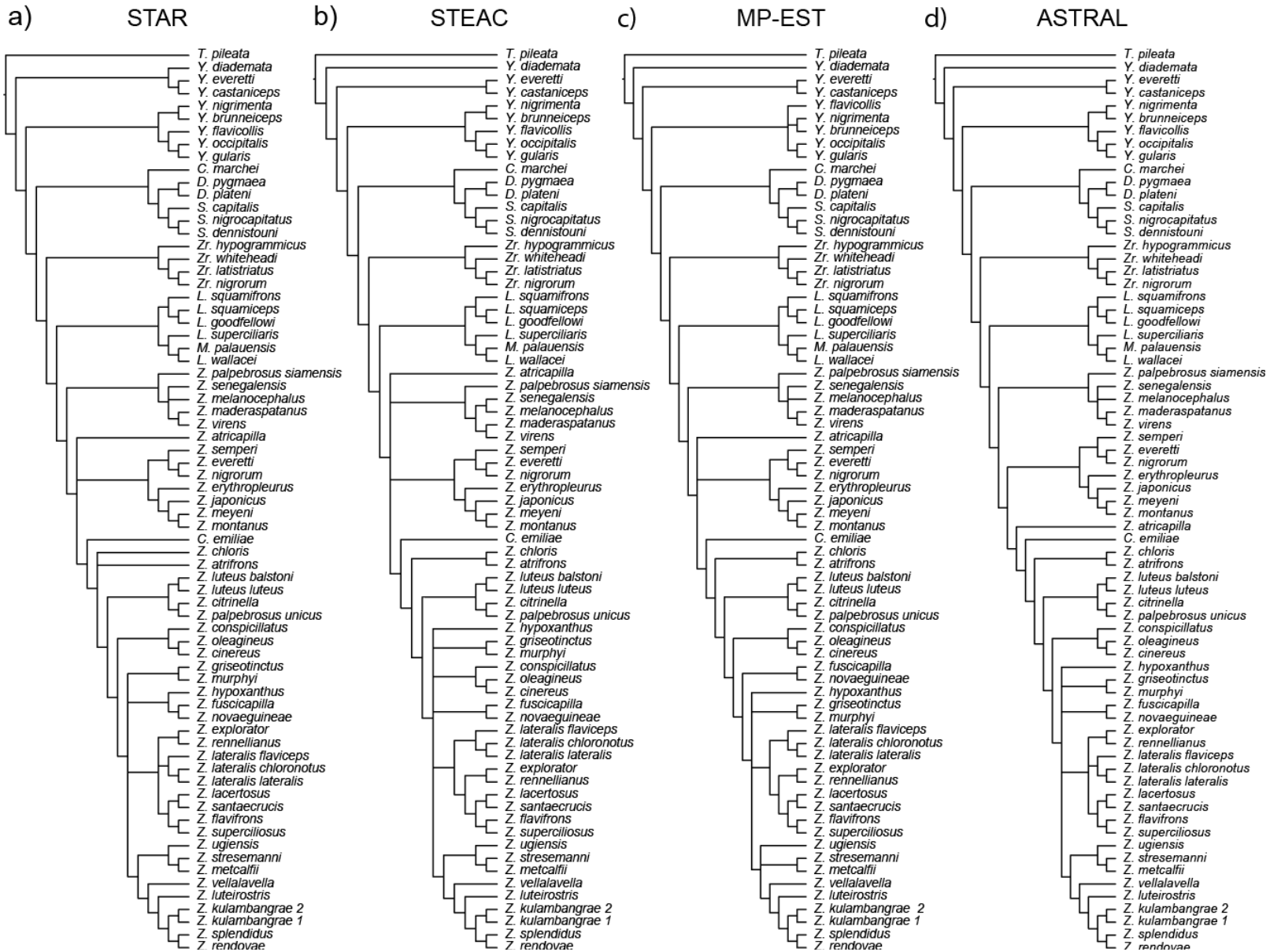

a)
STAR
b)
STEAC
c)
MP-EST
d)
ASTRAL
T. pileata
Y. diademata
Y. everetti
Y. castaniceps
Y. flavicollis
Y. nigrimenta
Y. brunneiceps
Y. occipitalis
Y. gularis
C. marchei
D. pygmaea
D. plateni
S. capitalis
S. nigrocapitatus
S. dennistouni
Zr. hypogrammicus
Zr. whiteheadi
Zr. latistriatus
Zr. nigrorum
L. squamifrons
L. squamiceps
L. goodfellowi
L. superciliaris
M. palauensis
L. wallacei
Z. palpebrosus siamensis
Z. senegalensis
Z. melanocephalus
Z. maderaspatanus
Z. virens
Z. atricapilla
Z. semperi
Z. everetti
Z. nigrorum
Z. erythropleurus
Z. japonicus
Z. meyeni
Z. montanus
C. emiliae
Z. chloris
Z. atrifrons
Z. luteus balstoni
Z. luteus luteus
Z. citrinella
Z. palpebrosus unicus
Z. conspicillatus
Z. oleagineus
Z. cinereus
Z. griseotinctus
Z. murphyi
Z. hypoxanthus
Z. fuscicapilla
Z. novaeguineae
Z. explorator
Z. rennellianus
Z. lateralis flaviceps
Z. lateralis chloronotus
Z. lateralis lateralis
Z. lacertosus
Z. santaecrucis
Z. flavifrons
Z. superciliosus
Z. ugiensis
Z. stresemanni
Z. metcalfii
Z. vellalavella
Z. luteirostris
Z. kulambangrae 2
Z. kulambangrae 1
Z. splendidus
Z. rendovae

Figure 1.9. Final species tree estimates from (a) STAR, (b) STEAC, (c) MP-EST, and (d) ASTRAL constructed from the combined results of datasets that used only the most informative loci. Nodes with < 70% BS are collapsed. BS values for each node vary across datasets and are provided in App. 1.1–1.13.

Four well-supported nodes found using GCM analyses were not recovered using ML analysis. In the first, ML results placed the Sumatra-Borneo endemic *Z. atricapilla*, the Bornean endemic *C. emiliae*, and a Pacific clade of 31 taxa in a three-way polytomy; however, all GCMs recovered *C. emiliae* sister to the Pacific clade (Fig. 1.9). Second, a clade of three West Pacific taxa (*Z. conspicillatus*, *Z. cinereus*, and *Z. oleagineus*) was situated in a six-way polytomy in the ML topology; but STAR, MP-EST, and ASTRAL increased resolution by placing it sister to a clade that contains the five other lineages in this polytomy. Resolving this polytomy further in two different ways comprise the third and fourth nodes recovered by GCMs but not ML. MP-EST placed one member of this polytomy, the pair of *Z. fuscicapilla* and *Z. novaeguinea* from the Louisiade archipelago and New Guinea, respectively, sister to a clade that consists of the four other lineages in the polytomy. On the other hand, STAR found the Bismarck Archipelago endemic, *Z. hypoxanthus*, sister to the *Z. fuscicapilla* and *Z. novaeguinea* clade.

DISCUSSION

Bias from Uninformative Loci

Uninformative loci have been shown to reduce bootstrap support values in GCM analysis (Liu et al. 2015). In this study, we have shown that they can also result in incorrect GCM species tree estimates with high support. However, this bias was observed in the placement of only two samples belonging to one species, *Z. kulambangrae* (Fig. 1.7) and only with STAR, MP-EST, and ASTRAL, but not with STEAC (App. 1.1–1.8). The strategy of analyzing only the most informative loci removed this bias but had varying effects on resolution. Using only the

most informative loci in GCM analysis significantly increased the proportion of well-supported nodes in MP-EST, but decreased slightly or had little effect on the proportion of well-supported nodes in STAR, STEAC, and ASTRAL (Fig. 1.6b). The limited improvement in resolution in some methods after applying this strategy may have been caused by filtering loci too stringently. In removing loci whose number of parsimony-informative sites and average bootstrap support values were less than their respective means, we discarded 55–86% of loci in each dataset, many of which contained some informative sites. Assembling datasets with fewer taxa increases the number of both informative and uninformative loci common to them. This large proportion of discarded loci is not too surprising considering that this study focuses on a fairly recently diverged group and uses relatively slow evolving UCE markers. We expect that in studies of more deeply diverged lineages a smaller proportion of loci will need to be discarded. However, it is remarkable that comparable resolution can be achieved with GCM analysis using only a small fraction of loci that are most informative. Further studies on selecting an optimal filtering scheme to maximize resolution in GCM species tree estimates and minimize bias from uninformative loci are needed.

GCMs achieve greater accuracy and resolution (higher bootstrap support) as the number of loci analyzed increases (Liu et al. 2009b, 2010; Song et al. 2012; Mirarab et al. 2014c) but this and another study (Liu et al. 2015) suggest that these advantages are achieved only with highly informative loci. The number of informative loci available for GCM analysis can be maximized in empirical datasets by choosing phylogenomic markers that evolve fast enough for the level of divergences in the group of interest. In sequence capture techniques, it may be worth targeting a high number of loci in hopes of obtaining more informative ones or increasing sequence capture success rates for each sample. Increasing information content of each loci can also be achieved

by shearing larger fragment sizes in library preparation and using longer sequencing reads (Faircloth et al. 2012).

Bias from Sequence Length Heterogeneity

The results from Iteration 0 demonstrate that species trees estimated by some GCMs can show high support for the wrong topology when using samples with sequence lengths that are consistently shorter than other samples. This bias is a result of gene tree estimation bias rather than a defect of the GCMs themselves. Our analyses in Iterations 1–3 show that GCMs can perform well with some sequence length heterogeneity (see Missing Data column in Table 1.2). Short sequence lengths should not bias GCM species tree estimates as long as: (1) enough phylogenetic signal is present in samples with short sequences so as not to bias gene tree estimation; or (2) samples have short sequences only in a small fraction of loci. In UCE loci, where informative sites occur towards the flanks of loci (Faircloth et al. 2012), the severity of the bias will depend on the depth of divergences between the taxa of interest, the average length of loci, the average length of the short sequences, and the fraction of loci in which consistently short sequences occur. In exon capture techniques where informative sites are present throughout the length of the locus, short sequences are expected to bias GCM species tree estimates to a lesser degree as they do with ultraconserved elements.

Analyzing samples with short sequence lengths separately with a subset of taxa requires prior information about their possible phylogenetic position. As was done in this study, this information can be taken from results of previous studies, the ML tree, and GCM species trees estimated with full sampling. Using a subset of taxa also considerably increases the number of loci available for GCM analysis (by up to 400%). In the case of *Z. oleagineus*, merely increasing the number of loci analyzed masked any bias introduced by short sequence lengths, even without

trimming sequences or removing uninformative loci (Figs. 1.8b, App. 1.10). However, for the other three samples sequence trimming was necessary to converge on consistent relationships. Trimming of UCE loci eliminates many informative sites (Table 1.2), but nonetheless resulted in well-supported relationships in the clades of interest (Figs. 1.8, App. 1.9–1.12). We expected to see biases introduced by uninformative loci in species tree estimates from trimmed datasets, but were surprised to see no well-supported conflicts between results obtained using all trimmed loci and those obtained using only the most informative loci for each of the four species with short sequences across all GCMs.

Bias in STAR analysis

The placement of *Y. diademata* as sister to the clade formed by *Y. everetti* and *Y. castaniceps* with high support by STAR in Iterations 0, 1, and 1i is quite unusual. This position is contradicted by results from the other three GCMs and ML, which recovered *Y. diademata* sister to the rest of the family. Interestingly, when more loci were analyzed in Iteration 2a, support for the sister relationship between *Y. diademata* and *Y. everetti* + *Y. castaniceps* in STAR analysis dropped from 99% in Iteration 1 to 68% in Iteration 2a (App. 1.1). Using only the most informative loci, support for this same relationship was only 82% in Iteration 1i; and when more loci were analyzed in Iteration 2ai, the alternative topology recovered by the other methods was also recovered by STAR, albeit with only 66% bootstrap support (App. 1.5). The drastic decline in bootstrap support and change in topology between analyses greatly suggest that the highly supported placement of *Y. diademata* as sister to *Y. everetti* + *Y. castaniceps* is erroneous and is a bias unique to STAR. We have observed that STAR places an early diverging lineage sister to the next diverging clade in another dataset with a large number of taxa. The possible causes of this bias, which may include the high number of taxa in analysis, the proximity of the taxon to

the root of the species tree, the number of loci analyzed, or the presence of uninformative loci should be further investigated.

Improving Resolution by Subsampling Taxa

This study shows that analyzing subsamples of taxa with GCMs improves resolution in their species tree estimates, but the improvement varies across methods. This strategy succeeds by increasing the number of loci analyzed by these statistically consistent methods. As has been found by Song et al. (2012), subsampling taxa does not seem to have an effect on the topology of species tree estimates, as shown by the absence of conflicts between results that were not attributed to known biases. This divide-and-conquer strategy will provide the largest benefits in datasets with high levels of missing data at the taxon level such as those collected with inefficient sequence capture techniques; however, this strategy will be of little value with datasets collected from complete genomes or from using highly efficient sequence capture techniques (e.g., Bi et al. 2012).

A divide-and-conquer approach has been developed for MP-EST that improves its efficiency and performance (Bayzid et al. 2014). This method implements an iterative process of estimating species trees from subsets of taxa, combining these estimates using a supertree method, and scoring the combined tree. This method did not take into account the biases that can be introduced by missing data presented here but it can be modified to implement the analytical strategies proposed in this paper.

Comparison of GCMs

Among the four GCMs used in this study, STEAC performed better than the others in three ways. First, the resolution of its initial species tree estimate in Iteration 0 was significantly higher than those obtained using the other GCMs, and in fact, only slightly less than that of its

final species tree estimate or the ML result (Fig. 1.6a). Second, STEAC was least affected by bias introduced by samples with short sequences. In Iteration 0 STEAC placed three out of the four samples with short sequences close to their final estimated position (Fig. 1.5). Third, bias introduced by uninformative loci did not affect STEAC species tree estimates. STEAC recovered a monophyletic *Z. kulambangrae* in all analyses. STEAC's use of branch length information from gene trees, in addition to topology, makes it more robust to these biases from short sequence lengths and uninformative loci. This GCM assumes a uniform substitution rate across lineages (i.e., the molecular clock; Liu et al. 2009b) but is robust to deviations from the molecular clock when the number of loci is large (Liu et al. 2009a). These advantages make STEAC a good GCM to employ in analyzing phylogenomic datasets.

Of the GCMs that make use only of topology information from gene trees, ASTRAL achieved the most resolved species trees across the different analyses, slightly better than STAR (Fig. 1.6b). MP-EST, one of the most popular GCMs used, appears to require more loci to achieve the same level of resolution in other GCMs (Fig. 1.6b). These three GCMs were sensitive to biases from both samples with short sequence lengths and from uninformative loci.

Missing Data and GCM Analysis

The bias introduced by samples with short sequence lengths shows the sensitivity of GCMs to missing data at the alignment level, a bias that does not influence concatenated analyses. These results highlight another case of coalescent methods being more sensitive to missing data than concatenation methods (Thomson et al. 2008). We have avoided any possible biases to GCM results caused by missing lineages by performing analysis on datasets with no missing lineages. A number of gene-tree based species-tree methods such as STEM, MP-EST, ASTRAL, and NJst can accommodate missing lineages in input gene trees (Liu et al. 2010; Liu

and Yu 2011; Hovmöller et al. 2013; Mirarab et al. 2014c). However, STEM requires that lineages are missing at random across loci (Hovmöller et al. 2013), and MP-EST and NJst can perform poorly with non-random missing lineages or many missing lineages, respectively (Liu et al. 2010; Liu and Yu 2011; Zhong et al. 2014). Species tree methods that are robust to non-random missing lineages, as is expected in empirical data, would be ideal to take full advantage of the size of phylogenomic datasets. Nonetheless, the potential biases that can be introduced by missing data to results of current GCMs emphasizes the importance of exploring the properties of phylogenomic datasets, including randomness of missing lineages, sequence length heterogeneity, and information content of loci, to inform the choice of methods and analytical strategies.

Zosteropidae Relationships

The resolution of the ML and GCM topologies represents a vast improvement over that obtained by Moyle et al. (2009) using three markers, underscoring the power of genome-wide sequence data to resolve difficult rapid radiations. The clades recovered in this study were generally circumscribed by known biogeographic regions in varying scales (from archipelagoes to continents). In addition, the species trees estimated from this study are generally compatible with the results obtained by Moyle et al. (2009), although a few well-supported conflicts between them exist and are discussed below.

Three instances of conflicts between the results of this study and that of Moyle et al. (2009) involve samples of three species (*Z. palpebrosus*, *Z. montanus*, and *Z. erythropleurus*) that were sourced from different localities. In the first species, Moyle et al. (2009) recovered one sample of *Z. palpebrosus palpebrosus* from Nepal as sister to *Z. japonicus*, which occurs in East Asia. On the other hand, this study placed one *Z. palpebrosus siamensis* sample from Vietnam

sister to a clade of *Zosterops* species from Africa and Madagascar, and another *Z. palpebrosus unicus* sample from the Lesser Sundas sister to the Lesser Sundas endemic *Z. citrinella*. In the second and third species, a sample of *Z. montanus* from Sulawesi and a captive individual of *Z. erythropleurus* were placed by Moyle et al. (2009) in a clade with a Bornean endemic (*C. emiliae*), a Sumatra-Borneo endemic (*Z. atricapilla*), and an E. African/Arabian peninsula endemic (*Z. abyssinica*). In this study, a *Z. montanus* sample from the Philippines was found to be sister to a Philippine endemic *Z. meyeni*, whereas a *Z. erythropleurus* sample from Vietnam was recovered sister to a clade containing *Z. japonicus*, and the two species from the Philippines (*Z. montanus* and *Z. meyeni*). Paraphyly in *Zosterops* species is evident in several species in this study and others (Warren et al. 2006; Moyle et al. 2009; Cox et al. 2014). It is likely that these conflicts reflect the true paraphyly in these three species that have fairly wide geographical distributions, but further molecular work with more intensive sampling of these widespread species is needed to confirm this and clarify species limits. One sample of *Z. ugiensis* from the same individual was used by both studies but placed in different clades. Moyle et al. (2009) placed this species endemic to the Solomon Islands in a clade with two species from the West Pacific, whereas this study recovered it in a clade with seven other species from the Solomon Islands. This conflict could possibly have arisen if the few genes used by Moyle et al. (2009) had a gene tree topology that was different from the species tree.

GCMs and Concatenation

One weakness of GCMs is their supposed difficulty in recovering clades that are easily recovered by concatenation methods. In this study, however, after using our three analytical strategies, all clades recovered by ML were recovered by at least one GCM method (Figs. 1.4,

1.9). Contrary to expectations, four well-supported clades found in GCM results were not found in ML results.

The congruence in species tree estimates of Zosteropidae obtained by GCM and concatenation approaches is remarkable. One of the primary reasons for using GCMs in a group like Zosteropidae is the potential for the ML species-tree estimate to be highly supported but incorrect if a section of the species tree is in the anomaly zone (Kubatko and Degnan 2007). However, given the high number of short internodes in the family, only two well-supported conflicting results were recovered between the GCM results and the ML result, and one involves a biased result by STAR. The internodes are extremely short in the African/Malagasy clade to which *Z. melanocephalus* belongs (Fig. 1.4), so this example is the only candidate case of a section of the Zosteropidae phylogeny falling within the anomaly zone. If there are others, then we have not detected them because either the GCMs or ML have not resolved them with high support. The accuracy of phylogenetic reconstructions of Zosteropidae derived here from GCMs and ML cannot be evaluated, but the consistency between topologies and comparable resolutions indicate that our estimates from concatenation and coalescent approaches are roughly as accurate as each other.

CHAPTER 2

Ultraconserved Elements Resolve Phylogenetic Relationships in Core Corvoidea Passerines

ABSTRACT

Oscine passerines in the core Corvoidea group, comprising 773 species that occur worldwide, represent one of the major radiations of perching birds (Passeriformes). Phylogenetic relationships among the 29 families in this enigmatic clade have been largely unresolved owing to a combination of inadequate character and taxon sampling in previous studies and short time intervals between lineage splitting events. We use sequence data from thousands of ultraconserved element loci obtained from 86 species to estimate higher level phylogenetic relationships in the group using maximum likelihood (ML) and coalescent methods. A highly resolved phylogeny of core Corvoidea is recovered, with a majority of nodes receiving high support from both ML and coalescent analyses. Four families restricted to or thought to have originated from the Australia-New Guinea region and New Zealand form the earliest branching lineages in the group. The other families are divided into three major clades each consisting of 8–9 families. Our results are vital to understanding the biogeographic history of the core Corvoidea, which has been considered an example of a major avian radiation that originated from an island setting and colonized almost the entire world.

INTRODUCTION

Passerine birds (order Passeriformes) are the largest avian radiation, representing almost 60% of avian diversity (Cracraft 2014). Our knowledge of higher level phylogenetic relationships in this ubiquitous order consisting of 6063 species of perching birds (Dickinson and Christidis 2014) has grown tremendously over the past decade (summarized in Cracraft 2014). It is now well known that Passeriformes consists of three main clades (Barker et al. 2002, 2004; Ericson et al. 2002). The New Zealand wrens (Acanthisittidae), which are represented by only 2 extant species, are sister to all other passerines, which themselves are subdivided into two large clades: the suboscines and the oscines. Suboscines, which comprise 21% of passerine diversity, occur in the tropical regions of the New World and Old World but a vast majority of species are restricted to the Neotropics. The oscines, or songbirds, represent a higher share of passerine diversity (79%) and have a cosmopolitan distribution.

One of the most enigmatic groups of passerines involves a clade of oscines referred to as the "core Corvoidea" (Barker et al. 2002) or infraorder Corvides (Cracraft 2014). The core Corvoidea comprises 773 species belonging to 29 families (Cracraft 2014; Dickinson and Christidis 2014), the largest of which contains 126 species with an almost cosmopolitan distribution (Corvidae). In contrast, four families are each represented by a single species with restricted ranges: Eulacestomatidae, Rhagologidae, and Ifritidae are endemic to the highlands of New Guinea, whereas Pityriasidae occurs only on Borneo. Recent molecular phylogenetic work (Barker et al. 2004; Norman et al. 2009; Jønsson et al. 2011; Aggerbeck et al. 2014; Selvatti et al. 2015) have had limited success in resolving interfamilial relationships in this clade (Fig. 2.1) owing to a combination of inadequate data (Cracraft 2014) and short intervals between lineage splitting events. The study that obtained the highest phylogenetic resolution recovered four

major clades: one containing only the New Zealand endemic Mohouidae, two clades each comprising 10 families, and one clade consisting of 9 families (Aggerbeck et al. 2014, Fig. 2.1c). However, support for interfamilial relationships in this study was attained mostly from Bayesian inference; the maximum likelihood topology was poorly resolved. In addition, it is unclear whether the high support obtained from Bayesian analysis of 22 concatenated loci could have resulted from concatenation approaches being misled by gene tree discordance in an anomaly zone, a scenario that can occur when sufficiently short successive internodes are present in the species tree (Degnan and Rosenberg 2006; Kubatko and Degnan 2007).

The core Corvoidea has been hypothesized to have originated in proto-Papuan islands in the late Eocene/Oligocene and has been held as an example of a major global radiation that was derived from an island setting and subsequently colonized and diversified across the globe (Jønsson et al. 2011; Aggerbeck et al. 2014). These conclusions, however, were based in part on phylogenies with limited resolution. A well-resolved phylogenetic estimate of the group is required for a more robust hypothesis of the tempo and mode of diversification in this major oscine radiation.

Here we use thousands of genome-wide loci to clarify phylogenetic relationships among families in the core Corvoidea. Concatenation and coalescent approaches to phylogenetic inference are used to take into account gene tree discordance among loci and to take advantage of the large size of phylogenomic data.

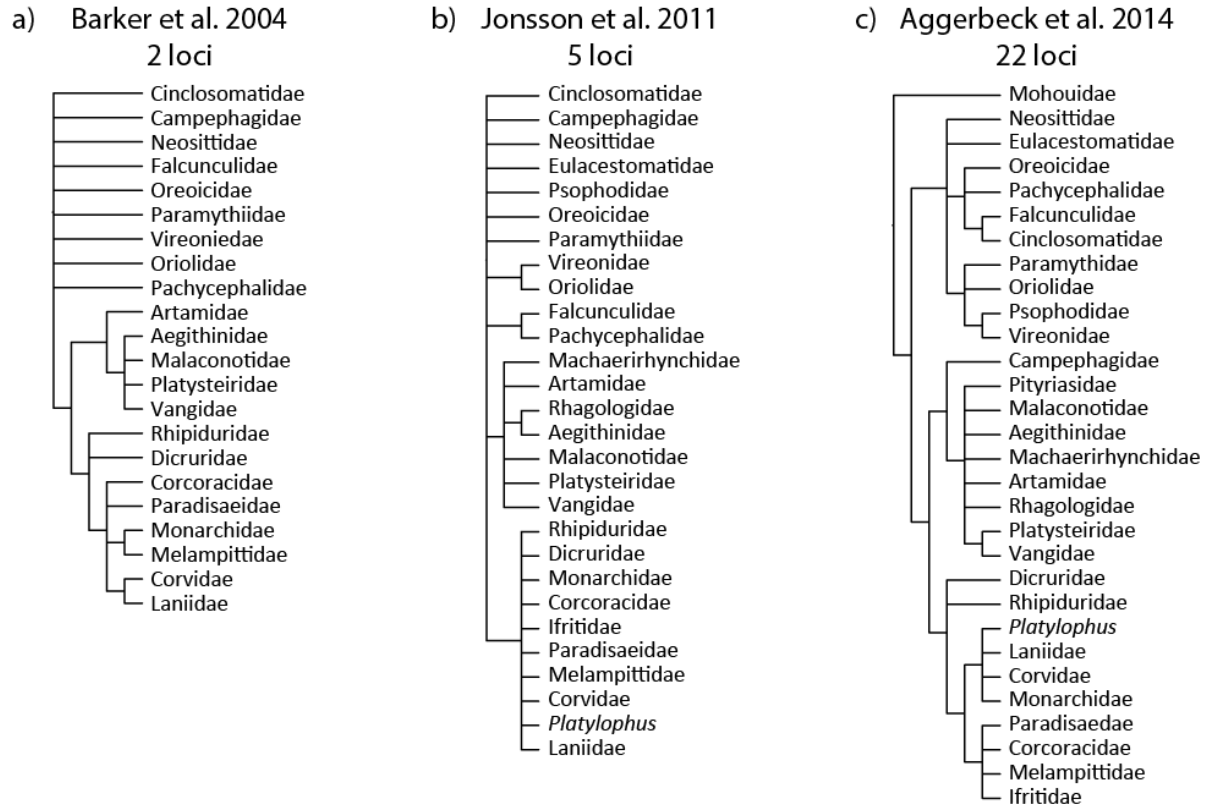


Figure 2.1. Estimate of interfamilial relationships in core Corvoidea from Bayesian inference in previous studies: (a) Barker et al. (2004), (b) Jonsson et al. (2011), and (c) Aggerbeck et al. (2014). The number of markers used in each study are indicated. Nodes with < 95% Bayesian posterior probability are collapsed.

METHODS

Sampling

We sampled 76 genera belonging to 28 out of the 29 recognized families in core Corvoidea (Dickinson and Christidis 2014). Only one family, the Bornean endemic Pityriasidae, was unsampled. All genera were represented by one species, with the exception of *Dicrurus*, which was sampled with 2 species (Table 2.1). Species from 9 families (Meliphagidae, Orthonychidae, Pomatostomidae, Cnemophilidae, Petroicidae, Cisticolidae, Stenostiridae, Passeridae, and Chloropseidae) were chosen as outgroups; thus, the total number of species in the study was 86.

Table 2.1. Sampling information.

Species	Family	Accession Number	Locality	Number of UCE loci enriched
Aegithina lafresnayei	Aegithinidae	KU 23213	Vietnam	4247
Aleadryas rufinucha	Oreoicidae	KU 16569	Papua New Guinea	3674
Arses telescophthalmus	Monarchidae	KU 12218	Papua New Guinea	4230
Artamus cinereus	Artamidae	KU 6183	Australia	4340
Batis senegalensis	Platysteiridae	KU 15403	Ghana	4161
Chaetorhynchus papuensis	Rhipiduridae	KU 16414	Papua New Guinea	4219
Chlorophoneus sulfureopectus	Malaconotidae	KU 15498	Ghana	3956
Cinclosoma punctatum	Cinclosomatidae	UWBM 57790	Australia	4344
Cissa hypoleuca	Corvidae	KU 23241	Vietnam	4300
Colluricincla harmonica	Pachycephalidae	KU 8866	Australia	4212
Coracina papuensis	Campephagidae	KU 27758	Papua New Guinea	4213
Corcorax melanoramphos	Corcoracidae	KU 10720	Australia	4082
Corvus corax	Corvidae	KU 30042	Alaska, USA	4579
Cyanocitta cristata	Corvidae	KU 2271	Kansas, USA	4088
Cyanocorax violaceus	Corvidae	KU 21655	Peru	4129
Cyanolyca viridicyanus	Corvidae	KU 17037	Peru	4113
Daphoenositta chrysoptera	Neosittidae	KU 23086	Australia	4547
Dendrocitta cinerascens	Corvidae	KU 17738	Borneo, Malaysia	4572
Dicrurus aeneus	Dicruridae	KU 23352	Vietnam	4381
Dicrurus paradiseus	Dicruridae	KU 23288	Vietnam	4290
Dryoscopus cubla	Malaconotidae	KU 26685	Botswana	4009
Edolisoma tenuirostre	Campephagidae	KU 23644	Palau	4250
Epimachus meyeri	Paradisaeidae	KU 16421	Papua New Guinea	4099
Erpornis zantholeuca	Vireonidae	KU 27942	Vietnam	4209
Eulacestoma nigropectus	Eulacestomidae	KU 16397	Papua New Guinea	3661
Falcunculus frontatus	Falcunculidae	UWBM 57627	Australia	4585
Grallina cyanoleuca	Monarchidae	KU 22946	Australia	4324
Gymnorhinus cyanocephalus	Corvidae	UNM NK165437	New Mexico, USA	4135
Hemipus picatus	Vangidae	KU 23303	Vietnam	4177
Hypothymis azurea	Monarchidae	KU 20189	Philippines	4657
Ifrita kowaldi	Ifritidae	KU 12106	Papua New Guinea	4389
Lalage maculosa	Campephagidae	KU 26428	Fiji	4302
Lamprolia victoriae	Rhipiduridae	KU 24329	Fiji	4265
Laniarius barbarus	Malaconotidae	KU 15451	Ghana	4009
Lanius excubitor	Laniidae	KU 28984	Mongolia	4585
Lophorina superba	Paradisaeidae	KU 16456	Papua New Guinea	4297
Machaerirhynchus nigripectus	Machaerirhynchidae	KU 4734	Papua New Guinea	4429
Melampitta lugubris	Melampittidae	KU 16552	Papua New Guinea	4477
Melanorectes nigrescens	Pachycephalidae	KU 18387	Papua New Guinea	4532

Melloria quoyi	Artamidae	KU 4831	Papua New Guinea	4008
Metabolus takatsukasae	Monarchidae	KU 22596	Northern Marianas	4522
Mohoua ochrocephala	Mohouidae	YPM 96717	New Zealand	3096
Myiagra azureocapilla	Monarchidae	KU 24306	Fiji	4402
Mystacornis crossleyi	Vangidae	FMNH 345860	Madagascar	4246
Oreocharis arfaki	Paramythiidae	KU 16440	Papua New Guinea	4445
Oreoica gutturalis	Oreoicidae	KU 8884	Australia	4179
Oriolus chinensis	Oriolidae	KU 10450	China	4587
Ornorectes cristatus	Oreoicidae	KU 4697	Papua New Guinea	3987
Pachycephala vitiensis	Pachycephalidae	KU 19410	Solomon Islands	4284
Paradisaea minor	Paradisaeidae	KU 16148	Papua New Guinea	4195
Parotia lawesii	Paradisaeidae	KU 16428	Papua New Guinea	3703
Pericrocotus divaricatus	Campephagidae	KU 10261	China	3766
Perisoreus canadensis	Corvidae	AMNH DOT15892	New York, USA	4304
Philentoma pyrhoptera	Vangidae	KU 12323	Borneo, Malaysia	4126
Phonygammus keraudrenii	Paradisaeidae	KU 12256	Papua New Guinea	4207
Pitohui dichrous	Oriolidae	KU 12200	Papua New Guinea	4432
Platylophus galericulatus	Corvidae	KU 24459	Borneo, Malaysia	4332
Platysteira cyanea	Platysteiridae	KU 29120	DR Congo	4255
Prionops plumatus	Vangidae	KU 26690	Botswana	4572
Pseudorectes ferrugineus	Pachycephalidae	KU 9610	Papua New Guinea	4134
Psophodes occidentalis	Psophodidae	KU 6204	Australia	4053
Pteruthius flaviscapis	Vireonidae	KU 17806	Borneo, Malaysia	4124
Ptilorrhoa leucosticta	Cinclosomatidae	KU 16514	Papua New Guinea	4458
Ptilostomus afer	Corvidae	LSU B39279	Ghana	4208
Pyrrhocorax pyrrhocorax	Corvidae	KU 28865	Mongolia	4309
Rhagologus leucostigma	Rhagologidae	KU 18382	Papua New Guinea	4546
Rhipidura javanica	Rhipiduridae	KU 17717	Borneo, Malaysia	4559
Sphecotheres viridis	Oriolidae	KU 10752	Australia	3990
Strepera graculina	Artamidae	KU 9660	Australia	4190
Struthidea cinerea	Corcoracidae	KU 10728	Australia	4313
Symposiachrus verticalis	Monarchidae	KU 27668	Papua New Guinea	4347
Tchagra senegalus	Malaconotidae	KU 15518	Ghana	4099
Tephrodornis virgatus	Vangidae	KU 30886	Vietnam	4206
Terpsiphone paradisi	Monarchidae	KU 23463	Vietnam	4342
Trochocercus nitens	Monarchidae	KU 15689	Ghana	4268
Urolestes melanoleuca	Laniidae	KU 26670	Botswana	4136
Vireo solitarius	Vireonidae	KU 25186	Kansas, USA	4189
Outgroup:				
Chelidorhynx hypoxanthus	Stenostiridae	KU 28022	Vietnam	4278
Chloropsis sonnerati	Irenidae	KU 24451	Borneo, Malaysia	4342
Cisticola anonymus	Cisticolidae	KU 29165	DR Congo	4317

Cnemophilus loriae	Cnemophilidae	KU 16529	Papua New Guinea	4644
Foulehaio carunculatus	Meliphagidae	KU 26344	Fiji	4549
Orthonyx temminckii	Orthonychidae	UWBM 76694	Australia	4503
Passer domesticus	Passeridae	KU 28860	Mongolia	4615
Petroica multicolor	Petroicidae	KU 24355	Fiji	4324
Pomatostomus superciliosus	Pomatostomidae	KU 8792	Australia	4388

Data Collection and Assembly

We extracted and purified DNA from fresh muscle or liver tissue (n = 85) or toe pad clips from museum specimens (n = 1) using the Qiagen DNeasy Blood and Tissue Kit following the manufacturer's protocol. Sequence capture of ultraconserved element (UCE) loci was performed targeting 5,060 UCE loci following the same process of library preparation, target capture, post-enrichment amplification, and sequencing described in Chapter 1. We followed the procedures from Chapter 1 in cleaning reads, assembling reads into contigs, compiling contigs into complete and incomplete datasets, aligning loci, trimming loci, and reformatting data to phylip format.

Phylogenetic Analyses

Maximum likelihood (ML) inference was performed on the concatenated loci of the incomplete dataset using RAxML ver. 8.1.3 (Stamatakis 2014) assuming a general time reversible model of rate substitution and gamma-distributed rates among sites. Node support was evaluated using 1000 rapid bootstraps.

Gene tree-based coalescent methods (GCM) were used to estimate relationships in core Corvoidea by applying strategies of subsampling taxa, filtering loci by information content, and trimming alignments that include short sequences to increase resolution and consistency among estimates (Chapter 2). We performed coalescent analysis on a total of five subsets of the 86 species. The first subset contained 21 species and focused on the basal families and major superfamilies in core Corvoidea, excluding *Mohoua albicilla*, the only sample that contained

substantially lower average contig length compared to the rest of the samples. The second subset consisted of the same species but this time included *Mohoua*. With this subset, alignments were trimmed so that no missing data was present at their flanks. The remaining three subsets focused on the three multi-family clades recovered by Aggerbeck et al. (2014; Fig. 2.1c). In this text we refer to these groups as "Orioloidea", "Malaconotoidea", and "Corvoidea," corresponding to clades X, Y, and Z in that study, respectively. Thus the last three subsets comprised 25 species with Orioloidea as ingroup, 25 species with Malaconotoidea as ingroup, and 43 species with Corvoidea as ingroup. Loci in which the number of parsimony-informative sites or average bootstrap support in gene tree inference were lower than their respective means were excluded from coalescent analysis.

For coalescent analyses, gene tree inference and bootstrapping were performed with RAxML ver. 8.1.3 (Stamatakis 2014) using the python package phyluce (Faircloth 2014). We modified the phyluce scripts to implement multi-locus bootstrapping (Seo 2008), and generated 500 multi-locus bootstrap replicate sets of gene trees for each dataset. On each replicate set of gene trees, we performed coalescent analyses using four GCMs: STAR and STEAC as implemented in the R package phybase ver. 1.3 (Liu et al. 2009b), MP-EST ver. 1.4 (Liu et al. 2010), and ASTRAL ver. 4.7.7 (Mirarab et al. 2014c). All programs were run with default options. Species trees inferred from the bootstrap replicates were summarized with 70% consensus trees using the sumtrees.py program in Dendropy (Sukumaran and Holder 2010). Command line, python, and R scripts used to process the data and run the species tree analyses are available at https://github.com/carloliveros/uce-scripts.

RESULTS

Dataset Attributes

We enriched an average of 4,257.5 UCE loci per sample (Table 2.1). The incomplete dataset used for ML analysis included 4121 loci with a mean locus length of 603.6 bp and total alignment length of 2,487,457 bp. Nucleotide data were present in 89.6% of the data matrix. GCM analyses were performed on 602–775 of the most informative loci in each dataset from an original 1,979–2,342 loci in common to all taxa in each dataset (Table 2.2). The average locus length ranged from 744.5–817.5 bp in the datasets that were not trimmed, but was less than half these lengths (291.9 bp) in the trimmed dataset. Similarly, the average number of parsimony-informative sites and average gene tree bootstrap support value both had considerably lower values in the trimmed dataset compared to the untrimmed datasets.

Table 2.2. Characteristics of datasets used in coalescent analyses

Dataset description	Number of taxa	Number of loci in common	Number of loci analyzed	Average locus length	Average number of parsimony-informative sites	Average gene tree bootstrap support value
Higher level core Corvoidea, excluding *Mohoua*	21	2342	775	766.6	52.6	34.8
Higher level core Corvoidea, including *Mohoua*, trimmed	22	2342	602	291.9	6.3	17.5
Orioloidea	25	2084	679	744.5	55.6	35.1
Malaconotoidea	25	2201	775	817.5	65.5	42.8
Corvoidea	43	1979	764	782.3	80.5	40.4

Phylogenetic Relationships

Both ML and coalescent analyses yielded well-resolved estimates of core Corvoidea phylogeny. Results obtained in both approaches largely agree so only the ML estimate is shown (Fig. 2.2). Results of coalescent analyses were consistent with each other (App. 2.1–2.2) and the few differences are discussed in the text. In ML results, 72 out of 76 nodes (95%) in core

Corvoidea were well-supported, (i.e., received > 70% bootstrap support; BS), 68 of these with 98–100% BS (Fig. 2.2). Among GCM results, resolution was comparable between methods with 65–67 well-supported nodes out of 76, 60–62 of these with 98–100% BS (App. 2.1–2.2).

Relationships within Families.—All families that were represented by more than one genus were recovered monophyletic with 100% BS across ML and four coalescent analyses, with the exception of Corvidae (see below). Estimates of intergeneric relationships within families that were represented by more than two genera other than Corvidae were all well-supported across ML and four coalescent analyses. All but one node received 98–100% BS with the exception of a clade of 5 genera in Monarchidae (*Symposiachrus, Metabolus, Grallina, Myiagra*, and *Arses*) which received 74% BS in ML and 85–93% in coalescent analyses. Only one case of well-supported conflicting relationship was recovered across all analyses: within Paradisaeidae, ML placed *Lophorina* sister to a clade containing *Paradisaea* and *Epimachus*, whereas all GCMs recovered *Paradisaea* sister to the pair of *Lophorina* and *Epimachus*.

Interfamilial Relationships.—For a clearer view of phylogenetic relationships among families in core Corvoidea, branches from ML and coalescent analyses were collapsed so that each tip represented a monophyletic family (Fig. 2.3). At the family level, the ML analysis achieved higher resolution (24 out of 28 nodes were well-supported) than coalescent analyses (17–19 well-supported nodes). ML inference recovered six major clades in core Corvoidea that branched sequentially from the base of the clade. The first clade, which was sister to the rest of core Corvoidea, consisted of a single family endemic to the Australia-New Guinea region, Cinclosomatidae. The second clade included only Campephagidae, which has a broad Old World distribution with members occurring from Africa to the Southwest Pacific. The third clade comprised two small families: Mohouidae, a New Zealand endemic, and Neosittidae, an

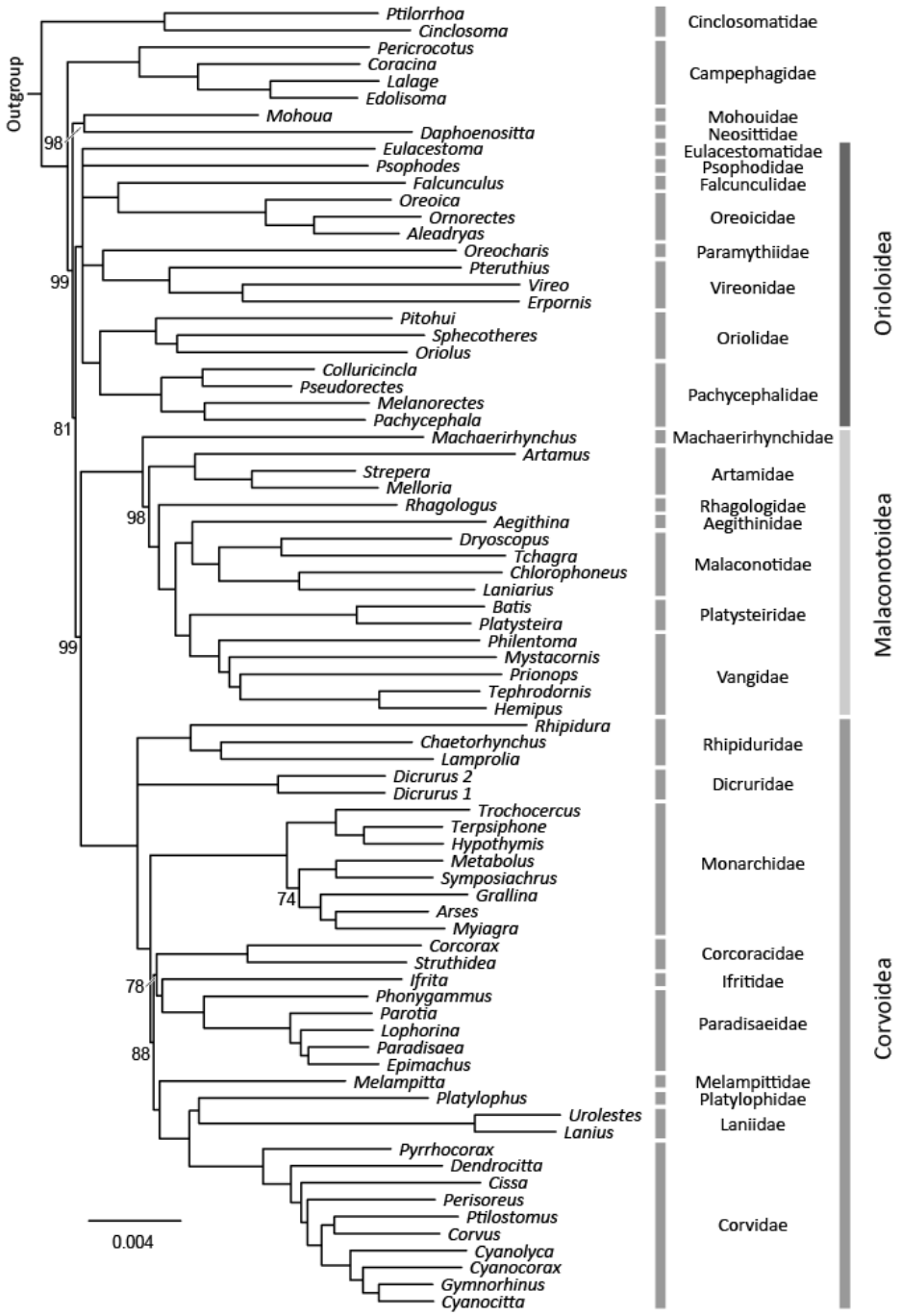


Figure 2.2. Maximum likelihood estimate of phylogenetic relationships in core Corvoidea inferred from 4121 UCE loci using RAxML. Numbers below branches indicate bootstrap support (BS) values from 1000 bootstrap replicates. Nodes with < 70% BS are collapsed whereas nodes with no labels indicate 100% BS.

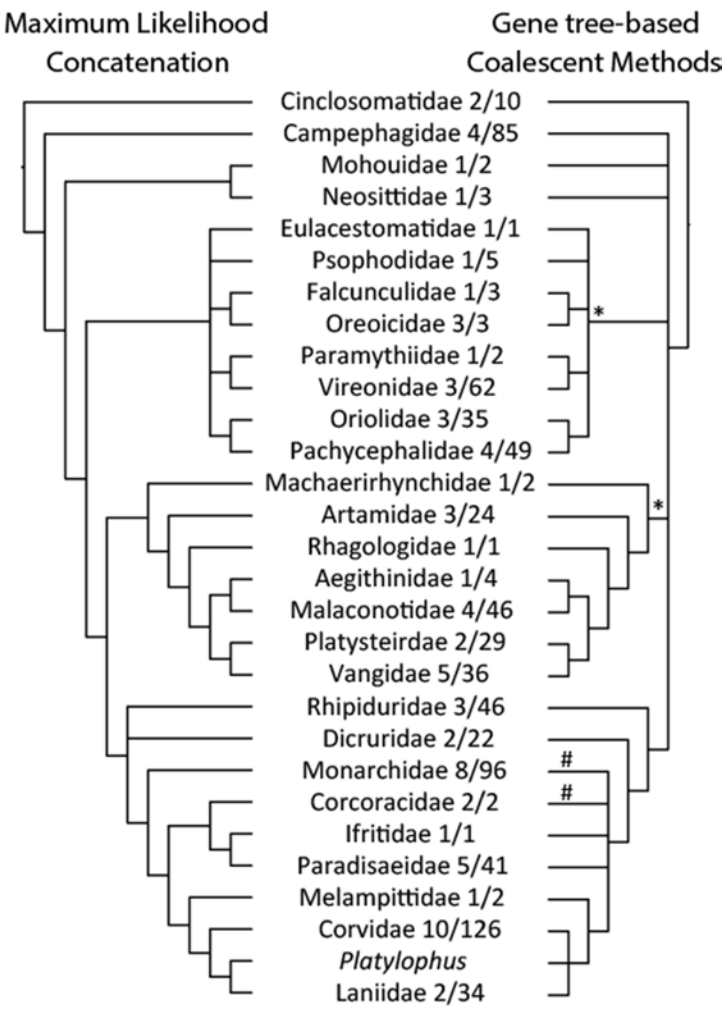


Figure 2.3. Estimate of interfamilial relationships in core Corvoidea based on maximum likelihood and ASTRAL analyses. Results of STAR, STEAC, and MP-EST were similar to that of ASTRAL; topological differences are indicated by symbols (*, #) and are discussed in the text. Nodes with < 70% BS are collapsed. BS values for each node vary across GCMs and subsampled datasets and are thus not shown, but they are provided in Appendices 2.1–2.2.

Australia-New Guinea-region endemic. The fourth clade is composed of 8 families: Eulacestomidae, Psophodidae, Falcunculidae, Oreoicidae, Paramythiidae, Vireonidae, Oriolidae, and Pachycephalidae. The fifth and sixth sister clades both contained several families. The fifth clade included seven families: Machaerirhynchidae, Artamidae, Rhagologidae, Aegithinidae, Malaconotidae, Platysteiridae, and Vangidae, whereas the sixth clade comprised nine families:

Rhipiduridae, Dicruridae, Monarchidae, Corcoracidae, Ifritidae, Paradisaeidae, Melampittidae, Corvidae, and Laniidae.

On the other hand, coalescent analyses recovered seven main clades in core Corvoidea among which relationships were largely not well-supported (Fig. 2.3). Similar to ML results, Cinclosomatidae was placed sister to the rest of core Corvoidea in coalescent analyses. However, the six other main clades formed a six-way polytomy with the following members: Campephagidae, Mohouidae, Neosittidae, and the same last three clades recovered in ML analysis (Orioloidea, Malaconotoidea, and Corvoidea). STAR and MP-EST recovered Orioloidea sister to Malaconotoidea in one of 5 subsets of taxa (App. 2.2a), albeit with only 76% and 74% BS, respectively.

Coalescent and ML analyses recovered identical but poorly resolved relationships within Orioloidea. This clade comprised five lineages forming a polytomy: Eulacestomidae, Psophodidae, the pair of Falcunculidae and Oreoicidae, the pair of Paramythiidae and Vireonidae, and the pair of Oriolidae and Pachycephalidae. Identical, highly resolved topologies were inferred by ML and coalescent analyses within Malaconotoidea. Machaerirhynchidae was recovered sister to the other members of Malaconotoidea. Next to branch out in succession were Artamidae, followed by Rhagologidae. The final clade within Malaconotoidea comprised the pair of Aegithinidae and Malaconotidae sister to the pair of Platysteiridae and Vangidae.

Within Corvoidea, ML and coalescent analyses resulted in slightly different, but compatible topologies (Fig. 2.3). In the ML result, Rhipiduridae, Dicruridae, and a clade containing the seven other families of Corvoidea formed a three-way polytomy. Coalescent analyses, however, recovered Rhipiduridae sister to all other Corvoidea families. ML fully resolved the clade of seven families with Monarchidae sister to a clade that in turn contained two

sister clades, each consisting of three families. In one of these triplets, Corcoracidae was sister to Ifritidae and Paradisaeidae; in the other, Melampittidae was recovered sister to Corvidae and Laniidae. Within the Corvidae and Laniidae group, the genus *Platylophus*, traditionally a member of Corvidae, was sister to Laniidae, and this clade was in turn sister to the rest of Corvidae. Coalescent analyses achieved much less resolution in this clade of 7 Corvoidea families, yielding a five-way polytomy with the following lineages: Monarchidae, Corcoracidae, Ifritidae, Paradisaeidae, and a clade containing the other three families. In the last clade, Melampittidae was recovered sister to a three-way polytomy of Corvidae, Laniidae, and *Platylophus*. STEAC's estimate of Corvoidea topology differed from that obtained by the other GCMs in two ways. First, it recovered Corcoracidae sister to Monarchidae, with 78% BS, a result compatible with other GCM results, but conflicting with ML results. Second, it placed *Platylophus* sister to the other members of Laniidae with 76% BS, marginally supporting Corvidae paraphyly. This finding is also compatible with other GCM results and congruent to the ML topology.

DISCUSSION

Core Corvoidea Relationships

This study provides a phylogenetic hypothesis for core Corvoidea with a vast improvement in resolution over previous work (Barker et al. 2004; Norman et al. 2009; Jønsson et al. 2011; Aggerbeck et al. 2014). The fact that most relationships were supported by both ML and coalescent analyses indicate that our results are robust and that the high support for most nodes in our ML estimate is not an artefact of the anomaly zone (Kubatko and Degnan 2007).

Our estimate of relationships among families in core Corvoidea is compatible with the well-supported relationships in Barker et al. (2004) and Jønsson et al. (2011; Figs. 2.1, 2.3). These two studies recovered the clades corresponding to Malaconotoidea and Corvoidea, with Barker et al. (2004) finding a sister relationship between these two clades. However, our results contradict these two studies in a few sister relationships among families that were found with high support in their Bayesian analyses. Barker et al. (2004) recovered Monarchidae sister to Melampittidae, whereas Jønsson et al. (2011) found sister relationships in the following pairs of families: Vireonidae and Oriolidae, Falcunculidae and Pachycephalidae, and Rhagologidae and Aegithinidae. In this study, each of these families were recovered in different clades with high support from both ML and coalescent methods.

Results of this study contradict many of the relationships recovered by Aggerbeck et al. (2014), the study that until this work included the densest character sampling and attained the highest resolution among core Corvoidea families. First, we find different basal lineages in the group. Aggerbeck et al. placed the New Zealand endemic Mohouidae sister to the rest of core Corvoidea whereas our data finds Cinclosomatidae, an endemic of the Australia-New Guinea region, in this position. Furthermore, ML results from our study recover Campephagidae, a widespread family thought to have originated in the Australia-New Guinea region (Jønsson et al. 2010), and the families Mohouidae and Neosittidae among the early lineages to break off from the group. Aggerbeck et al. (2014) placed Campephagidae within our Malaconotoidea and Neosittidae within our Orioloidea. Second, we find conflicting relationships within the Orioloidea clade. Our data recovered only three sets of well-supported sister families in this clade, but each one contradicts well-supported relationships in Aggerbeck et al. (2014). For example, we found Falcunculidae sister to Oreoicidae but Aggerbeck et al. recovered

Falcunculidae sister to Cinclosomatidae. Third, well-supported conflicts also occur within the Corvoidea clade. ML and coalescent analyses in this study recover Melampittidae, Corvidae, *Platylophus*, and Laniidae forming a clade, whereas Aggerbeck et al. group the last three lineages in a clade with Monarchidae. The high number of conflicting results between this study and that of Aggerbeck et al. (2014) may have been caused by the sparse taxon sampling in the latter. Aggerbeck et al. sampled only one individual in each family whereas this study used multiple genera in each family where possible. These conflicts, especially those that involve lineages at the base of core Corvoidea and the major clades, may have important implications in ancestral range reconstruction in the group.

Concatenation and coalescent methods

Three of the major clades in core Corvoidea illustrate the varying degrees by which concatenation and coalescent approaches can resolve clades with many short internodes. Within Orioloidea, both approaches yield identical results but with poor resolution. The low support values in ML analyses can be expected in cases of high gene tree heterogeneity, a low number of informative sites for this section of core Corvoidea phylogeny, or both. On the other hand, within Malaconotoidea, both approaches produce the same topology with high resolution, indicating little gene tree discordance in this group. Within Corvoidea, two contrasting patterns are observed. In the first pattern, coalescent approaches resolve a node that concatenation does not. Specifically, all GCMs found Dicruridae sister to the eight other families of Corvoidea with 91–96% BS (App. 2.2), whereas ML recovered the same relationship with only 64% BS. In this case, slight to moderate levels of gene tree discordance, likely due to incomplete lineage sorting, could have pulled down support values for this node in ML analysis, but not in GCM analyses, which are expected to perform well in this type of situation. In the second pattern, nodes are

resolved with moderate to high support by the concatenation approach, but poorly resolved by coalescent methods, such as in the clade containing Monarchidae, Corcoracidae, Ifritidae, Paradisaeidae, and the clade containing Melampittidae, Corvidae, Laniidae, and the genus *Platylophus*. In this situation, overall signal from nucleotides yields support in concatenated analysis, but that signal is not apparent in many genes, unlike the case in Malaconotoidea. Moreover, unlike in the Dicruridae pattern, GCM support values were low. This pattern is indicative of high gene tree estimation errors in GCM analysis.

One justification for using coalescent-based approaches in phylogenetic analysis is to detect highly-supported, incorrect topologies from concatenation approaches that are expected in species trees that fall within the anomaly zone (Kubatko and Degnan 2007). In this study, we have shown that not only do coalescent methods increase our confidence for highly supported nodes in concatenation results when they agree, as with the vast majority of nodes in our species tree estimate, coalescent methods can also provide high support for an otherwise weakly or moderately well-supported node in concatenation results. This result was demonstrated in the placement of Dicruridae sister to the rest of Corvoidea.

The conflicting results between ML and GCMs in estimating relationships among the genera *Lophorina*, *Paradisaea*, and *Epimachus* in the family Paradisaeidae is an apparent artefact of bias in GCMs caused by consistently short sequence lengths for the *Paradisaea* sample (Chapter 2). Further GCM analysis on a dataset containing only members of Paradisaeidae and 3 outgroups, with alignments trimmed to eliminate flanking missing data, and using only the most informative loci recovers the same result as in ML, with *Lophorina* sister to *Paradisaea* and *Epimachus* (not shown). Several other samples have shorter average sequence lengths than the *Paradisaea* sample, including *Pericrocotus, Eulacestoma, Psophodes,*

Aleadryas, *Corcorax*, and *Parotia*. The congruence of ML and GCM results for the phylogenetic position of these taxa indicates that missing sequence data in these samples did not bias GCM results. Either sufficient parsimony-informative sites were present in these short sequences or only a small proportion of loci lacked informative sites for these samples. In addition, genetic divergences between *Lophorina*, *Paradisaea*, and *Epimachus* were shallow compared to those between the other samples with short sequences and their closely related taxa. Thus, missing sequences in *Paradisaea* likely resulted in more severe gene tree estimation error in its clade.

The sister relationship between Orioloidea and Malaconotoidea recovered by STAR and MP-EST in the analysis of one subset of taxa should be interpreted with caution for two reasons. First, bootstrap support values for this node (74% and 76%) were on the low end of our definition of well-supported. Second, a support value > 70% for this node was found only in 2 out of 20 GCM analyses that contained both clades. In these two instances, Orioloidea formed part of the outgroup and had sparse taxon sampling. This relationship could be an artefact of sparse taxon sampling in this subset of the data. STEAC was the only GCM that recovered *Platylophus* sister to Laniidae, agreeing with ML results, and placed Corcoracidae sister to Monarchidae, contradicting ML results. The positions of these taxa were ambiguous in the results of other GCMs. Considering that the support values for these nodes were only 76% and 78% respectively, these results should also be considered with caution.

STEAC, the only GCM among the four used in this study that uses branch length information from gene trees, assumes the molecular clock across lineages (Liu et al. 2009b). The high congruence of the STEAC results with the other three GCMs indicates the robustness of this method to deviations from this assumption in this relatively deeply diverging and diverse group (Liu et al. 2009a).

Taxonomy

We propose two updates to the higher-level classification of Infraorder Corvides of Cracraft (2014) based on relationships recovered with high support from ML and coalescent approaches in this study. The superfamily Orioloidea can be erected to group Eulacestomatidae, Psophodidae, Falcunculidae, Oreoicidae, Paramythiidae, Vireonidae, Orioloidae, and Pachycephalidae. The best estimate of the position of the genus *Platylophus* is the one provided by ML and STEAC, which places it sister to Laniidae, rendering Corvidae paraphyletic. Thus, either *Platylophus* can be moved to Laniidae, or a new monotypic family can be erected.

CHAPTER 3

Phylogenomic Comparisons Clarify Relationships and Biogeographic History of a Rapid Pantropical Bird Radiation: the Trogons

ABSTRACT

Historical biogeographic inference is only as accurate and reliable as the phylogenetic hypotheses on which it is based. In the avian family of trogons (Trogonidae) multiple, conflicting biogeographic histories have been proposed to explain their pan-tropical distribution because of differing estimates of inter-generic relationships. Phylogenetic reconstruction in the trogons has been problematic because of successive short internodes at the base of this radiation that cause gene tree discordance and a long branch leading to the family that makes root placement tricky. We re-estimate inter-generic relationships in this rapid bird radiation using data from thousands of ultraconserved element (UCE) loci and employ analytical methods that consider gene tree discordance. We then examine the origin, tempo, and mode of diversification of the group. We recover the first well-supported hypothesis of relationships among trogon genera. Trogons comprise three clades, each confined to one of three biogeographic regions: Africa, Asia, and the Neotropics, with the African clade sister to the rest of trogons. This topology, combined with the trogon fossil record, geologic, and climatic data, suggests an Old World origin for the group. Continental connections during the warm Late Oligocene/Early Miocene facilitated dispersion between Eurasia, North America, and Africa, and subsequent global cooling plausibly caused divergence between main trogon lineages.

INTRODUCTION

Two main hypotheses attempt to explain the origins of bird lineages with disjunct distributions in portions of the Old World and New World. One hypothesis attributes these distributions to vicariance caused by the Cretaceous break-up of Gondwana (Cracraft 1973, 2001). The other hypothesis ascribes these disjunct distributions to a Laurasian origin in the Early Paleogene, during which bird lineages assumed a broad distribution under favorable Laurasian climatic conditions and suitable land connections before subsequently becoming isolated with the deterioration of climatic conditions and the severing of land bridges (Cracraft 1973; Olson 1989). Evaluating these hypotheses for a specific lineage requires a robust estimate of phylogenetic relationships and divergence times within the group and its close relatives, as well as a careful consideration of the fossil record.

The avian family Trogonidae (trogons) represents a challenge for reconstructing biogeographic history. The family, which is placed in its own order Trogoniformes, consists of 43 species in 7 genera distributed throughout much of the Old World and New World tropics. Multiple, conflicting biogeographic histories have been published for the group because of differences in the estimates of the family's phylogeny (Espinosa de los Monteros 1998; Johansson and Ericson 2005; Moyle 2005; Ornelas et al. 2009; Hosner et al. 2010). For instance, some studies recovered single clades each for African, Asian, and the Neotropical taxa (Espinosa de los Monteros 1998; Johansson and Ericson 2005; Ornelas et al. 2009), whereas others suggested the Neotropical taxa are paraphyletic (Moyle 2005; Hosner et al. 2010). These conflicting findings likely resulted from the use of different sets of only a few molecular markers, each supporting different topologies or providing no support at all. In all of these studies support for inter-generic relationships was low, with the exception of one (Ornelas et al.

2009) in which Bayesian posterior probabilities were high (100%) but maximum likelihood bootstrap support values were < 50%. Moreover, with the exception of one study (Hosner et al. 2010), previous work did not include the highly divergent Asian genus *Apalharpactes,* which may not be closely related to other Asian trogons.

Despite conflicting phylogenetic estimates for trogons, previous molecular studies have rejected a Gondwanan origin for this group based on estimated dates of divergence among crown trogons that are much younger than the sundering of Gondwana (Espinosa de los Monteros 1998; Moyle 2005). A Laurasian derivation is evident but these two studies offer two conflicting hypotheses of trogon origins. Espinosa de los Monteros (1998) concluded an African origin based on his results. He further noted that his hypothesis is supported by the fact that stem trogon fossils from the Paleocene have been found in Europe, whereas the oldest known fossil trogons from the New World are extant taxa from Pleistocene deposits (Espinosa de los Monteros 1998). On the other hand, Moyle's (2005) results supported a derivation of trogons in the New World, which had been previously proposed because of the higher trogon diversity in the region (Mayr 1946). Since the publication of these two studies, more information on the fossil record of trogons and close relatives have been revealed (Kristoffersen 2002; Mayr 2003, 2005, 2009; Weidig 2006; Ksepka and Clarke 2009) that has implications for understanding trogon origins. Nonetheless, a critical step to understanding trogon origins is clarifying phylogenetic relationships in the group.

The difficulty in resolving trogon phylogeny likely stems from two main factors. First, the base of the trogon phylogeny is characterized by short successive internodes, indicating a rapid initial burst of lineage splitting. Phylogenies with such structure are difficult to resolve with limited data because of the low probability of informative substitutions occurring along

these short internodes (Lanyon 1988) and the high probability of gene tree discordance (Degnan and Rosenberg 2006). Second, trogons have no close living sister-group such that a very long branch leads to this rapid radiation. This long branch, combined with subsequent short internodes at the base of the trogon clade, makes rooting the trogon phylogeny difficult (Johansson and Ericson 2005; Moyle 2005), which in turn has a large impact on the possible biogeographic conclusions. An additional factor that may be confounding phylogeny estimation in trogons is that the presence of successive short internodes may place the species tree within the anomaly zone, a situation in which the most common gene trees inferred from sampled loci are expected to be incongruent with the species tree, and as such, methods based on the multi-species coalescent may be necessary to resolve the species tree (Degnan and Rosenberg 2006; Kubatko and Degnan 2007).

Advances in sequencing technology have made collection of hundreds to thousands of genome-wide sequences feasible for phylogenetic studies (Faircloth et al. 2012; Lemmon et al. 2012). Recent studies using sequence capture of ultraconserved elements and their flanking regions (UCE loci) have clarified phylogenetic relationships in regions of the avian tree of life with successive short internodes (McCormack et al. 2013; Sun et al. 2014). Additionally, several analytical methods that are statistically consistent under the multi-species coalescent can be harnessed to estimate species trees from hundreds to thousands of loci (Liu et al. 2009b, 2010; Mirarab et al. 2014c). In this paper, we leverage these advantages to re-examine phylogenetic relationships in the rapid pan-tropical radiation of trogons using thousands of UCE loci. We then estimate the timing of diversification and evaluate hypotheses of the origin, tempo, and mode of diversification in this pan-tropical group using recent advances in our knowledge of the trogon fossil record.

METHODS

Sampling

Monophyly of all seven trogon genera has been established in previous molecular studies (DaCosta and Klicka 2008; Ornelas et al. 2009; Hosner et al. 2010). Because the focus of this study is on inter-generic relationships and broad-scale biogeographic patterns, we selected 11 trogon species that included at least one representative of each genus and one or two additional species from species-rich genera as ingroup taxa (Table 3.1). Samples from five closely related orders were used as outgroups.

Laboratory techniques

We extracted and purified DNA from fresh muscle or liver tissue using the Qiagen DNeasy Blood and Tissue Kit following the manufacturer's protocol. Sequence capture of ultraconserved element (UCE) loci was performed targeting 5,060 UCE loci following the same process of library preparation, target capture, post-enrichment amplification, and sequencing described in Chapter 1.

Table 3.1. Sampling, read, and contig characteristics.

Species	Accession Number	Locality	Total number of reads	Percentage of ≥ Q30 bases	Mean read quality score	Total number of contigs	Number of UCE contigs	Mean UCE contig length	Mean UCE coverag
Ingroup									
Apalharpactes mackloti	LSUMNS B49104	Indonesia	4,262,870	88.88	35.29	7310	4192	957.9	38.6
Apaloderma aequatoriale	KUNHM 8461	Equatorial Guinea	3,239,034	90.63	35.73	7222	4082	790.6	35.5
Euptilotis neoxenus	AMNH DOT11081	USA	4,834,786	89.94	35.53	8793	4078	881.3	42.3
Harpactes ardens	KUNHM 26958	Philippines	4,498,186	89.49	35.44	8860	4288	890.8	37.4
Harpactes erythrocephalus	KUNHM 9970	China	3,270,822	89.75	35.48	7294	3882	792.6	32.1
Harpactes oreskios	KUNHM 23185	Vietnam	2,996,054	87.9	35.04	7031	3695	799.3	26.2
Pharomacrus antisianus	LSUMNS B22870	Bolivia	4,868,516	88.64	35.26	7331	4201	1017.5	42.5
Priotelus roseigaster	KUNHM 6363	Dominican Republic	3,937,650	89.56	35.43	7362	4105	684.8	41.6
Priotelus temnurus	ANSP 5565	Cuba	5,680,700	88.34	35.16	8164	4264	1003	47.7
Trogon personatus	AMNH DOT4266	Venezuela	3,742,850	88.41	35.16	7938	3935	911.7	34.1
Trogon violaceus	AMNH DOT11951	Venezuela	3,491,602	85.66	34.46	9095	3195	748.7	31.1
Outgroup									
Berenicornis comatus	AMNH DOT14737	Captive	5,483,676	91.63	35.95	10962	4342	644.0	51.1
Ceyx argentatus	KUNHM 19269	Philippines	6,084,638	89.97	35.58	9033	4278	951.8	55.6
Leptosomus discolor	FMNH 449184	Madagascar	5,031,310	89.37	35.46	9321	4306	945.8	37.2
Otus elegans	KUNHM 10975	Philippines	5,732,988	89.5	35.45	9332	4225	955.2	46.8
Colius striatus	Downloaded from gigadb.org (Jarvis et al. 2014)	Captive	-	-	-	-	4872	2113.8	-

Acronyms: AMNH (American Museum of Natural History), ANSP (Academy of Natural Sciences of Drexel University), KUNHM (University of Kansas Natural History Museum), LSUMNS (Louisiana State University Museum of Natural Science).

Data assembly

Raw reads were de-multiplexed using CASAVA ver. 1.8.2. Low-quality bases and adapter sequences were trimmed from reads using illumiprocessor ver. 1 (https://github.com/faircloth-lab/illumiprocessor). Subsequent data processing was performed using the python package phyluce (Faircloth 2014) and outlined below. Cleaned reads were assembled into contigs using the program Trinity (Grabherr et al. 2011). Contigs matching UCE loci were extracted for each taxon. For *Colius striatus* UCE loci were obtained by *in silico* alignment of UCE probes with the full genome sequence (Jarvis et al. 2014) and taking the matched region along with 1000 bp of flanking nucleotides on each side of the matched region. Two datasets were then assembled: one "incomplete dataset" containing UCE loci that were present in at least 75% of taxa and a "complete dataset" containing UCE loci that were present in all taxa. Each locus was aligned using MAFFT (Katoh and Standley 2013) and trimmed using Gblocks (Castresana 2000) using default parameters with the exception of the minimum number of sequences for a flank position in Gblocks, which we set at 65% of taxa. The alignments were formatted to phylip and nexus files for phylogenetic analysis.

Phylogenetic analysis

Maximum likelihood (ML) and Bayesian inference were performed on the concatenated loci of the incomplete dataset. ML tree searches were carried out using RAxML ver. 8.1.3 (Stamatakis 2014) on the unpartitioned dataset assuming a general time reversible model of rate substitution and gamma-distributed rates among sites. Node support was evaluated using 500 rapid bootstraps. For Bayesian analysis, we first estimated substitution models for each locus using Cloudforest (Crawford and Faircloth 2014) and grouped loci with the same substitution model into separate partitions. We performed two sets of Bayesian analysis using ExaBayes ver.

1.3 (Aberer et al. 2014): one with 4 independent runs each with two coupled chains, and another with 8 independent runs with no coupled chains. Both MCMCs were run for 10^7 generations sampling every 5×10^3 generations. Convergence of likelihood and parameter estimates was assessed using the program Tracer ver. 1.6.0 (Rambaut and Drummond). Effective sample sizes of all parameters in all runs were > 200. Convergence of the independent runs was evaluated using the sdsf program of ExaBayes.

We used gene tree-based coalescent methods (GCM) to estimate the species tree from the complete dataset. Gene tree inference and bootstrapping were performed with RAxML ver. 8.1.3 (Stamatakis 2014) using the python package phyluce (Faircloth 2014). We modified the phyluce scripts to implement multi-locus bootstrapping (Seo 2008), i.e., sampling with replacement of loci and sites, and generated 500 multi-locus bootstrap replicate sets of gene trees for each dataset. On each replicate set of gene trees, we ran four GCMs: species tree estimation using average ranks of coalescences (STAR) and species tree estimation using average coalescence times (STEAC) (Liu et al. 2009b) as implemented in the R package phybase ver. 1.3, maximum pseudo-likelihood for estimating species trees (MP-EST) ver. 1.4 (Liu et al. 2010), and accurate species tree algorithm (ASTRAL) ver. 4.7.7 (Mirarab et al. 2014c). All programs were run with the default options. Consensus trees were generated for each method using the sumtrees.py program in Dendropy (Sukumaran and Holder 2010). Command line and R scripts used to process the data and run the species tree analyses are available at https://github.com/carloliveros/uce-scripts.

Loci with weak or no phylogenetic signal can lower bootstrap support values in species tree estimation (Liu et al. 2015) and can lead to incorrect GCM-estimated species trees with high support (Chapter 1). Thus, in addition to performing GCM analyses with all 1828 loci common

to all taxa in the dataset, we also performed three other sets of GCM analyses on subsets of loci consisting of 1469, 1095, and 741 loci with the highest number of parsimony-informative sites.

In order to assess the robustness of rooting the trogon phylogeny, we carried out additional ML tree searches using RAxML ver. 8.1.3 (Stamatakis 2014) evaluated with 500 rapid bootstraps on the incomplete dataset with four alternative sets of outgroups: (a) *Ceyx argentatus*, a member of the sister clade of trogons, as a single outgroup; (b) *Otus elegans*, the most distant taxon to trogons in our sampling, as a single outgroup; (c) *Ceyx argentatus* and *Berenicornis comatus*, two members of the sister clade of trogons, as the outgroup, and (d) *Ceyx argentatus* and *Otus elegans* as the outgroup.

We used the program MCMCtree in the PAML ver. 4.8 package (dos Reis and Yang 2011) to estimate absolute divergence times between trogon genera. The complete dataset was concatenated and treated as a single locus in the analysis. The tree topology was fixed to the well-supported topology inferred from both concatenated and species tree analyses. No fossils are known from the trogon crown clade (Mayr 1999, 2005; Kristoffersen 2002; Mayr and Smith 2013); thus two secondary time calibrations from a recent chronogram of avian bird orders based on genome-wide data and multiple fossil calibrations (Jarvis et al. 2014) were utilized. Normally distributed priors were assigned to both calibrations: one with mean of 59.942 Mya with s.d. 1.675 My at the split of Coliiformes and its sister clade; and the other with mean 51.985 Mya and s.d. 1.55 My at the split of Bucerotiformes and its sister clade. We used means and standard deviations in our calibration priors that reflected their posterior distribution in (Jarvis et al. 2014). A model of independent rates between lineages drawn from a lognormal distribution was used with gamma hyper priors for the mean and variance of rates. Hyper prior parameters (rgene_gamma and sigma2_gamma) were selected using estimates of the overall rate on the tree

obtained by the PAML program baseml. A birth-death-sampling model of lineage diversification was used with sampling rate of 0.25. Date estimates were robust to changing initial values for the hyper prior parameters and the birth and death rates. An HKY85 substitution model with gamma-distributed rates in 5 categories was chosen to account for substitution rate variation between sites. Two independent MCMC chains were run with parameters sampled every 5000 generations 10^4 times after discarding 2×10^5 generations as burn-in. Convergence of likelihood and parameters was assessed by examining trace plots using the program Tracer ver. 1.6.0 (Rambaut and Drummond) and comparing results between independent runs.

RESULTS

Sequence attributes

We obtained a total of 6.71×10^7 raw Illumina reads with each individual yielding an average of 4.47×10^6 reads (Table 3.1). For each individual, an average of 8337 contigs were assembled, of which roughly half corresponded to UCE loci. The average UCE contig length was 865 bp with an average coverage of 40×. The incomplete dataset included data from 4,011 loci with a total alignment length of 3,339,138 bp, yielding a mean locus length of 832.50. Nucleotide data were present for 88% of the data matrix. On the other hand, the complete dataset included data from 1,828 loci with a total alignment length of 1,653,366 bp, resulting in an average locus length of 904.47. In this data matrix, 92% represented nucleotide data. UCE sequence data are available at GenBank (accession numbers to be provided upon acceptance of manuscript).

Phylogenetic relationships

Bayesian, ML, and GCMs yielded the same estimate of topology with high support in most nodes (Fig. 3.1). The earliest split in crown-clade trogons involves the African genus *Apaloderma*, which is sister to two radiations, one in Asia and the other in the New World tropics, each of which is monophyletic. Within the Asian clade *Apalharpactes* is sister to *Harpactes*. In the New World, the quetzal genera *Euptilotis* and *Pharomachrus* form a clade sister to *Trogon* and *Priotelus*. With two exceptions, all nodes in the ingroup had 100% bootstrap and Bayesian posterior-probability support across the different analytical approaches, and across the subsets of loci. The node that unites the Asian and New World clades (Node 1, Fig. 3.1) received 100% Bayesian posterior-probability and ML bootstrap support, and 91–97% bootstrap support across GCM results based on all 1828 loci. In constrast, the node that unites the Asian genera *Apalharpactes* and *Harpactes* (Node 2, Fig. 3.1) received 100% Bayesian posterior-probability, 61% ML bootstrap support, and 60–83% bootstrap support across GCM analysis on all loci. GCM analyses performed on subsets of informative loci yielded the same ingroup topology (not shown) but with lower support values for Nodes 1 and 2 (Fig. 3.2). Bootstrap support for Node 2 dropping below 50% for STEAC, MP-EST, and ASTRAL when the 741 most informative loci were analyzed. Support values for the other nodes in the ingroup remained constant at 100% in these sets of GCM analyses.

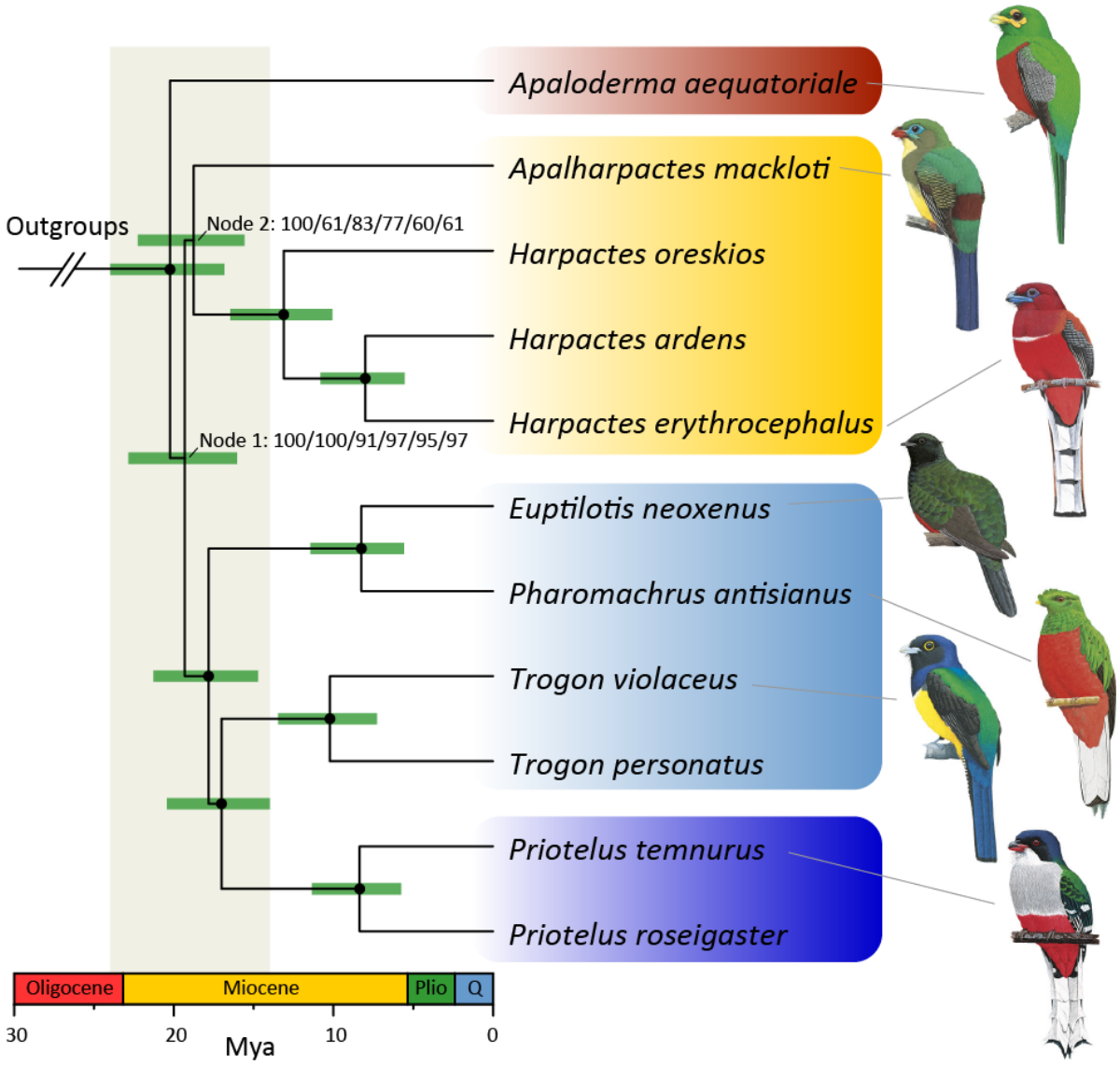


Figure 3.1. Trogonidae generic level phylogeny. Estimate of phylogenetic relationships of trogons based on concatenated and coalescent approaches. Nodes with dots correspond to 100% Bayesian posterior probability and bootstrap support from all methods of analysis. Numbers next to nodes indicate support from Bayesian, ML, STAR, STEAC, MP-EST, and ASTRAL analyses, respectively. Chronogram based on divergence time estimation with MCMCTree. Color highlights on species names correspond to different geographic regions: red = Africa, yellow = Asia, light blue = continental Neotropics, dark blue = Caribbean. Light gray shading shows short time window of diversification among continents. Abbreviations in geologic time scale are Plio = Pliocene, Q = Quaternary.

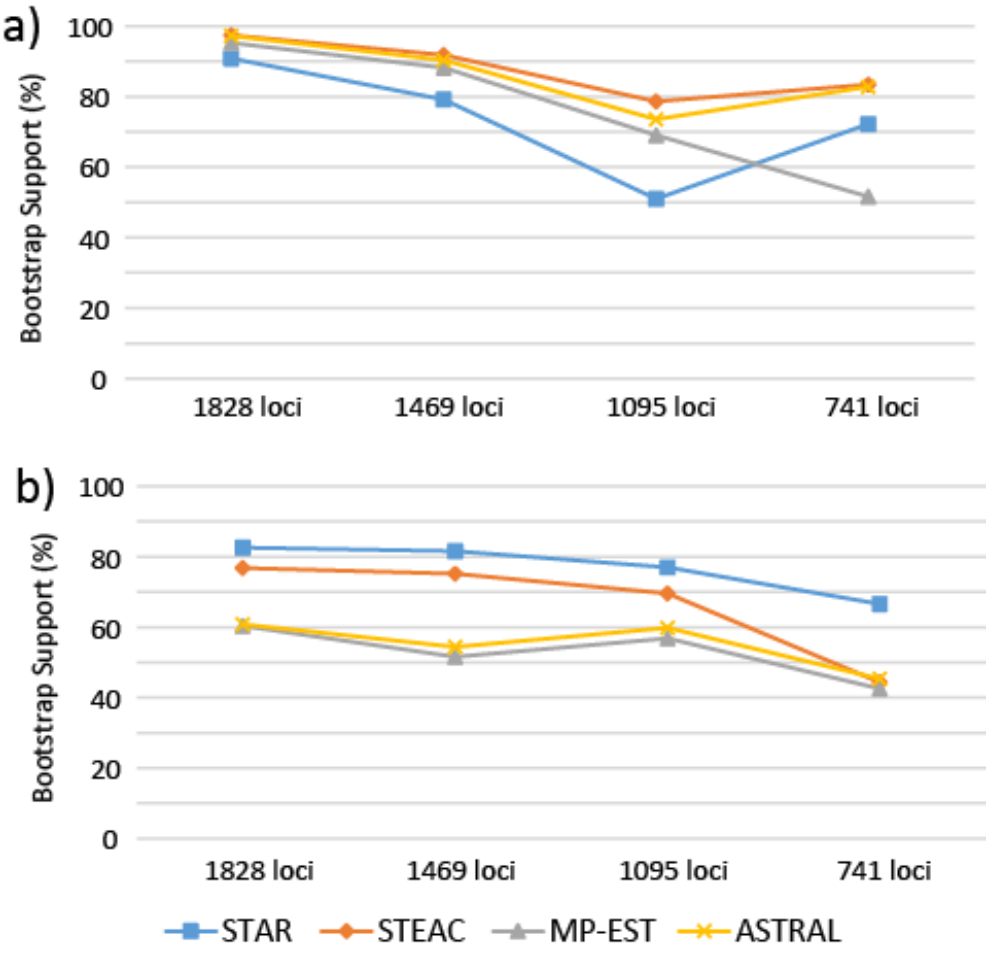


Figure 3.2. Effect of the number of loci analyzed on bootstrap support values. a) Support values for Node 1 in Figure 1. b) Support values for Node 2 in Figure 1.

The trogon family was found to be sister to a clade consisting of Bucerotiformes (hornbills) and Coraciiformes (rollers, bee-eaters and allies; Fig. 3.1) in all analyses except in STAR, which placed the trogons in a three-way polytomy with Bucerotiformes and Coraciiformes. These clades were sister to Leptosomiformes (cuckoo-roller) and in turn sister to Coliiformes (mousebirds) based on most phylogenetic analyses. STEAC, however, recovered a

topology in which the placement of Leptosomiformes and Coliiformes was switched with high support.

Almost all analyses using smaller sets of outgroup taxa produced the same rooting of trogon phylogeny with high support, that is, along the edge that leads to the African lineage *Apaloderma* (not shown), similar to the result when all five outgroup taxa were used. The only exception, the case in which *Otus elegans* was used as the single outgroup, rooted the tree along the edge that connected a clade containing *Apaloderma* and *Apalharpactes* relative to another clade consisting of all the other trogon genera (not shown). This alternative root presented a different ingroup topology as well but the alternative clades received little bootstrap support values (< 51%).

DISCUSSION

Genomic data and phylogenetic inference

This study joins a growing number that uses hundreds to thousands of loci to clarify phylogenetic relationships that have previously been difficult to resolve with fewer data. The congruence of results among various concatenation and coalescent approaches indicates that our estimate of relationships among trogon genera is the most robust to date. The short internodes in the topology represent a soft polytomy, which was resolved with the use of genome-scale data. Although the internodes at the base of the trogon phylogeny are extremely short, they do not appear to be in the anomaly zone. If the outgroup taxa are excluded, the most common gene tree topology inferred from the complete dataset (n = 19) is congruent with the species tree (Table 3.2).

Table 3.2. Distribution of unique gene tree topologies from the complete dataset if outgroup taxa are excluded.

Frequency of gene tree topology	Number of gene tree topologies that have given frequency
1	1267
2	77
3	27
4	11
5	11
6	6
7	2
8	2
9	4
10	3
11	3
12	1
14	1
17	1
19	1*

* Gene tree topology congruent to the inferred species tree topology

Support for most of the nodes in the ingroup was high (>90% bootstrap support or 100% Bayesian posterior probability) across both concatenation and coalescent approaches with one exception. Support for the node uniting the Asian genera *Apalharpactes* and *Harpactes* was low in ML (61%), MP-EST (60%), and ASTRAL (61%), moderate in STEAC (77%) and STAR (83%), and strong in Bayesian inference (100%). The high support value from Bayesian inference compared with those from other approaches is not surprising considering that Bayesian posterior probabilities are generally higher than ML bootstrap support values (Alfaro et al. 2003) and can be excessively liberal (Suzuki et al. 2002). The higher support value in STEAC relative to MP-EST and ASTRAL is notable. This GCM tends to be least affected by biases from short sequence lengths and uninformative loci because of its use of branch length information from gene trees (Chapter 1). Despite some low support values, this study provides some support (>70% from both STAR and STEAC, and 100% from Bayesian analysis) for the position of *Apalharpactes* from both concatenation and coalescent approaches, whereas the only other study

that included this genus was unable to estimate its position with any confidence (Hosner et al. 2010).

The root of the trogon phylogeny along the edge leading to *Apaloderma* appears to be appropriate because this positioning is supported in reconstructions with different numbers and combinations of outgroups. The only alternative root placement occurred when a single distant outgroup, *Otus elegans*, was used, albeit with weak support. It is well known that the use of only distant outgroups can cause incorrect rooting (Wheeler 1990).

The utility of GCMs at deep phylogenetic time scales has recently been questioned based in part on a re-analysis of UCE loci in estimating relationships within placental mammals (Gatesy and Springer 2014). In our study, however, GCMs were effective, probably for a variety of reasons. First, the average locus length in this study was longer (twice that of the mammal dataset in (Gatesy and Springer 2014)). Longer alignments generally improves accuracy of GCMs (Mirarab et al. 2014c), especially for UCE loci because variability in UCE flanking regions increases with increasing distance from the center of the ultraconserved element (Faircloth et al. 2012). Second, more loci were used for GCM analyses in this study than the mammal study (10 × as many). The GCMs used here are statistically consistent, i.e., they are expected to be highly accurate as the number of gene trees analyzed goes to infinity (Liu et al. 2009b, 2010; Mirarab et al. 2014c). Lastly, this study involves a relatively few lineages separated by short internodes. As the number of short internodes in a phylogeny increases, the likelihood and complexity of deep coalescences between lineages also increases.

Results in previous studies indicate that filtering out uninformative loci in GCM analyses can increase bootstrap support values of species tree estimates (Liu et al. 2015; Chapter 1). However, the lower bootstrap values obtained when fewer loci were used in GCM analyses in

this study (although only for Nodes 1 and 2 in Fig. 3.1) indicate the opposite result. These conflicting results can partly be explained by the large discrepancy in information content in loci between the studies. In Chapter 1, in which the oldest divergence between taxa occurred relatively recently (~8 Ma; Moyle et al. 2009), the average number of parsimony-informative characters in each locus was 42.7 bp. In contrast, trogons are an older clade, and a locus in the present study contained an average of 73.5 informative characters. In our study the exclusion of hundreds of loci in GCM analysis had the effect of removing informative gene trees, leading to lower support values, whereas in Chapter 1 the exclusion of many loci resulted in minimizing spurious gene trees, leading to higher support values. Filtering fewer uninformative loci than was done in our study could potentially have produced higher support values for Nodes 1 and 2 but this possibility was not further explored. GCMs are statistically consistent under the multi-species coalescent (Liu et al. 2009b, 2010; Mirarab et al. 2014c) and as such perform better with more loci. Thus the criteria for filtering uninformative loci should consider levels of divergence in the taxa of interest.

Trogon relationships and biogeography

This study is the first to estimate trogon phylogeny based on genome-wide data and the first to produce a well-resolved tree of trogon genera. The main results of our study, demonstrating the monophyly of trogons from each geographic region—Africa, Asia, and the Neotropics—and establishing hierarchical relationships among them, are similar to results obtained in earlier studies (Espinosa de los Monteros 1998; Ornelas et al. 2009), but now with strong nodal support throughout the tree. Our results contradict the outcome of earlier studies that found Asian (Hosner et al. 2010) and New World trogons (Moyle 2005; Hosner et al. 2010) to be paraphyletic. These earlier problems were probably caused by weak conflicting

phylogenetic signal from markers used in these studies. For example, analysis of mitochondrial Cyt-b gene sequences showed the African genus *Apaloderma* to be the sister of all other trogons, with low support (Espinosa de los Monteros 1998; Johansson and Ericson 2005). Another mitochondrial gene, ND2, and the nuclear RAG-1 gene each indicated *Priotelus* and the quetzal genera *Euptilotis* and *Pharomachrus* as the first lineages to sequentially diverge from the base of the tree, again with weak support (Moyle 2005). Interestingly, the study that combined all these markers arrived at a topology similar to ours (Ornelas et al. 2009), suggesting that more data is better than fewer.

Our dating analysis estimated the trogon lineages from Africa, Asia, and the New World diverged rapidly within a few million years in the Early Miocene (Fig. 3.1). The divergence of the Caribbean lineage from the continental lineage in the New World is estimated to have taken place 1–2 million years later. These age estimates are within the confidence intervals of previous estimates (Espinosa de los Monteros 1998; Moyle 2005) but are on the younger end of these ranges. The younger date estimates in this study are not surprising considering that our calibrations are based on a study that, in general, placed a younger date of diversification of modern birds than previously thought (Jarvis et al. 2014). As found in two previous trogon studies (Espinosa de los Monteros 1998; Moyle 2005), a Gondwanan origin for the crown group can be rejected based on our date-estimates and a Laurasian origin appears to be a more plausible scenario.

With respect to the placement of trogons among related bird orders, our results are similar to those of recent phylogenomic studies of birds (Hackett et al. 2008; Jarvis et al. 2014). The trogons are members of the clade that includes Bucerotiformes, Coraciiformes, and Piciformes (App. 3.1). Bucerotiformes presently occurs in the Old World tropics, whereas

Coraciiformes and Piciformes each has a contemporary pan-tropical distribution. However, Bucerotiform fossils from the Miocene and Eocene are known from Europe; Eocene/Oligocene Coraciiform fossils are also known from Europe; and Piciforms are known from the Eocene of North America and the Miocene of Europe (Brodkorb 1971). The groups closely related to the clade comprising Trogoniformes, Bucerotiformes, Coraciiformes, and Piciformes include Leptosomiformes, a Malagasy endemic order, and Coliiformes, an African endemic order. However, the fossil record indicates that these two orders had much wider distributions in North America and Europe during the Eocene (Weidig 2006; Ksepka and Clarke 2009). All unambiguously identified trogon fossils from the Paleogene are from Europe (e.g., Kristoffersen 2002; Mayr 2005, 2009) and these fossil taxa are viewed as successive sister taxa of crown trogons (Mayr 2009). Thus Mayr (2009) believed that the fossil record is consistent with an Old World origin of crown trogons. Our phylogenetic estimate of trogon relationships is also consistent with this view. However, Mayr (2009) also cautioned that an unpublished record of a putative stem trogon from Early Eocene North America exists (Weidig 2003), albeit from a partial skeleton, which if confirmed can possibly erode support for an Old World origin hypothesis for this group.

The timing of trogon diversification is much younger than the Early Paleogene opening of the North Atlantic Ocean (Courtillot et al. 1999), precluding the North Atlantic land bridge playing a role in crown trogon diversification. However, Western North America and Eastern Asia were connected by the Beringian land bridge from the mid-Cretaceous until the Pliocene (see summary in Sanmartín et al. 2001). In the Eocene, when climates were much warmer than the present, a belt of boreotropical forest stretched over this land bridge permitting the exchange of terrestrial fauna and flora. By the Oligocene, global climates began cooling and the

boreotropical forest across Beringia were replaced by boreal forest vegetation. Evidence exists for a short bout of Late Oligocene warming, followed by a period of cooling in the Early Miocene, before another climatic optimum in the mid-Miocene (Zachos et al. 2001). Our estimates of crown trogon rapid diversification fall within these last two climatic optima. Coincidentally, the African continent established a definitive connection with Eurasia through the Middle East in the Early Miocene (Gheerbrant and Rage 2006). Thus geologic and climate data, along with our divergence date estimates, support the view that crown trogons assumed a wide distribution during the Early Miocene, likely crossing through the Beringian land bridge and the African-Eurasian collision zone. The aridification of North Africa and the Middle East, along with continued increase of the latitudinal temperature gradient in the Northern Hemisphere would have isolated trogons in African, SE Asia, and the New World. This hypothesis assumes that suitable vegetation permitted the dispersion of trogons between Eurasia, North America, and Africa. The presence of Early Miocene trogon fossils in Europe (Brodkorb 1971) indicate that early trogons persisted during this period in higher latitudes, possibly including the Beringian land bridge.

A review of the origins of Caribbean vertebrates suggests colonization occurred via overwater dispersal for most lineages (Hedges 1996). A similar study with a focus on mammals indicate that the Gaarlandia land connection to South America in the Late Eocene/Early Oligocene played a role in the establishment of some mammal lineages in the Caribbean (Dávalos 2004). No land bridges are known to have connected the Caribbean islands with North America in the Miocene (Iturralde-Vinent 2006). Thus long distance dispersal may have played a role in the colonization of the Caribbean by ancestors of the genus *Priotelus*. This finding seems counterintuitive given that trogons are considered poor dispersers (Moyle 2005).

However, trogons have colonized islands in the Philippines that have no known land connections to the Asian continent (Hall 2002), so trogon dispersal ability may be underestimated.

LITERATURE CITED

Aberer A.J., Kobert K., Stamatakis A. 2014. ExaBayes: massively parallel bayesian tree inference for the whole-genome era. Mol. Biol. Evol. 31:2553–6.

Aggerbeck M., Fjeldså J., Christidis L., Fabre P.H., Jønsson K.A. 2014. Resolving deep lineage divergences in core corvoid passerine birds supports a proto-Papuan island origin. Mol. Phylogenet. Evol. 70:272–285.

Alfaro M.E., Zoller S., Lutzoni F. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. 20:255–266.

Alstrom P., Jonsson K.A., Fjeldsa J., Odeen A., Ericson P.G.P., Irestedt M. 2015. Dramatic niche shifts and morphological change in two insular bird species. R. Soc. Open Sci. 2:140364.

Barker F.K., Barrowclough G.F., Groth J.G. 2002. A phylogenetic hypothesis for passerine birds: taxonomic and biogeographic implications of an analysis of nuclear DNA sequence data. Proc. R. Soc. B Biol. Sci. 269:295–308.

Barker F.K., Cibois A., Schikler P., Feinstein J., Cracraft J. 2004. Phylogeny and diversification of the largest avian radiation. Proc. Natl. Acad. Sci. U. S. A. 101:11040–5.

Bayzid M.S., Hunt T., Warnow T. 2014. Disk covering methods improve phylogenomic analyses. BMC Genomics. 15:S7.

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

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.

Brandley M.C., Bragg J.G., Singhal S., Chapple D.G., Jennings C.K., Lemmon A.R., Lemmon E.M., Thompson M.B., Moritz C. 2015. Evaluating the performance of anchored hybrid enrichment at the tips of the tree of life: a phylogenetic analysis of Australian *Eugongylus* group scincid lizards. BMC Evol. Biol. 15:62.

Brodkorb P. 1971. Catalogue of fossil birds: part 4 (Columbiformes through Piciformes). Bull. Florida State Museum, Biol. Sci. 15:163–266.

Castresana J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Mol. Biol. Evol. 17:540–552.

Chaudhary R., Fernandez-Baca D., Burleigh J.G. 2014. MulRF: a software package for phylogenetic analysis using multi-copy gene trees. Bioinformatics. 31:432–433.

Cibois A. 2003. Sylvia is a babbler: taxonomic implications for the families Sylviidae and Timaliidae. Bull. Ornithol. Club. 123:257–260.

Courtillot V., Jaupart C., Manighetti I., Tapponnier P., Besse J. 1999. On casual links between flood basalts and continental breakup. Earth Planet. Sci. Lett. 166:177–195.

Cox S.C., Prys-Jones R.P., Habel J.C., Amakobe B. a., Day J.J. 2014. Niche divergence promotes rapid diversification of East African sky island white-eyes (Aves: Zosteropidae). Mol. Ecol. 23:4103–4118.

Cracraft J. 1973. Continental drift, paleoclimatology, and the evolution and biogeography of birds. J. Zool. 169:455–545.

Cracraft J. 2001. Avian evolution, Gondwana biogeography and the Cretaceous-Tertiary mass extinction event. Proc. Biol. Sci. 268:459–469.

Cracraft J. 2014. Avian higher-level relationships and classification: Passeriforms. In: Dickinson E., Christidis L., editors. The Howard & Moore Complete Checklist of the Birds of the World. 4th Edition. Vol. 2. Eastbourne, U.K.: p. xvii–xlv.

Crawford N.G., Faircloth B.C., McCormack J.E., Brumfield R.T., Winker K., Glenn T.C. 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. 8:783–6.

Crawford N.G., Faircloth B.C. 2014. CloudForest: Code to calculate species trees from large genomic datasets. DOI:10.5281/zenodo.12259.

DaCosta J.M., Klicka J. 2008. The Great American Interchange in birds: a phylogenetic perspective with the genus Trogon. Mol. Ecol. 17:1328–43.

Dávalos L.M. 2004. Phylogeny and biogeography of Caribbean mammals. Biol. J. Linn. Soc. 81:373–394.

Degnan J.H., Rosenberg N. a. 2006. Discordance of Species Trees with Their Most Likely Gene Trees. PLoS Genet. 2:e68.

Dickinson E., Christidis L. 2014. The Howard & Moore complete checklist of the birds of the world. 4th Edition, Vol. 2. Eastbourne, U.K.: Aves Press.

Edwards S. V. 2009. Is a new and general theory of molecular systematics emerging? Evolution (N. Y). 63:1–19.

Ericson P.G.P., Christidis L., Cooper A., Irestedt M., Jackson J., Johansson U.S., Norman J.A. 2002. A Gondwanan Origin of Passerine Birds Supported by DNA Sequences of the Endemic New Zealand Wrens. Proc. R. Soc. B Biol. Sci. 269:235–241.

Espinosa de los Monteros A. 1998. Phylogenetic relationships among the trogons. Auk. 115:937–954.

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. 61:717–26.

Faircloth B.C., Sorenson L., Santini F., Alfaro M.E. 2013. A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs). PLoS One. 8:e65923.

Faircloth B.C. 2014. phyluce: phylogenetic estimation from ultraconserved elements. DOI:10.6079/J9PHYL.

Gatesy J., Springer M.S. 2014. Phylogenetic Analysis at Deep Timescales: Unreliable Gene Trees, Bypassed Hidden Support, and the Coalescence/Concatalescence Conundrum. Mol. Phylogenet. Evol. 80:231–266.

Gelang M., Cibois A., Pasquet E., Olsson U., Alström P., Ericson P.G.P. 2009. Phylogeny of babblers (Aves, Passeriformes): major lineages, family limits and classification. Zool. Scr. 38:225–236.

Gheerbrant E., Rage J.C. 2006. Paleobiogeography of Africa: How distinct from Gondwana and Laurasia? Palaeogeogr. Palaeoclimatol. Palaeoecol. 241:224–246.

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–652.

Hackett S.J., Kimball R.T., Reddy S., Bowie R.C.K., Braun E.L., Braun M.J., Chojnowski J.L., Cox W.A., Han K.-L., Harshman J., Huddleston C.J., Marks B.D., Miglia K.J., Moore W.S., Sheldon F.H., Steadman D.W., Witt C.C., Yuri T. 2008. A phylogenomic study of birds reveals their evolutionary history. Science. 320:1763–8.

Hall R. 2002. Cenozoic geological and plate tectonic evolution of SE Asia and the SW Pacific: computer-based reconstructions, model and animations. J. Asian Earth Sci. 20:353–431.

Hedges S.B. 1996. Historical Biogeography of West Indian Vertebrates. Annu. Rev. Ecol. Syst. 27:163–196.

Hosner P.A., Sheldon F.H., Lim H.C., Moyle R.G. 2010. Phylogeny and biogeography of the Asian trogons (Aves: Trogoniformes) inferred from nuclear and mitochondrial DNA sequences. Mol. Phylogenet. Evol. 57:1219–25.

Hovmöller R., Lacey Knowles L., Kubatko L.S. 2013. Effects of missing data on species tree estimation under the coalescent. Mol. Phylogenet. Evol. 69:1057–1062.

Iturralde-Vinent M.A. 2006. Meso-Cenozoic Caribbean paleogeography: implications for the historical biogeography of the region. Int. Geol. Rev. 48:791–827.

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., 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., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., 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., Alfaro-Nunez 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., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., 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. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science (80). 346:1320–1331.

Jetz W., Thomas G.H.H., Joy J.B.B., Hartmann K., Mooers a. O.O. 2012. The global diversity of birds in space and tim. Nature. 491:1–5.

Johansson U.S., Ericson P.G.P. 2005. A re-evaluation of basal phylogenetic relationships within trogons (Aves: Trogonidae) based on nuclear DNA sequences. J. Zool. Syst. Evol. Res. 43:166–173.

Jønsson K. a, Fabre P.-H., Ricklefs R.E., Fjeldså J. 2011. Major global radiation of corvoid birds originated in the proto-Papuan archipelago. Proc. Natl. Acad. Sci. U. S. A. 108:2328–2333.

Jønsson K. a., Bowie R.C.K., Nylander J. a a, Christidis L., Norman J. a., Fjeldså J. 2010. Biogeographical history of cuckoo-shrikes (Aves: Passeriformes): Transoceanic colonization of Africa from Australo-Papua. J. Biogeogr. 37:1767–1781.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–80.

Kristoffersen A. V. 2002. An early Paleogene trogon (Aves: Trogoniformes) from the Fur Formation, Denmark. J. Vertebr. Paleontol. 22:661–666.

Ksepka D.T., Clarke J. a. 2009. Affinities of Palaeospiza bella and the phylogeny and biogeography of mousebirds (Coliiformes). Auk. 126:245–259.

Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics. 25:971–973.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.

Lanyon S.M. 1988. The stochastic mode of molecular evolution: what consequences for systematic investigations? Auk. 105:565–573.

Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61:727–44.

Liu L., Xi Z., Wu S., Davis C., Edwards S. V. 2015. Estimating phylogenetic trees from genome-scale data. Ann. N. Y. Acad. Sci. DOI:10.1111/nyas.12747.

Liu L., Yu L., Edwards S. V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10:302.

Liu L., Yu L., Kubatko L., Pearl D.K., Edwards S. V. 2009a. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53:320–328.

Liu L., Yu L., Pearl D.K., Edwards S. V. 2009b. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58:468–77.

Liu L., Yu L. 2011. Estimating species trees from unrooted gene trees. Syst. Biol. 60:661–667.

Mayr E. 1946. History of the North American bird fauna. Wilson Bull. 58:3–41.

Mayr G., Smith T. 2013. Galliformes, Upupiformes, Trogoniformes, and other avian remains (?Phaethontiformes and ?Threskiornithidae) from the Rupelian stratotype in Belgium, with comments on the identity of “Anas” benedeni Sharpe, 1899. Proc. 8th Int. Meet. Soc. Avian Paleontol. Evol.:23–36.

Mayr G. 1999. A new trogon from the Middle Oligocene of Céreste, France. Auk. 116:427–434.

Mayr G. 2003. On the phylogenetic relationships of trogons (Aves, Trogonidae). J. Avian Biol. 34:81–88.

Mayr G. 2005. New trogons from the early Tertiary of Germany. Ibis (Lond. 1859). 147:512–518.

Mayr G. 2009. A well-preserved second trogon skeleton (Aves, Trogonidae) from the Middle Eocene of Messel, Germany. Palaeobiodiversity and Palaeoenvironments. 89:1–6.

McCormack J.E., Faircloth B.C., Crawford N.G., Gowaty P.A., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res. 22:746–54.

McCormack J.E., Harvey M.G., Faircloth B.C., Crawford N.G., Glenn T.C., Brumfield R.T. 2013. A phylogeny of birds based on over 1,500 Loci collected by target enrichment and high-throughput sequencing. PLoS One. 8:e54848.

Mirarab S., Bayzid M.S., Boussau B., Warnow T. 2014a. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science. 346:1250463.

Mirarab S., Bayzid M.S., Warnow T. 2014b. Evaluating Summary Methods for Multilocus Species Tree Estimation in the Presence of Incomplete Lineage Sorting. Syst. Biol. 0:1–15.

Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014c. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics. 30:i541–8.

Mossel E., Roch S. 2010. Incomplete lineage sorting: Consistent phylogeny estimation from multiple loci. IEEE/ACM Trans. Comput. Biol. Bioinforma. 7:166–171.

Moyle R.G., Andersen M.J., Oliveros C.H., Steinheimer F.D., Reddy S. 2012. Phylogeny and biogeography of the core babblers (Aves: Timaliidae). Syst. Biol. 61:631–51.

Moyle R.G., Filardi C.E., Smith C.E., Diamond J.M. 2009. Explosive Pleistocene diversification and hemispheric expansion of a “great speciator.” Proc. Natl. Acad. Sci. U. S. A. 106:1863–1868.

Moyle R.G. 2005. Phylogeny and biogeographical history of Trogoniformes, a pantropical bird order. Biol. J. Linn. Soc. 84:725–738.

Norman J.A., Ericson P.G.P., Jønsson K.A., Fjeldså J., Christidis L. 2009. A multi-gene phylogeny reveals novel relationships for aberrant genera of Australo-Papuan core Corvoidea and polyphyly of the Pachycephalidae and Psophodidae (Aves: Passeriformes). Mol. Phylogenet. Evol. 52:488–97.

Nyári Á.S., Joseph L. 2013. Comparative phylogeography of Australo-Papuan mangrove-restricted and mangrove-associated avifaunas. Biol. J. Linn. Soc. 109:574–598.

Olson S.L. 1989. Aspects of global avifaunal dynamics during the Cenozoic. Acta XIX Congr. Int. Ornithol. 2:2023–2029.

Ornelas J.F., González C., Espinosa de los Monteros a. 2009. Uncorrelated evolution between vocal and plumage coloration traits in the trogons: a comparative study. J. Evol. Biol. 22:471–84.

Rambaut A., Drummond A.J. 2007. Tracer v.1.5. http://tree.bio.ed.ac.uk/software/tracer/.

Dos Reis M., Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol. Biol. Evol. 28:2161–72.

Roch S., Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor. Popul. Biol. 100:56–62.

Sanmartín I., Enghoff H., Ronquist F. 2001. Patterns of animal dispersal, vicariance and diversification in the Holarctic. Biol. J. Linn. Soc. 73:345–390.

Selvatti A.P., Gonzaga L.P., Russo C.A.D.M. 2015. A Paleogene origin for crown passerines and the diversification of the Oscines in the New World. Mol. Phylogenet. Evol. 88:1–15.

Seo T.K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol. Biol. Evol. 25:960–971.

Song S., Liu L., Edwards S. V., Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. 109:14942–14947.

Springer M.S., Gatesy J. 2014. Land plant origins and coalescence confusion. Trends Plant Sci. 19:267–269.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–3.

Sukumaran J., Holder M.T. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics. 26:1569–71.

Sun K., Meiklejohn K.A., Faircloth B.C., Glenn T.C., Braun E.L., Kimball R.T. 2014. The evolution of peafowl and other taxa with ocelli (eyespots): a phylogenomic approach. Proc. Biol. Sci. 281.

Suzuki Y., Glazko G. V., Nei M. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. 99:16138–16143.

Thomson R.C., Shedlock A.M., Edwards S. V, Shaffer H.B. 2008. Developing markers for multilocus phylogenetics in non-model organisms: A test case with turtles. Mol. Phylogenet. Evol. 49:514–25.

Toews D.P.L., Campagna L., Taylor S.A., Balakrishnan C.N., Baldassere D.T., Deane-Coe P.E., Harvey M.G., Hooper D.M., Irwin D.E., Judy C.D., Mason N.A., McCormack J.E., McCracken K.G., Oliveros C.H., Safran R.J., Scordato E.S.C., Stryjewski K.F., Tigano A., Uy J.A.C., Winger B. In review. Genomic approaches to understanding the early stages of population divergence and speciation in birds. Auk.

Tonini J., Moore A., Stern D., Shcheglovitova M., Ortí G. 2015. Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions. PLOS Curr. Tree Life.:1–14.

Warren B.H., Bermingham E., Prŷs-Jones R., Thébaud C. 2006. Immigration, species radiation and extinction in a highly diverse songbird lineage: white-eyes on Indian Ocean islands. Mol. Ecol. 15:3769–86.

Weidig I. 2003. Fossil birds from the Lower Eocene Green River Formation (USA). Ph.D. Thesis. Johann-Wolfgang-Goethe-Universität.

Weidig I. 2006. The first New World occurrence of the Eocene bird Plesiocathartes (Aves: ?Leptosomidae). Paläontologische Zeitschrift. 80:230–237.

Wheeler W.C. 1990. Nucleic acid sequence phylogeny and random outgroups. Cladistics. 6:363–367.

Whitfield J.B., Lockhart P.J. 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22:258–65.

Wu Y. 2012. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evolution (N. Y). 66:763–775.

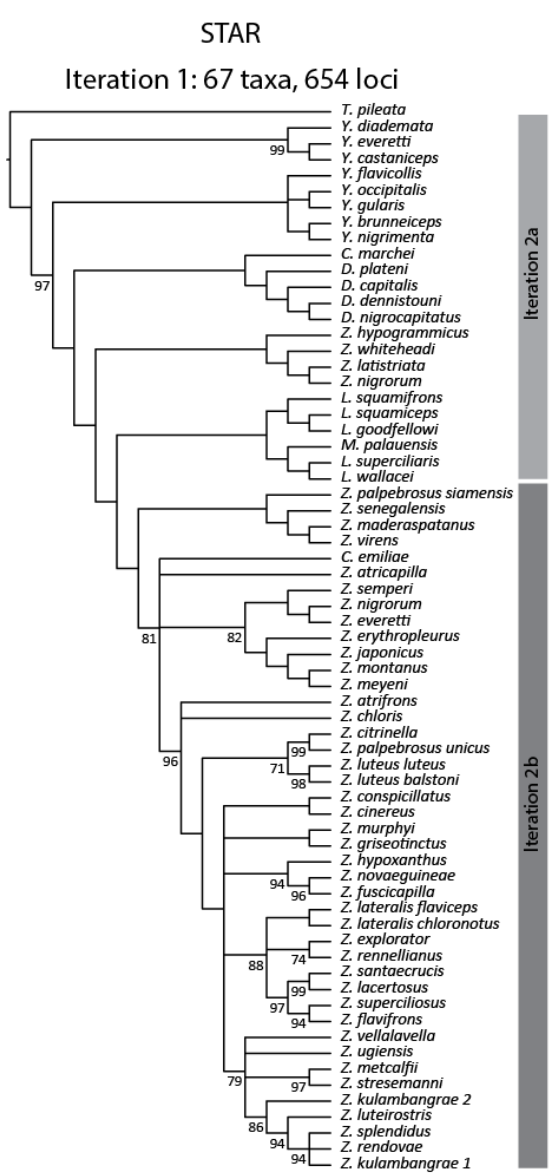
Zachos J., Pagani M., Sloan L., Thomas E., Billups K. 2001. Trends, rhythms, and aberrations in global climate 65 Ma to present. Science. 292:686–694.

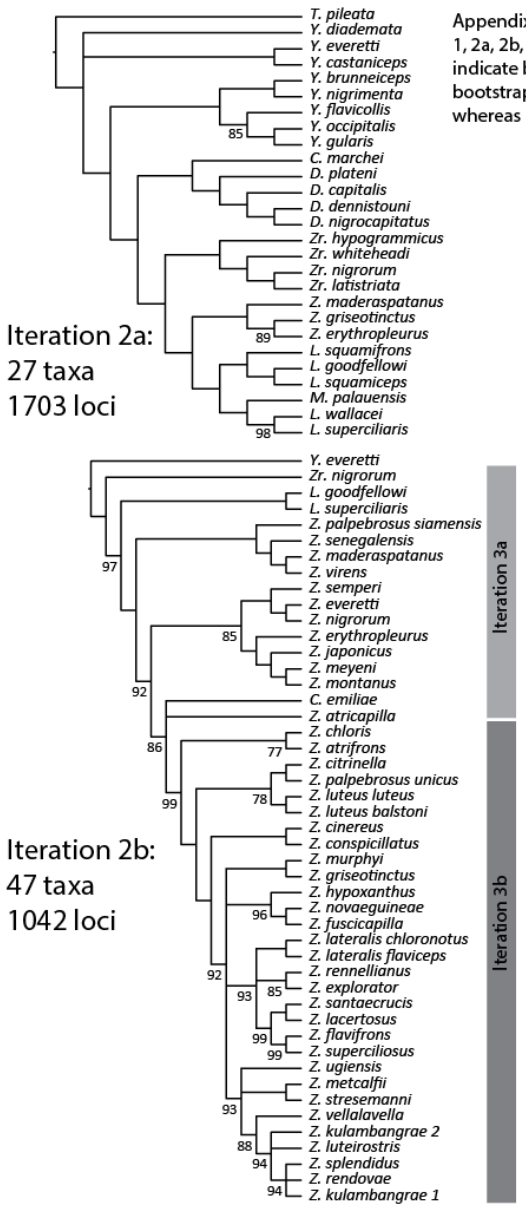
Zhong B., Liu L., Penny D. 2014. The multispecies coalescent model and land plant origins: A reply to Springer and Gatesy. Trends Plant Sci. 19:270–272.

Zhong B., Liu L., Yan Z., Penny D. 2013. Origin of land plants using the multispecies coalescent model. Trends Plant Sci. 18:492–495.

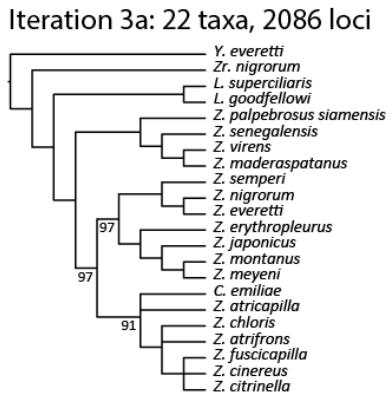
APPENDICES

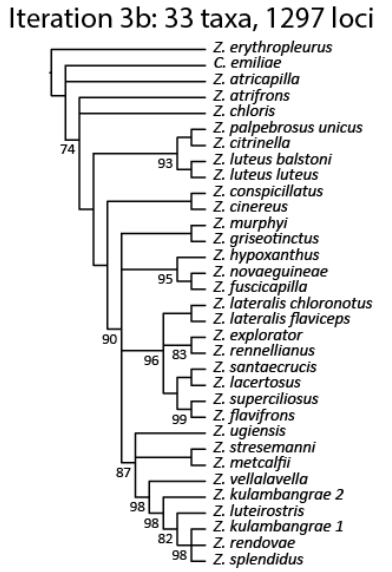
(left intentionally blank)

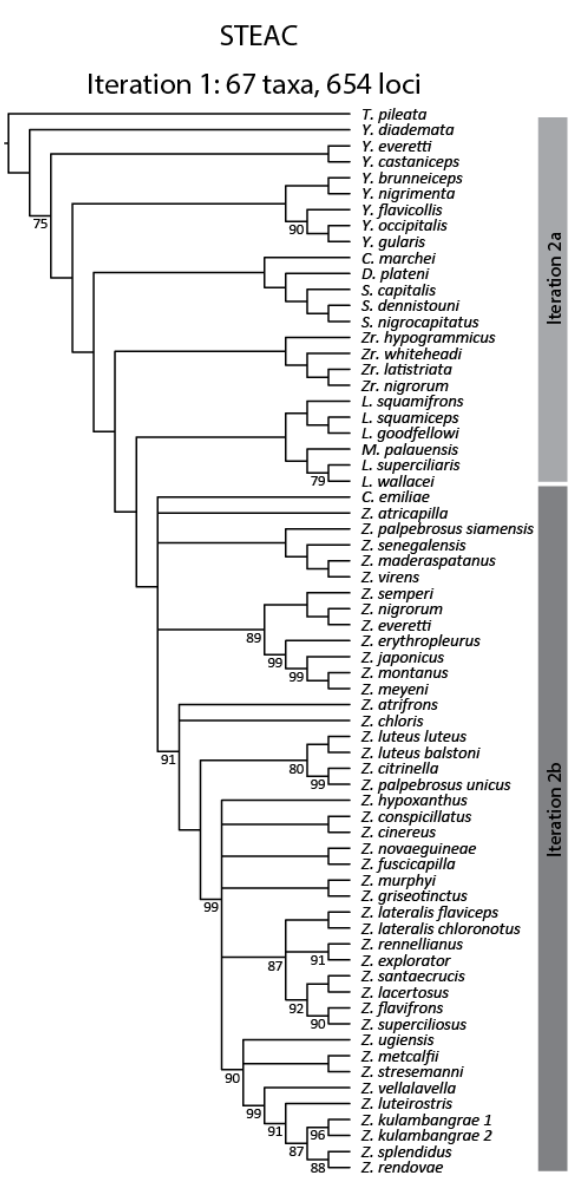


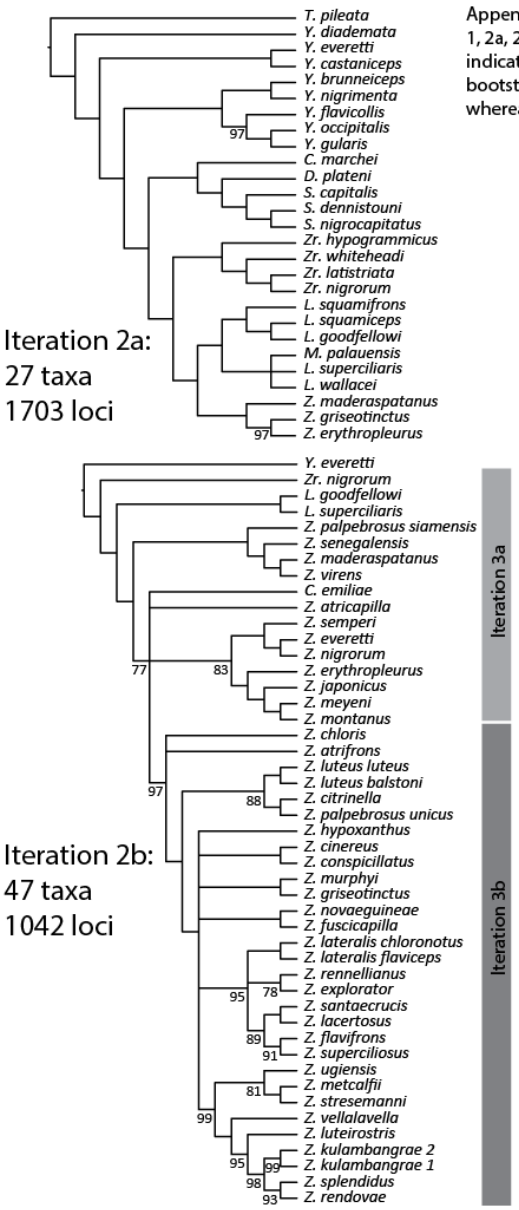


Appendix 1.1. Species tree estimates of STAR for Iterations 1, 2a, 2b, 3a, 3b using all loci. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed, whereas nodes with no labels indicate 100% BS.

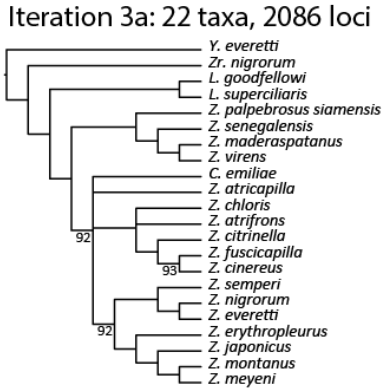


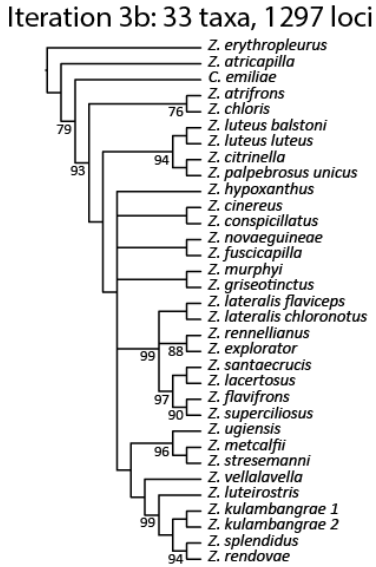


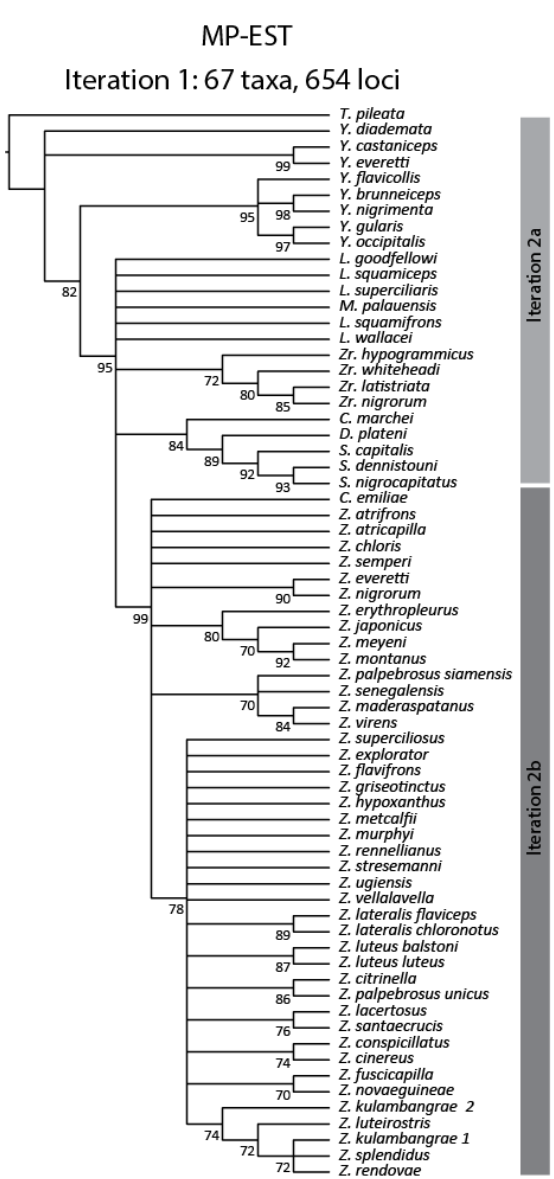


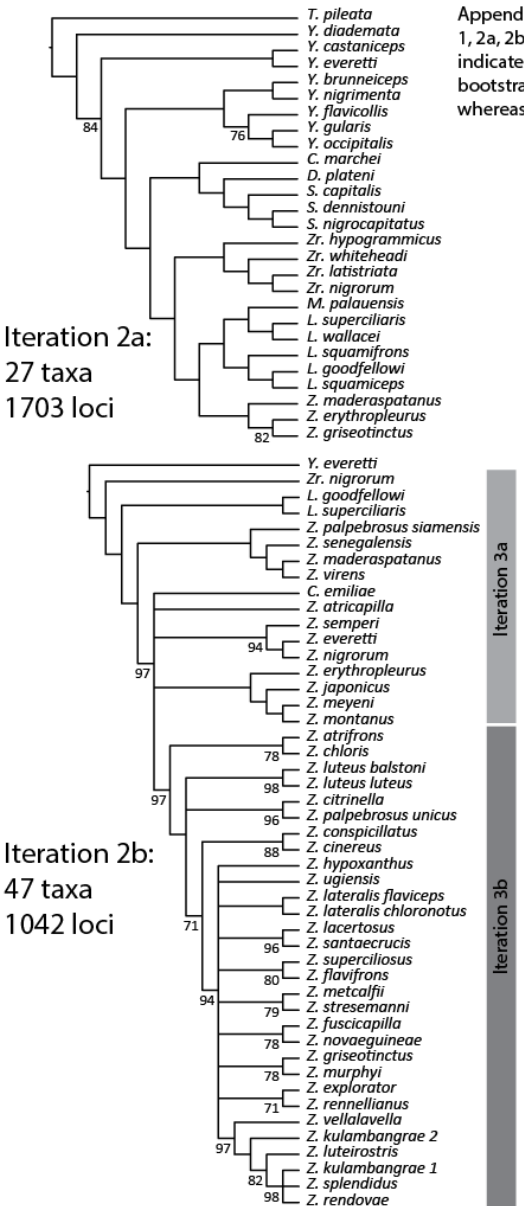


Appendix 1.2. Species tree estimates of STEAC for Iterations 1, 2a, 2b, 3a, 3b using all loci. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed, whereas nodes with no labels indicate 100% BS.

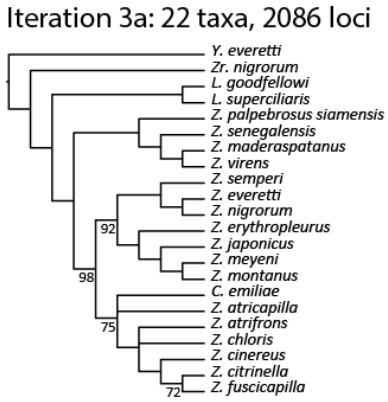


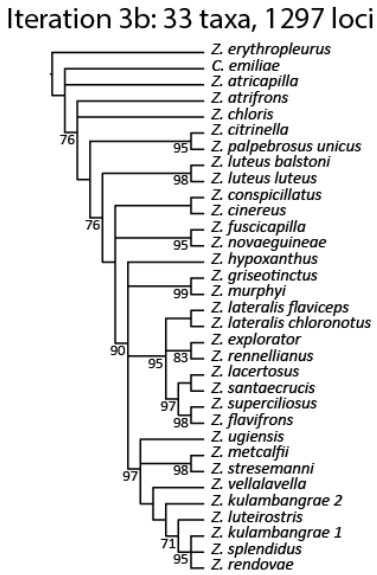


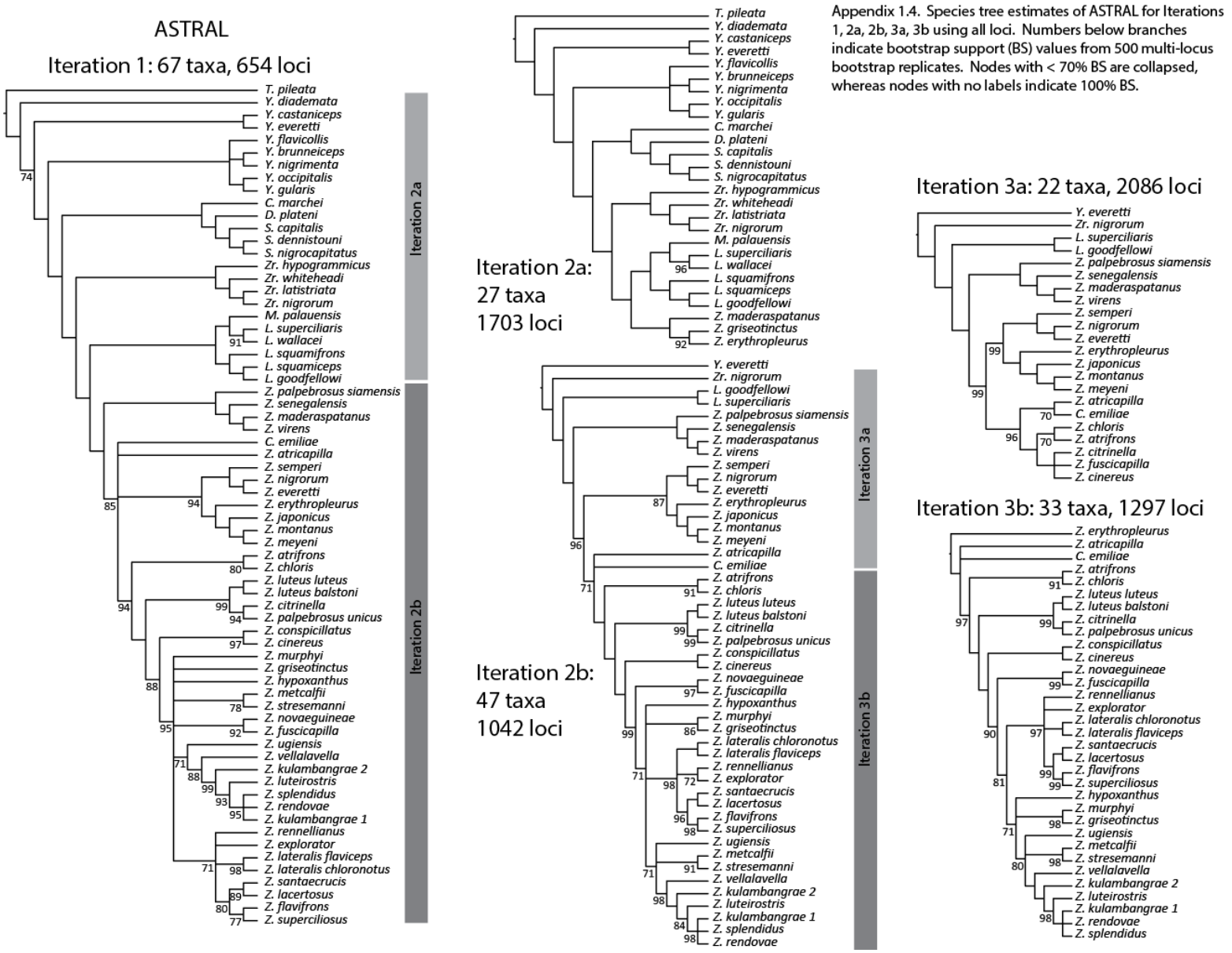




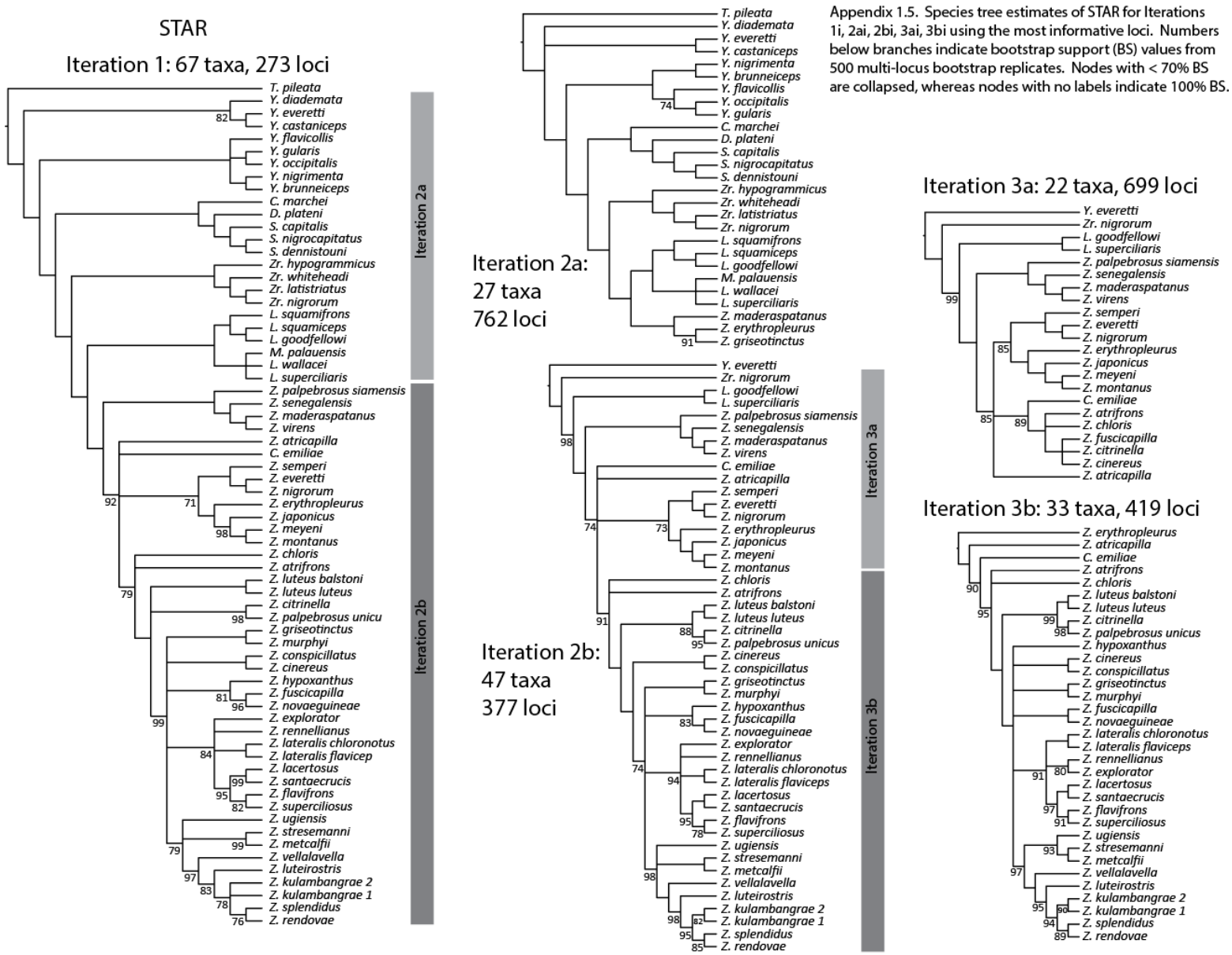
Appendix 1.3. Species tree estimates of MP-EST for Iterations 1, 2a, 2b, 3a, 3b using all loci. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed, whereas nodes with no labels indicate 100% BS.



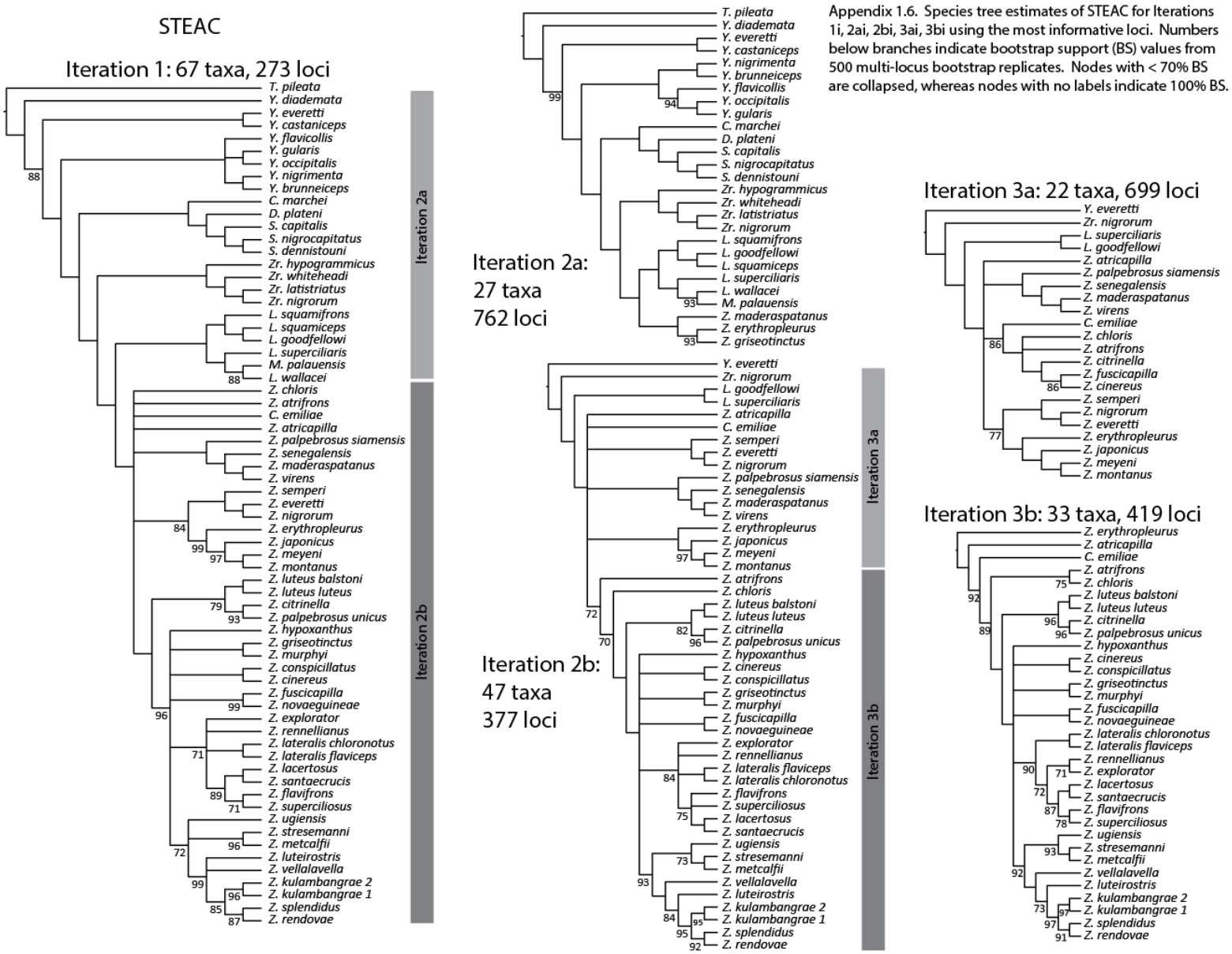




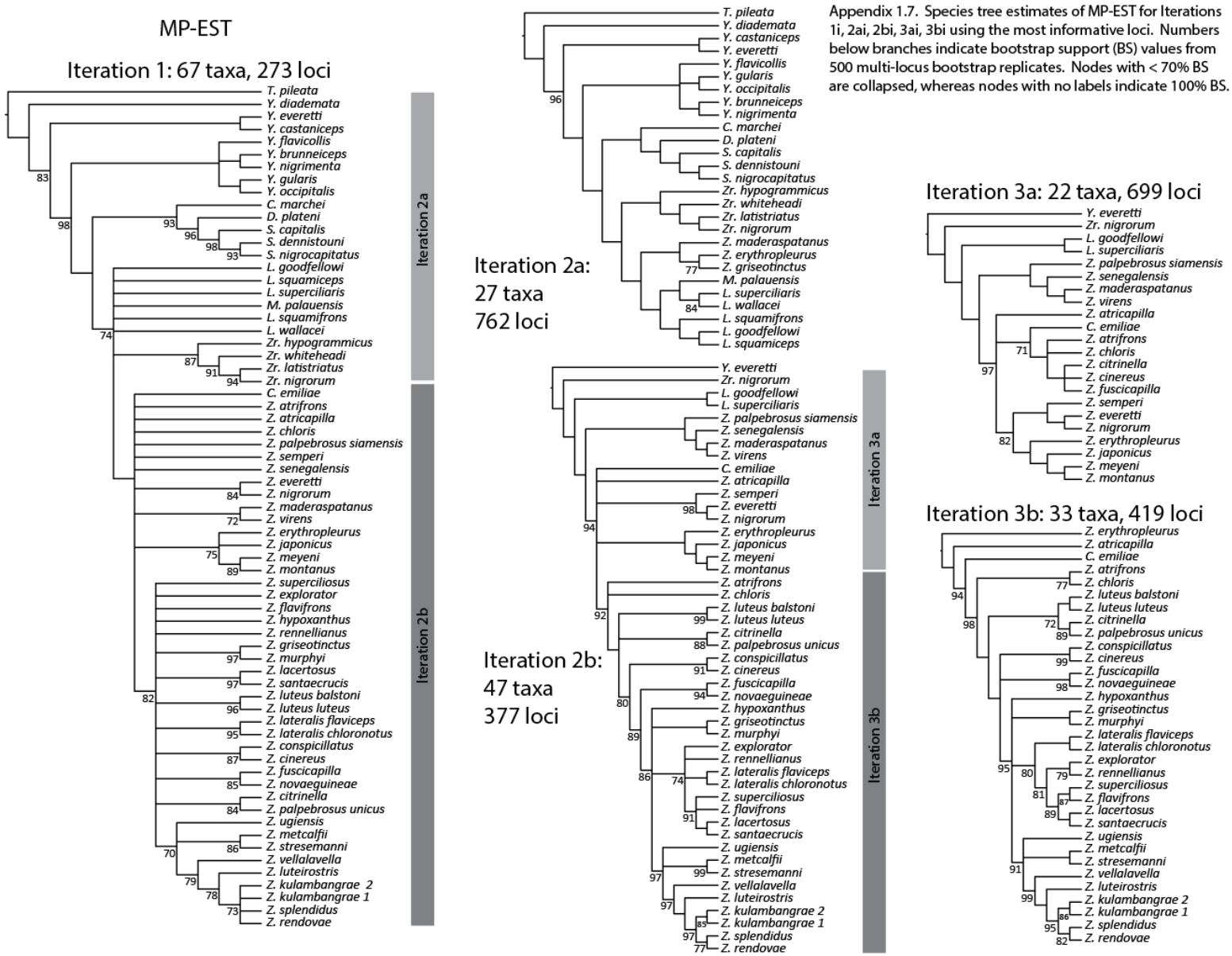
Appendix 1.4. Species tree estimates of ASTRAL for Iterations 1, 2a, 2b, 3a, 3b using all loci. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed, whereas nodes with no labels indicate 100% BS.



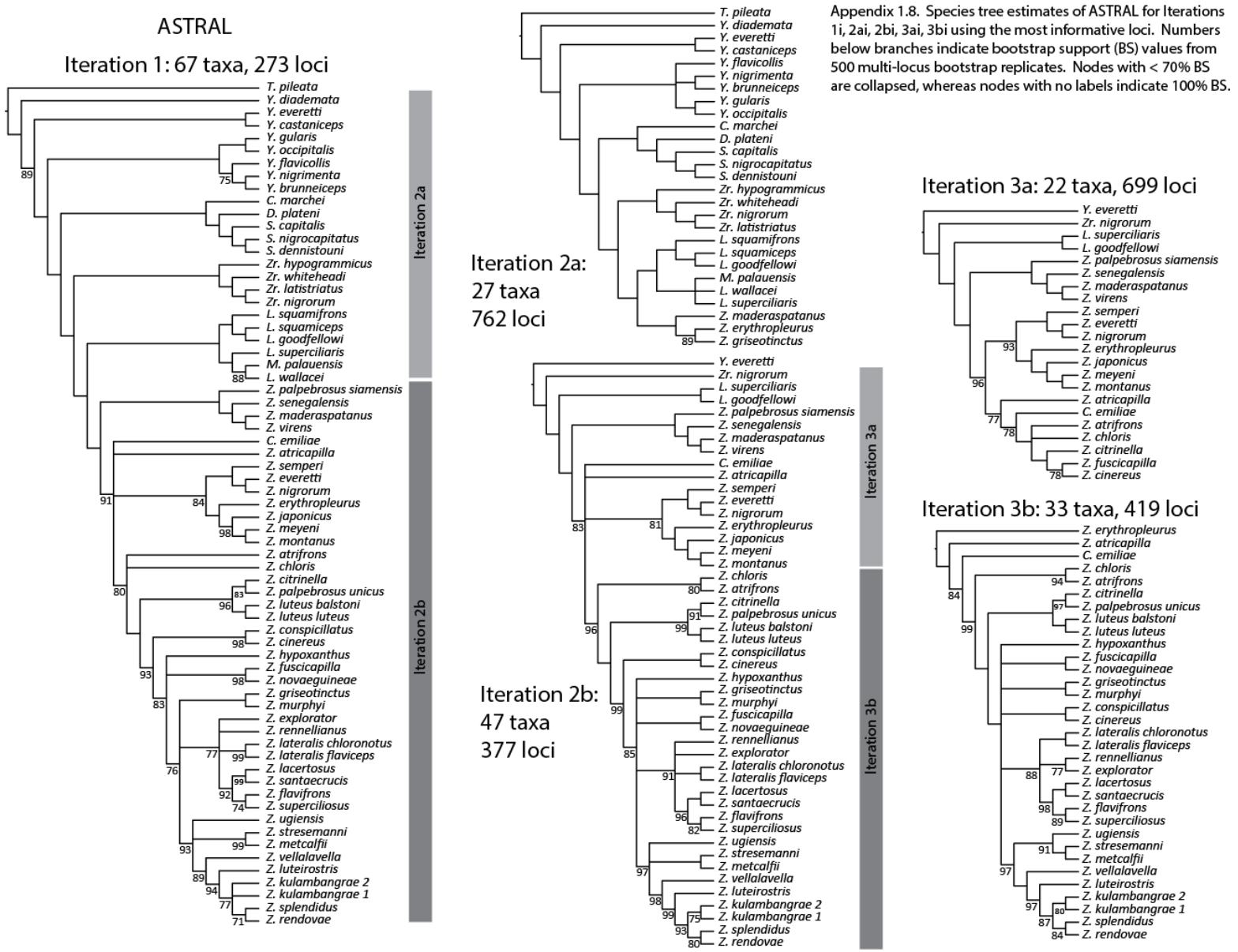
Appendix 1.5. Species tree estimates of STAR for Iterations 1i, 2ai, 2bi, 3ai, 3bi using the most informative loci. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed, whereas nodes with no labels indicate 100% BS.



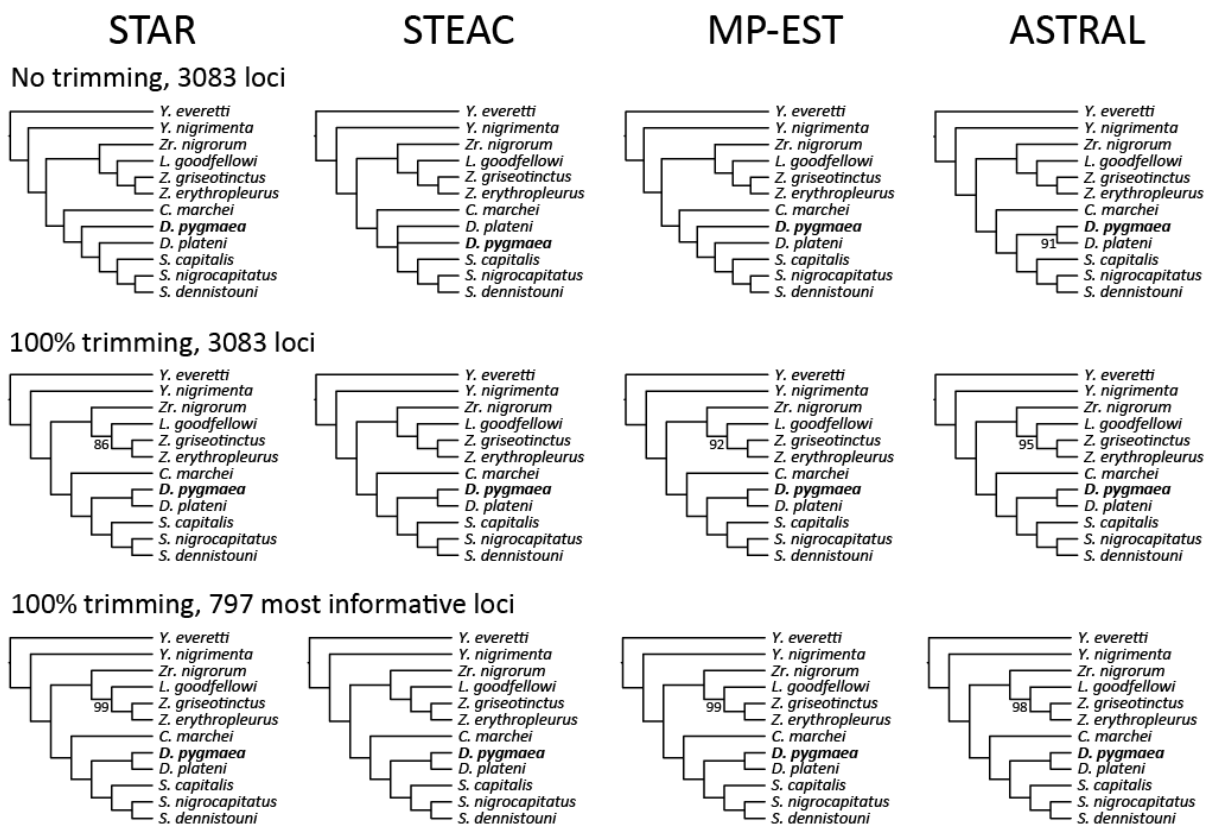
Appendix 1.6. Species tree estimates of STEAC for Iterations 1i, 2ai, 2bi, 3ai, 3bi using the most informative loci. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed, whereas nodes with no labels indicate 100% BS.



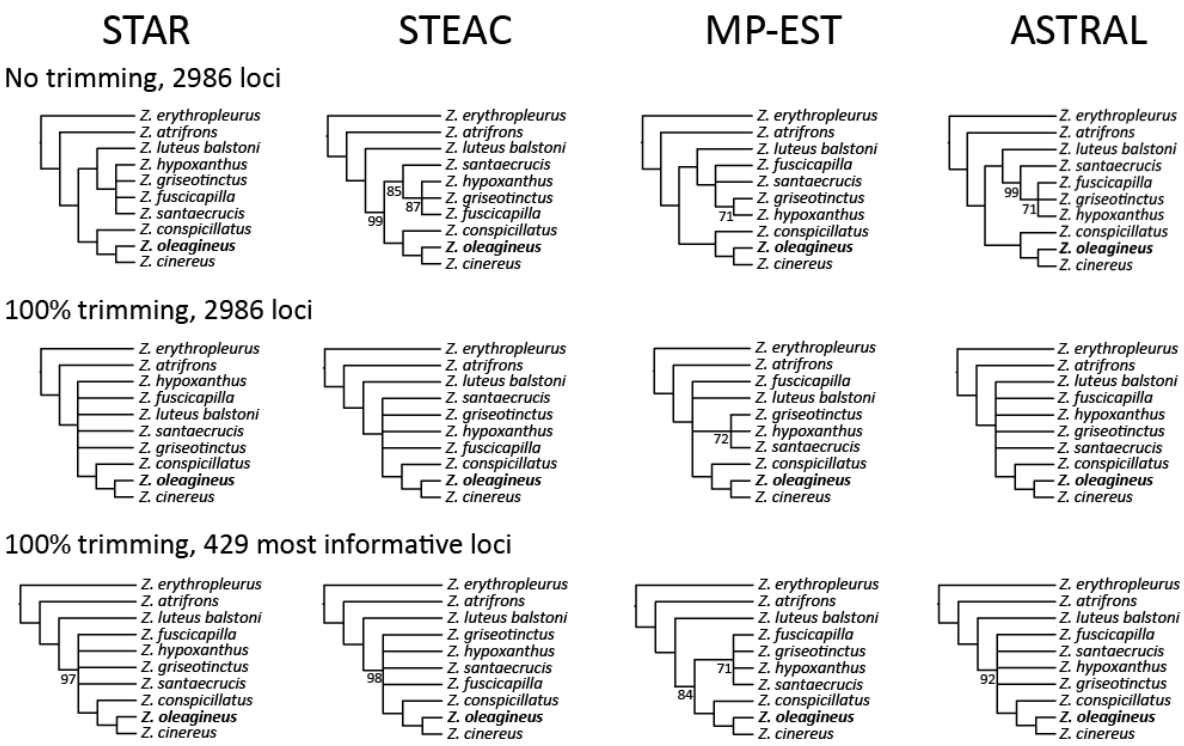
Appendix 1.7. Species tree estimates of MP-EST for Iterations 1i, 2ai, 2bi, 3ai, 3bi using the most informative loci. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed, whereas nodes with no labels indicate 100% BS.



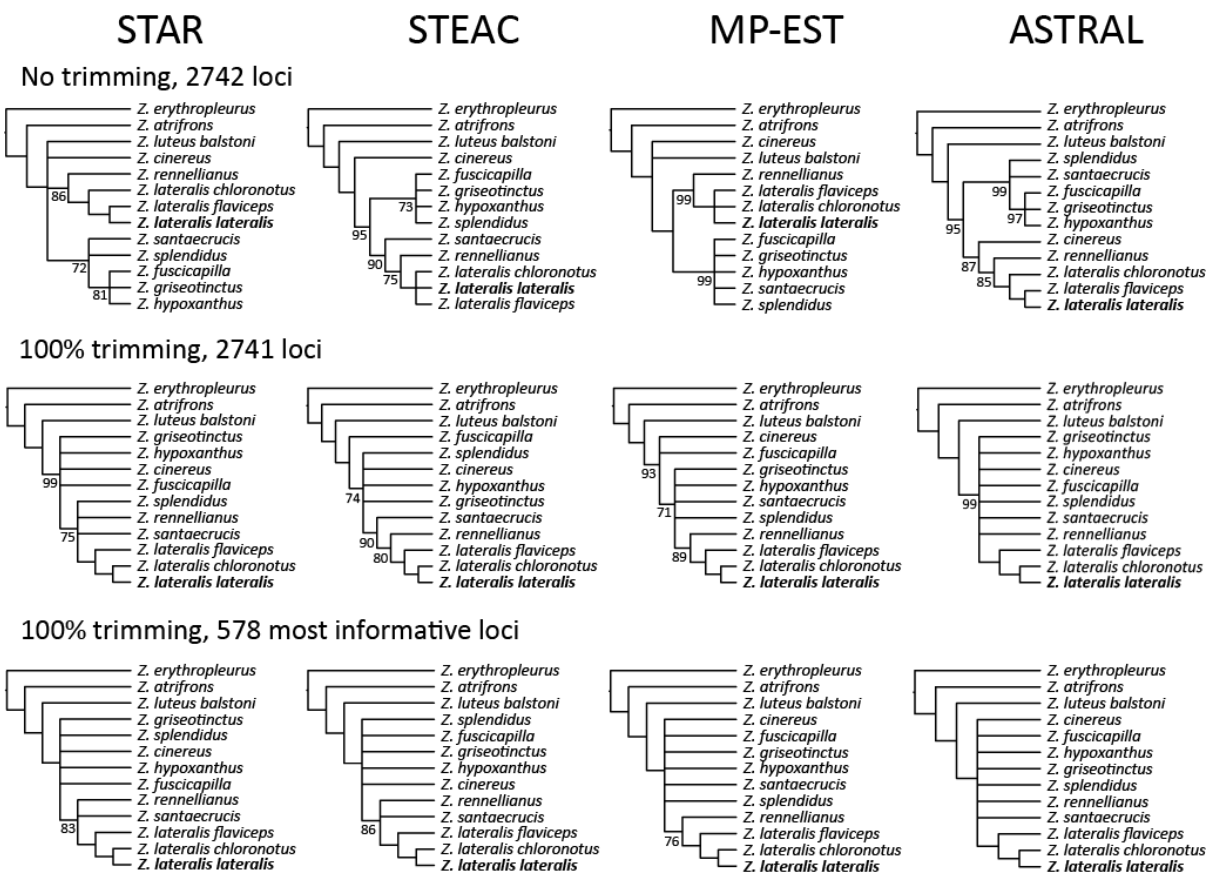
Appendix 1.8. Species tree estimates of ASTRAL for Iterations 1i, 2ai, 2bi, 3ai, 3bi using the most informative loci. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed, whereas nodes with no labels indicate 100% BS.



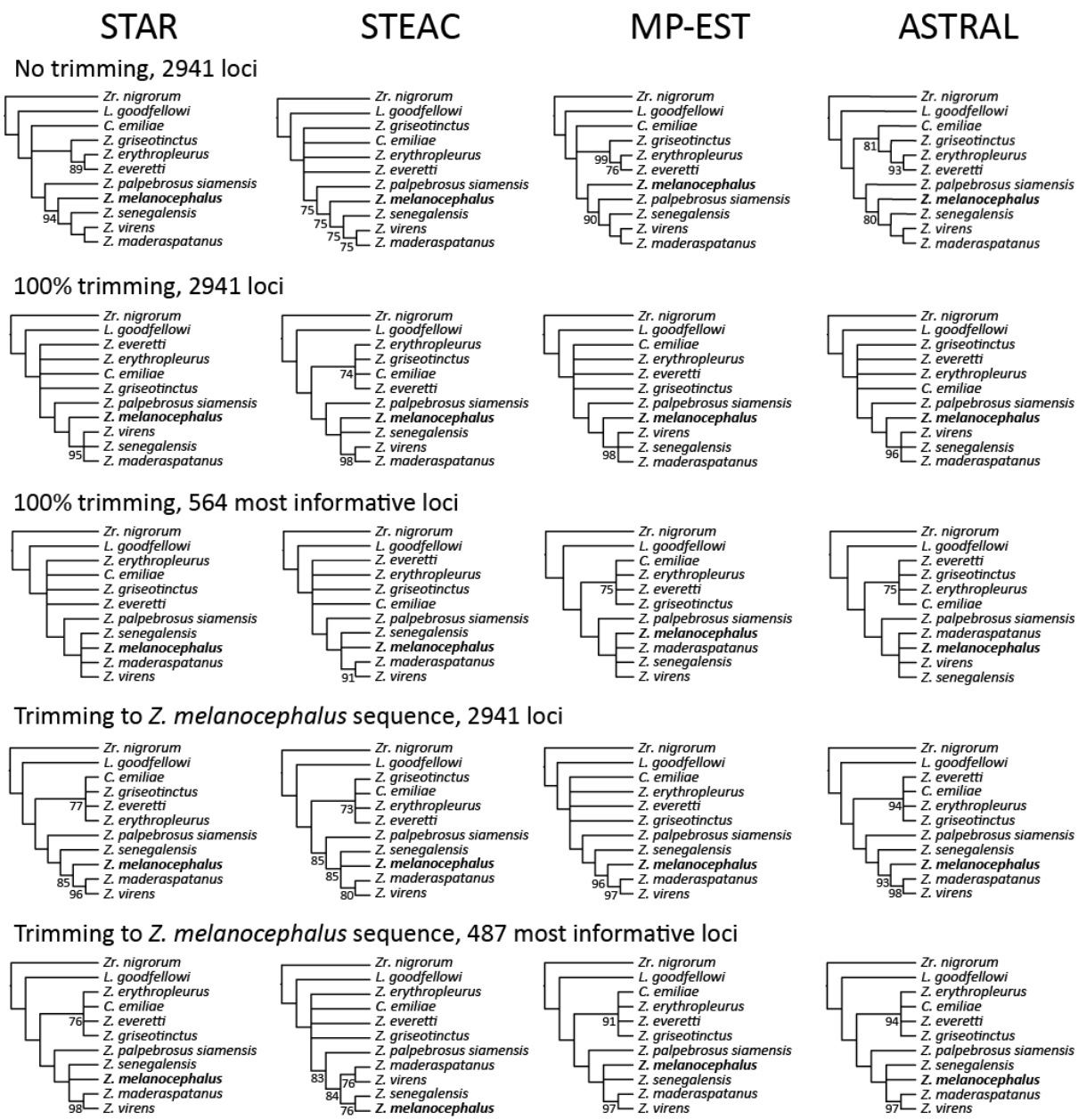
Appendix 1.9. Phylogenetic position of *Dasycrotapha pygmaea* as estimated by STAR, STEAC, MP-EST, and ASTRAL on the untrimmed dataset with all loci, 100% trimmed dataset with all loci, and 100% trimmed dataset with the most informative loci. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed, whereas nodes with no labels indicate 100% BS.



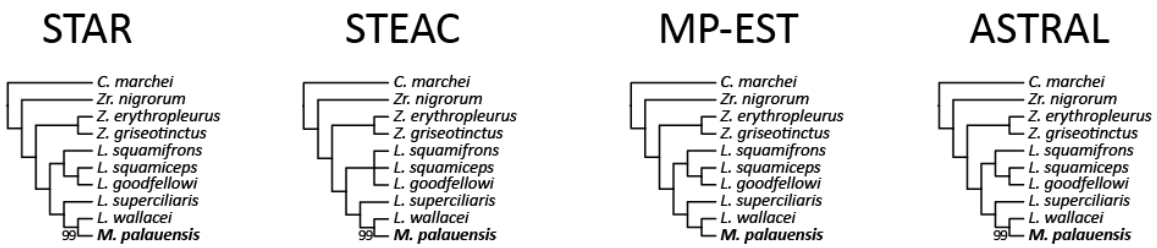
Appendix 1.10. Phylogenetic position of *Zosterops oleagineus* as estimated by STAR, STEAC, MP-EST, and ASTRAL on the untrimmed dataset with all loci, 100% trimmed dataset with all loci, and 100% trimmed dataset with the most informative loci. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed, whereas nodes with no labels indicate 100% BS.



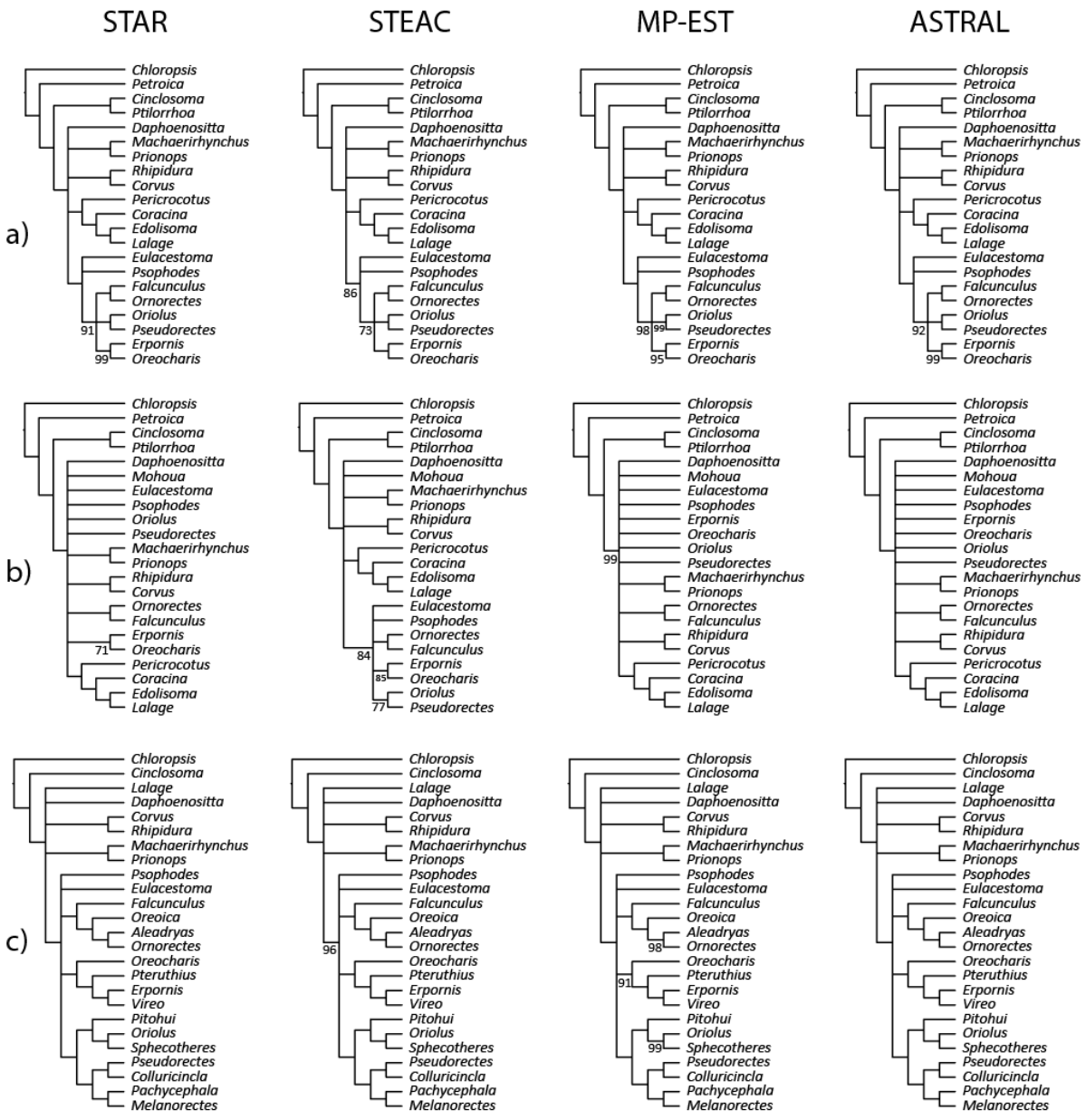
Appendix 1.11. Phylogenetic position of *Zosterops lateralis lateralis* as estimated by STAR, STEAC, MP-EST, and ASTRAL on the untrimmed dataset with all loci, 100% trimmed dataset with all loci, and 100% trimmed dataset with the most informative loci. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed, whereas nodes with no labels indicate 100% BS.



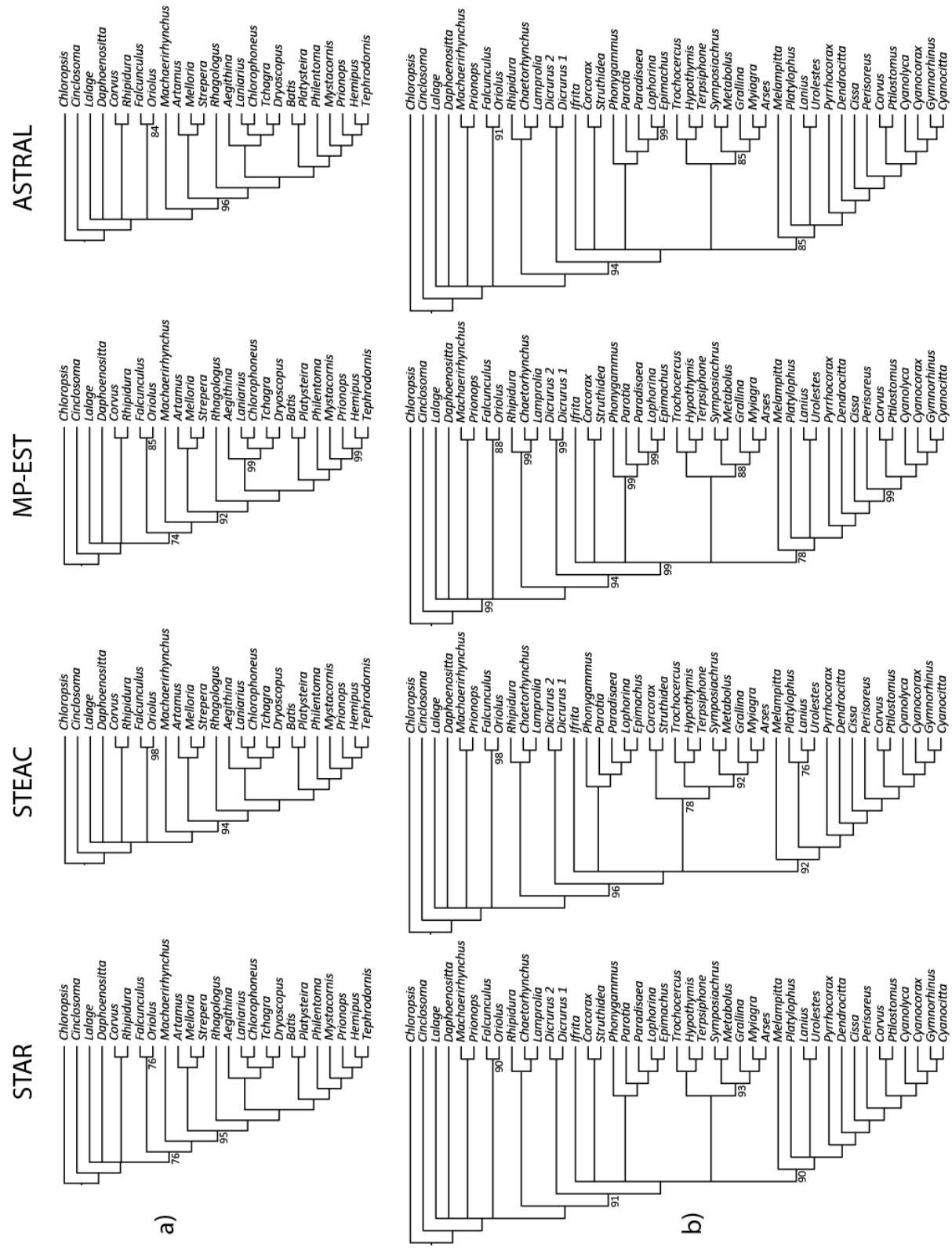
Appendix 1.12. Phylogenetic position of *Zosterops melanocephalus* as estimated by STAR, STEAC, MP-EST, and ASTRAL on the untrimmed dataset with all loci, 100% trimmed dataset with all loci, 100% trimmed dataset with the most informative loci, dataset trimmed to *Z. melanocephalus* sequence length with all loci, and dataset trimmed to *Z. melanocephalus* sequence length with the most informative loci. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed, whereas nodes with no labels indicate 100% BS.



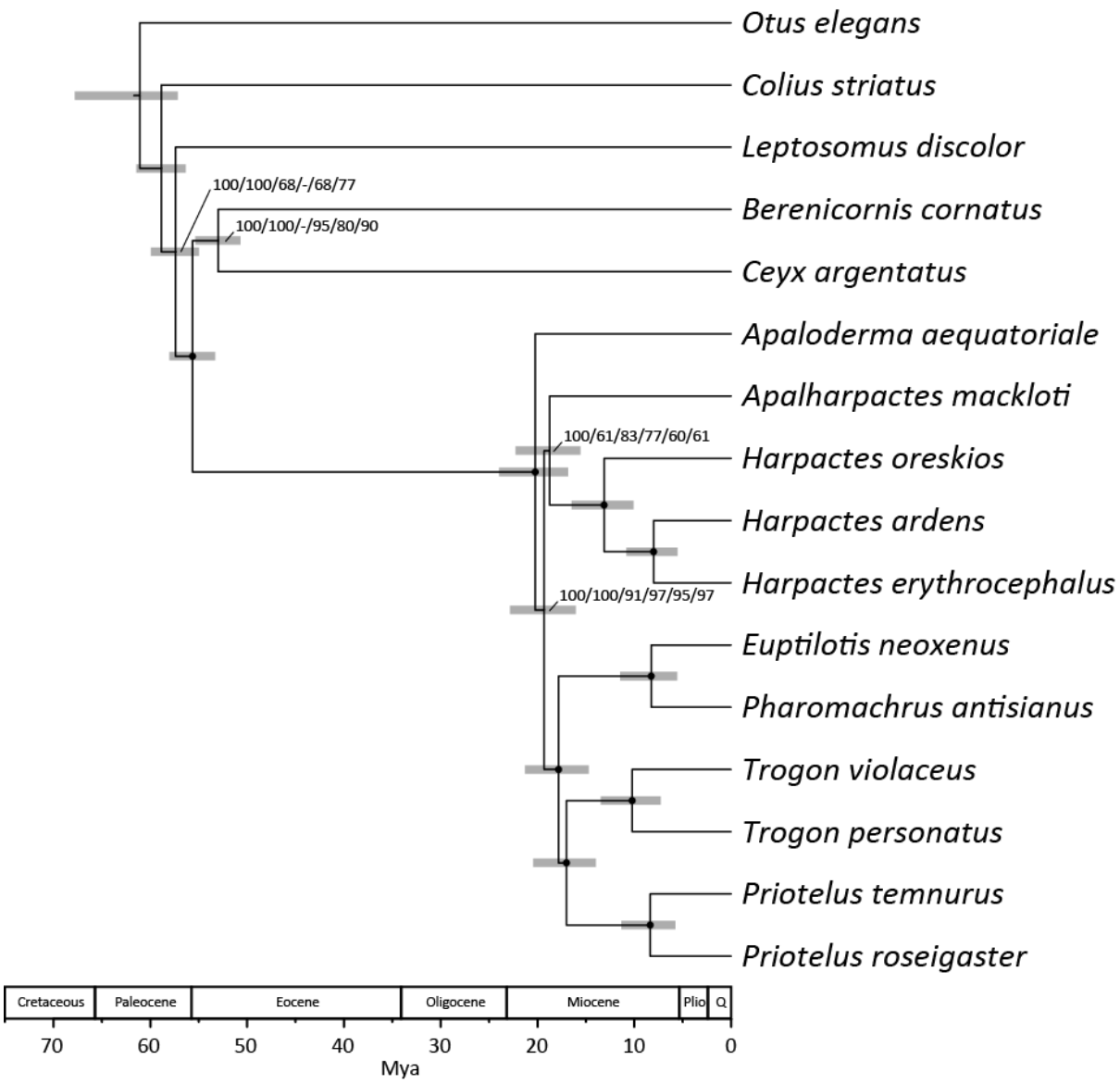
Appendix 1.13. Phylogenetic position of *Megazosterops palauanensis* as estimated by STAR, STEAC, MP-EST, and ASTRAL on the untrimmed dataset with all loci, 100% trimmed dataset with all loci, and 100% trimmed dataset with the most informative loci. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed, whereas nodes with no labels indicate 100% BS.



Appendix 2.1 Species tree estimates of STAR, STEAC, MP-EST, and ASTRAL among: (a) the basal families and major superfamilies in core Corvoidea; (b) the basal families and major superfamilies in core Corvoidea including *Mohoua albicilla*, with sequence alignments trimmed to eliminate missing data at flanks; and (c) Orioloidea. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed whereas nodes with no labels indicate 100% BS.



Appendix 2.2. Species tree estimates of STAR, STEAC, MP-EST, and ASTRAL among (a) Malaconotoidea and (b) Orioloidea. Numbers below branches indicate bootstrap support (BS) values from 500 multi-locus bootstrap replicates. Nodes with < 70% BS are collapsed whereas nodes with no labels indicate 100% BS.



Appendix 3.1. Phylogenetic placement of Trogoniformes. Estimate of phylogenetic relationships of trogons and closely-related orders based on concatenated and coalescent species tree approaches. Nodes with dots correspond to 100% Bayesian posterior probability and bootstrap support from all analysis methods. Numbers next to nodes indicate support from Bayesian, ML, STAR, STEAC, MP-EST, and ASTRAL analyses, respectively. Chronogram based on divergence time estimation with MCMCTree.