RUNNING HEADER: Missing data produces biased loci

TITLE: Uneven missing data skews phylogenomic relationships within the lories and lorikeets

Brian Tilston Smith[1]*, William M. Mauck III[1,2], Brett Benz[1], and Michael J. Andersen[3]

[1]*Department of Ornithology, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, USA*
[2]New York Genome Center, New York, NY 10013, USA.
[3]*Department of Biology and Museum of Southwestern Biology, University of New Mexico, Albuquerque, New Mexico, 87131, USA*

*Corresponding Author: Brian Tilston Smith, *Department of Ornithology, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, USA; Email address: briantilstonsmith@gmail.com*

Keywords.—bird, phylogeny, parrot, likelihood, museum specimen, ancient DNA

Abstract.—Resolution of the Tree of Life has accelerated with massively parallel sequencing of genomic loci. To achieve dense taxon sampling within clades, it is often necessary to obtain DNA from historical museum specimens to supplement modern genetic samples. A particular challenge that arises with this type of sampling scheme is an expected systematic bias in DNA sequences, where older material has more missing data. In this study, we evaluated how missing data influenced phylogenomic relationships in the brush-tongued parrots, or the lories and lorikeets (Tribe: Loriini), which are distributed across the Australasian region. We collected ultraconserved elements from modern and historical material representing the majority of described taxa in the clade. Preliminary phylogenomic analyses recovered clustering of samples within genera, where strongly supported groups formed based on sample type. To assess if the aberrant relationships were being driven by missing data, we performed an outlier loci analysis and calculated gene-likelihoods for trees built with and without missing data. We produced a series of alignments where loci were excluded based on Δ gene-wise log-likelihood scores and inferred topologies with the different datasets to assess whether sample-type clustering could be altered by excluding particular loci. We found that the majority of questionable relationships were driven by particular subsets of loci. Unexpectedly, the biased loci did not have higher missing data, but rather more parsimony informative sites. This counterintuitive result suggests that the most informative loci may be subject to the highest bias as the most variable loci can have the greatest disparity in phylogenetic signal among sample types. After accounting for biased loci, we inferred a more robust phylogenomic hypothesis for the Loriini. Taxonomic relationships within the clade can now be revised to reflect natural groupings, but for some groups additional work is still necessary.

Introduction

Historical and ancient DNA from museum specimens are widely employed for incorporating rare and extinct taxa in phylogenetic studies (e. g., Thomas et al. 1989; Mitchell et al. 2014; Fortes et al. 2016). The inclusion of these samples has helped discover and delimit species (Helgen et al. 2013; Paijmans et al. 2017), resolve phylogenetic relationships (Mitchell et al. 2016), and clarify biogeographic history (Kehlmaier et al. 2017; Yao et al. 2017). DNA obtained from dry and alcohol-preserved museum specimens has been collected using a range of techniques, including Sanger sequencing (Sorensen et al. 1999), restriction site associated DNA sequencing (Tin et al. 2015), and sequence capture of reduced (McCormack et al. 2016; Ruane et al. 2017; Linck et al. 2017) or whole genomes (Hung et al. 2014; Enk et al. 2014). However, the DNA sequences collected from these museum specimens are subject to errors associated with contamination (Malmström et al. 2005), DNA degradation (Briggs et al. 2007; Sawyer et al. 2012), and low coverage in read depth (Tin et al. 2015), which all present challenges in distinguishing evolutionary signal from noise.

Sequence capture of ultraconserved elements (UCEs) is a popular approach for collecting orthologous genomic markers in phylogenomic studies (Faircloth et al. 2015; Chakrabarty et al. 2017; Esselstyn et al. 2017) and is increasingly used for historical specimens (McCormack et al. 2016; Hosner et al. 2016; Ruane et al. 2017). A common finding is that the length and number of loci recovered are typically shorter and smaller in older samples (McCormack et al. 2016; Hosner et al. 2016; Ruane et al. 2017). Shorter loci are potentially problematic because the UCE sequence includes the invariable core and a limited portion of the flanking region that contains polymorphic sites. Although some studies only use DNA sequences collected from historical or ancient samples (e. g., Hung et al. 2014), most phylogenetic studies combine data from historical and modern samples. For studies that use DNA from both sample types, additional challenges in downstream analyses may arise due to an asymmetry in the phylogenetic signal caused by missing data (e. g., Hosner et al. 2016).

The impact of missing data on phylogenetic inference remains contentious (Lemmon et al. 2009; Simmons 2012; Wiens and Morrill 2011; Hovmöller et al. 2013; Simmons 2014; Streicher et al. 2015). Findings suggest that even when a large proportion of sites have no data, phylogenetic signal is retained if enough characters are present (Philippe et al. 2004; Roure et al. 2012; Shavit Grievink et al. 2013; Molloy and Warnow 2017). In contrast, missing data has also been shown to bias phylogenetic relationships, particularly when the missing data is non-randomly distributed (e. g., Lemmon et al. 2009; Simmons 2012; Simmons 2014). Bias may manifest as inflated support values or erroneous branch lengths, or by producing inconsistencies when using different optimality criteria or phylogenomic approaches (i.e., concatenation vs the multi-species coalescent). The increased availability in phylogenomic data has provided a more nuanced look at missing data's effect on phylogenetic inference (Philippe et al. 2004; Huang et al. 2014; Streicher et al. 2015; Xi et al. 2015). One means of accounting for missing data in phylogenomic datasets is to filter loci based on the proportion of either missing data or missing species in the dataset (Hosner et al. 2016). However, this approach may not directly target problematic regions of an alignment, and phylogenetically informative signal may be discarded unnecessarily. A more direct approach would entail identifying which specific sites or genes are influenced by missing data.

Analyses of outlier loci in phylogenomic data indicate that a small number of genes can

have a large impact on a topology (Shen et al. 2017; Arcila et al. 2017; Brown et al. 2018; Walker et al. 2018). These conflicting genealogies can be due to biological processes (e. g., incomplete lineage sorting; introgression; horizontal gene transfer) or to spurious phylogenetic signal caused by poor alignments, paralogy, and/or sequencing error. Putative outlier loci have been identified using topology tests (Arcila et al. 2017; Esselstyn et al. 2017), Bayes factors (Brown and Thomson 2017), and site/gene-wise log-likelihood differences among alternative topologies (Shen et al. 2017; Walker et al. 2018). Support for particular phylogenetic hypotheses may be driven by a small subset of loci (Brown and Thomson 2017), and the targeted removal of outlier loci has been shown to decrease conflict among alternative topologies (Walker et al. 2018). Phylogenetic outlier analyses may provide a framework for assessing how missing data from historical samples can impact a topology. In this study, we used gene-wise likelihoods to assess how the expected missing data in DNA sequences sourced from historical specimens impacted the estimated topologies in our focal group, the Loriini.

Lories and lorikeets, commonly known as the brush-tongued parrots, are a speciose clade (Tribe: Loriini) of colorful birds that are widely distributed across the Australasian region (Forshaw et al. 1989). The characterization of distributional ranges, phenotypic variation, and systematics of the clade were the product of expansive biological inventories that peaked during the early 1900s (Forshaw et al. 1989). The geographical extent of this work encompasses thousands of large and small islands spread across many countries in the Australasian region. Given these immense logistical constraints, modern collecting expeditions that aim to produce voucher specimens with genetic samples for continued systematic work (e. g., Kratter et al. 2006, Andersen et al. 2017) have been much more focused in scope relative to the pioneering work of the 20th century that produced extensive series of specimens across species entire ranges (e. g., Mayr 1933; Mayr 1938; Mayr 1942; Amadon 1943). Thus, the lack of modern genetic samples, phylogenetic relationships in many groups, like the Loriini, remain unresolved. To get around this constraint, phylogenomic studies have sourced DNA from historical specimens (Moyle et al. 2016; Andersen et al. 2018).

Prior phylogenetic work on the Loriini showed evidence for at least three paraphyletic genera and highlighted the need for increased taxon and genomic sampling to more fully resolve relationships among taxa (Schweizer et al. 2015). To this end, we collected UCEs from 94% of the 112 described taxa in the Loriini. Our sampling design used DNA isolated from fresh tissues and historical specimens up to 100 years old (Fig. 1). We anticipated challenges with processing and analyzing UCEs from historical specimens, so we produced alignments using both a standard pipeline and alignments using more stringent filtering. Preliminary estimates of phylogenetic relationships showed unusual but strongly-supported relationships where historical samples grouped together. We explored whether this pattern was caused by missing data by estimating gene likelihoods for topologies estimated with and without missing data, and identified which loci had the largest likelihood differences. We then assessed the impact of filtering the identified loci to verify the stability of nodes within the Loriini phylogeny. After assessing the impact of missing data, we propose a phylogenetic hypothesis for the lories and lorikeets.

Materials & Methods

We sampled all 12 genera, 55/56 species, and 105 (Dickinson and Remsen 2013) or 106 (Clements et al. 2017) of 112 named taxa within the Loriini. *Charmosyna diadema* is the only species not included in our study, which is extinct and known from a single female specimen

(Forshaw et al. 1989). Three additional taxa (*Eos histrio talautensis*, *E. squamata riciniata, and Lorius lory viridicrissalis*) produced limited data and were excluded from final analyses. We did not obtain samples from the following taxa: *Trichoglossus haematodus brooki*, *Psitteuteles iris rubripileum*, *Neopsittacus pullicauda socialis*, *E. histrio challengeri*, *Pseudeos fuscata fuscata*, and *C. rubronotata kordoana*. When possible, we sampled more than one individual per species to verify the phylogenetic position of a taxon. For outgroups, we used *Melopsittacus undulatus*, *Psittaculirostris edwardsii*, and *Cyclopsitta diophthalma*, which together with the Loriini form the clade Loriinae (Joseph et al. 2012; Provost et al. 2018). Specimen details and locality information are available in Supplementary Table S1 available on Dryad (Dryad link pending).

We extracted total genomic DNA from muscle tissue using QIAamp DNeasy extraction kits (Qiagen, Valencia, CA). For historical samples from museum specimens we used a modified DNeasy extraction protocol that used QIAquick PCR filter columns that size selected for smaller fragments of DNA. The modified protocol also included washing the sample with H_2O and EtOH prior to extracting as well as extra time for digestion. DNA extraction from historical samples was done in a dedicated lab for working with degraded samples to reduce contamination risk. We quantified DNA extracts using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific). Library preparation of UCEs and enrichment, and Illumina sequencing were performed by RAPiD Genomics (Gainesville, FL). The Tetrapod UCE 5K probe set was used to enrich for 5,060 UCE loci (Faircloth et al. 2012). The wetlab component of this study was carried out over three years and the number of individuals multiplexed per lane ranged from 48–384. Sequencing was done on an Illumina HiSeq 2500 PE 125 or HiSeq 3000 PE 150. Fastq files are available on the Short Read Archive (SRA numbers pending).

We used a modified data-processing pipeline that incorporated PHYLUCE (Faircloth 2015), a software package developed for analyzing UCE data, and seqcap_pop (Smith et al. 2014; Harvey et al. 2017). Low-quality bases and adaptor sequences were trimmed from multiplexed fastq files using illumiprocessor v1 (Faircloth 2013; Bolger et al. 2014). Next, reads were assembled into contigs with Trinity v2.0.6 (Grabherr et al. 2011) and contigs were mapped to UCE probes. We chose the sample that produced the largest number of UCEs as the reference for subsequent mapping for all individuals. We generated an index of the reference sequence and independently mapped reads from each sample to the same reference sequence using BWA v0.7.13-r1126 (Li and Durbin 2009). SAM files produced from the BWA mapping were converted to BAM files and sorted with SAMtools (Li et al. 2009). Then, we used the mpileup function in SAMtools to get a summary of the coverage, bcftools and vcfutils to call variant sites, and seqtk to convert fastq files to fasta. These collective steps produced single fasta files containing all UCE loci for each individual sample. Then we concatenated fasta files of each sample and used MAFFT (Katoh & Standley 2013) to align the concatenated sequence. Next, we retained loci where 75% of the samples were present in a locus, from which we assembled a concatenated alignment using PHYLUCE and a SNP alignment using SNP-sites (Page et al. 2016). We produced both a minimum (here after unfiltered) and more stringently filtered (here after filtered) datasets. The extra filtering included masking sites with less than 6x coverage with bcftools, trimming alignment ends during the concatenation alignment step, and removing individuals from each locus that had more than 30% missing sites. All downstream analyses were completed for both the unfiltered and filtered datasets.

Phylogenomic outlier loci analysis

To examine how missing data may influence phylogenetic relationships, we compared topologies estimated with and without missing data. The number of sites with missing data increased with the number of individuals in the alignment. To retain phylogenetic information after removing sites with missing data in an alignment, we performed clade-based analyses where we estimated topologies using a concatenated alignment with (topology T_1) and without missing data (topology T_2). The six clades were based on preliminary phylogenetic analysis and were 1) *Eos*, *Trichoglossus*, *Psitteuteles iris* and *Glossopsitta*, 2) *Chalcopsitta* and *Pseudeos*, 3) *Lorius*, 4) *Psitteuteles versicolor* and *Parvipsitta*, 5) *Neopsittacus*, and 6) *Charmosyna*, *Vini*, and *Phigys*. To produce a concatenated alignment for a clade, we followed the same steps listed above. To retain more sites in the larger clades we did not include redundant taxa or samples. We also had to divide the speciose clade containing *Eos*, *Trichoglossus*, *Psitteuteles iris*, and *Glossopsitta* into two separate alignments with different samples, because the amount of missing data in this clade was high. We calculated alignment statistics for each locus and for alignments partitioned into historical or modern samples using AMAS (Borowiec 2016). To estimate a maximum likelihood tree for each clade's concatenated alignments with and without missing data, we used the software package IQ-TREE v. 1.5.5 using the best-fit substitution model (Nguyen et al. 2014).

We performed a two-topology, site-specific log-likelihood test that estimated the site-likelihoods based on the gene partition of the full concatenated alignment using two alternative topologies (T_1 and T_2) in RAxML. Site-likelihoods were converted to gene-wise log-likelihoods by summing the site-likelihoods for each gene using the scripts in Walker et al. (2018). We then estimated the Δ gene-wise log-likelihoods (Δ gene-wise log-likelihood = T_1 log-likelihood - T_2 log-likelihood). Higher, and positive, Δ gene-wise log-likelihoods for gene partitions indicate that missing data changed phylogenetic relationships. We built a neural net in the R package caret v. 6.0.79 (Kuhn 2008) to test whether the Δ gene-wise log-likelihood of each gene partition could be predicted by the alignment statistics. The alignment statistics (alignment length, percent of missing data, number of parsimony informative sites, number of variable sites, and GC content) were specified as the input neurons, and the output neuron was the Δ log-likelihood. The input data was scaled to the min and max for each statistic, and the percentage of training/test data was set to 75%/25%, respectively. We performed this analysis on the six sub-clades. To assess whether the information content varied among historical and modern samples for the datasets produced from gene-wise log-likelihood analysis, we plotted the number of variable sites versus percentage of missing data for each locus and calculated 95% confidence interval ellipses around the points for each dataset using the R package car v. 2.1.6. We then calculated the area of overlap among ellipses using SIBER v. 2.1.3.

To explore how these putatively biased loci impacted phylogenetic inference, we grouped loci into classes representing Δ gene-wise log-likelihoods of greater than 2, 10, and 20 and successively excluded these loci from the global alignments. In this assessment we also built trees that included all samples, one individual per taxon, and one per species to examine how excluding loci impacted trees with varying levels of tip-sampling. We then estimated concatenated and species trees to assess how sensitive phylogenetic relationships were to excluding loci in each of these approaches. For each alignment, we used PartitionFinder 2.1.1 (Lanfear *et al.* 2016) to determine the best-fitting substitution model per gene partition using the rclusterf algorithm (Lanfear et al. 2014) and RAxML option (Stamatakis 2014). We performed concatenated tree analyses in IQ-TREE, with 1000 rapid bootstraps (Nguyen et al. 2014), on the

full dataset and the reduced datasets produced from the gene likelihood analyses. To estimate species trees using SVDquartets (Chifman and Kubatko 2014) in PAUP* test-version (Swofford 2003) and performed 100 bootstrap replicates. We did not estimate species trees for the dataset containing all samples because some of the samples were from captive birds that could not be accurately assigned to a taxon below the species-level. To visualize how different the trees were, we measured the distance among 100 bootstrap trees using Robinson-Foulds distances (Robinson and Foulds 1981) with the multiRF function in phytools (Revell 2012) and used multidimensional scaling to plot the distances in two-dimensional space. Trees were plotted using the R packages phytools and ggtree (Yu et al. 2017).

Results

Data characteristics

We sequenced 176 unique samples, including 16 that were re-sequenced to improve the amount of data recovered. We dropped six individuals that produced limited data and had a final dataset of 170 unique individuals (167 ingroup, three outgroup). In the unfiltered and filtered datasets we recovered 4,256 (mean: 761 range: 224-2,310) and 4,159 (mean: 489 range: 101-1,900) loci, respectively. The alignment lengths for unfiltered and filtered were 3,237,881 bps with 101,742 parsimony informative sites, and 2,033,655 bps with 36,515 parsimony informative sites, respectively. The mean and range number of taxa was 159 unfiltered (41-170) and 145 filtered (3-171).

Outlier loci

For each dataset, we inferred topologies for the six clades listed above using alignments with and without missing data. A comparison of gene-wise log-likelihoods showed that up to 48% of the loci had a Δ log-likelihood of ≥ 2 (Fig. 2). There were 68/98 (unfiltered/filtered), 406/553, and 1,992/1,525 loci in the three bins (Δ gene-wise log-likelihood of 20+, 10+ and 2+), respectively (Fig. 2). In the neural net models, alignment statistics were generally poor predictors of Δ gene-wise log-likelihood scores (Supplementary Table S2), but the number of parsimony informative sites was typically the most important variable. Modern samples had loci with more variable sites than the historical samples (Fig. 3). In the unfiltered dataset, the ellipses of included loci had the highest number of variable sites, and in the filtered dataset, the excluded loci were the highest (Fig. 3). The proportion of overlap among ellipses was lower in the unfiltered versus the filtered datasets (Fig. 3; Supplementary Table S2). The lowest overlap proportions were in comparisons among historical and modern samples of included versus excluded loci (Fig. 3; Supplementary Table S2). Removing putative outlier loci did not qualitatively alter the degree of overlap among ellipses (Fig. 3; Supplementary Table S2).

From the gene likelihood analyses we produced 12 separate concatenated alignments for both the filtered and unfiltered data, and inferred 24 maximum likelihood concatenated trees and 16 species trees. Multidimensional scaling of Robinson-Foulds' distances among topologies showed that excluding loci changed the topology (Fig. 4). The variance in Robinson-Foulds' distances among bootstrap trees was lower than among the trees produced with varying levels of outlier loci removed (Fig. 4). For most of the topologies, the greatest variances in distances were among species and concatenated trees, and among trees estimated from the unfiltered and filtered alignment, irrespective of the outlier loci removed (Fig. 4). There were several important

distinctions between the trees built with the filtered and unfiltered alignments. There was greater disparity in branch lengths within the unfiltered trees and apparent spurious phylogenetic relationships (Fig. 5). Long branches in the unfiltered dataset remained even after removing loci detected using gene likelihoods (Fig. 5). The unfiltered tree exhibited clustering of historical and modern samples, rendering strong support for some non-monophyletic species and genera (Supplemental Figs. S3-S12). Species trees were generally concordant, albeit with lower support, with the concatenated topology (Supplemental Figs. S3-S12).

Impacts on phylogenetic relationships

The backbone phylogeny we inferred for the Loriini generally had high support and was stable across the filtering schemes, but relationships towards the tips showed varying levels of stability (Fig. 6; Supplemental Figs. S3-S12). *Oreopsittacus* was sister to all other ingroup taxa, then *Charmosyna* was sister to the clade containing *Neopsittacus*, *Lorius*, *Pseudeos*, *Chalcopsitta*, *Psitteuteles*, *Glossopsitta*, *Eos*, *Trichoglossus*, and *Parvipsitta*. The placement of *Neopsittacus*, *Lorius*, *Pseudeos*, and *Chalcopsitta* was stable and well-supported, and each of these genera were monophyletic. *Trichoglossus*, *Charmosyna*, and *Psitteuteles* were not monophyletic (Fig. 6; Supplemental Figs. S3-S12). Relationships among a monophyletic *Eos* and an apparent paraphyletic *Trichoglossus* were poorly resolved. *Vini* and *Phigys* are strongly supported as nested within *Charmosyna*. *Psitteuteles* is found in three separate places in the tree (Fig. 6; Supplemental Figs. S3-S12): *P. versicolor* is sister to the recently erected genus *Parvipsitta*, *P. iris* is nested within a clade of *Trichoglossus* from Indonesia, and *P. goldei* is sister to a clade containing *Glossopsitta*, *Eos*, and *Trichoglossus,* and *P. iris*. Excluding loci with high gene-wise log-likelihood scores had the greatest impact in the clade containing *Trichoglossus*, *Eos*, *Psitteuteles iris*, and *Glossopsitta concinna* (Fig. 6; Supplemental Figs. S3-S12), but some relationships remained unstable. A putative clade containing *P. iris*, *T. ornatus*, *T. flavoviridis*, and *T. johnstoniae,* which was sister to *Eos,* had varying support across the analyses, but collectively our findings suggest the taxa are in a clade separate from the other *Trichoglossus*. Stable phylogenetic placements with high support in the remaining *Trichoglossus* taxa was not recovered in the unique taxa and species only datasets (Supplemental Figs. S4 and S5). Particularly, the placement of *T. h. caeruleiceps, T. h. micropteryx, T. h. flavicans, and T. h. nigrogularis,* which were all taxa that came from historical specimens, varied among filtered and unfiltered datasets. The placement of this group of taxa ranged from being sister to *Trichoglossus* proper to sister to the entire clade of *Trichoglossus* and *Eos*, albeit with low support. *Trichoglossus rubiginosus*, an aberrantly colored taxon in the clade, could not be accurately placed and moved around the tree depending on the dataset (Fig. 6; Supplemental Figs. S3-S12).

The trees based on complete sampling best captured the sensitivity of biased loci to phylogenetic relationships (Supplemental Fig. S3 and S8). These trees, which had all 170 samples and no loci removed had a higher frequency of non-monophyletic species and genera. As the number of taxa was reduced to all unique taxa or species only, the problematic relationships remained in the unfiltered dataset (Supplemental Figs. S4 and S5). The cases where species were not monophyletic often showed samples clustered based on sample type. The filtering schemes indicated that some of these clusters could be broken up by removing loci. For example, samples of *Eos bornea* are not monophyletic in any of the unfiltered trees, but when outlier loci were removed in the filtered dataset the taxon became monophyletic (Supplemental Figs. S3, S4, S8, and S9). Similar clustering among historical and modern samples was observed

in *Psitteuteles iris* and *Chalcopsitta atra*, but these relationships were not altered when outlier loci were removed. *Trichoglossus haematodus* was not monophyletic and the apparent clustering of samples within this species was associated with sample type. Although *Trichoglossus* was not a well-supported clade, all trees indicated *T. euteles*, *T. forsteni*, *T. capistratus*, *T. weberi*, and *T. rubritorquis* are nested within *T. haematodus*. *Trichoglossus forsteni stresemanni* is more closely related to *T. capistratus* then to other *Trichoglossus forsteni* taxa.

Discussion

We showed that systematic bias in missing data caused by sampling DNA from modern and historical specimens produces aberrant phylogenetic relationships. To obtain dense taxon sampling in our focal group, the Loriini, we leveraged samples collected over the last 100 years and assessed how this sampling scheme impacted phylogenetic relationships by filtering alignments based on sequence quality, missing data, and loci. We found clear examples of where strongly supported phylogenetic relationships and branch lengths differed among our datasets. There were cases of clustering within genera where strongly supported groups formed based on sample type and these relationships were driven by subsets of loci that had higher variation and missing data. Branch lengths from historical material, in some samples, were atypically long. After accounting for biased loci, we inferred a more robust phylogenetic hypothesis for the Loriini. Taxonomic relationships within the clade can now be revised to reflect natural groupings, but for some groups additional work is still necessary.

Massively-parallel sequencing technology has provided systematists with an unprecedented amount of information for inferring phylogenetic relationships (McCormack et al. 2013). However, the data has come faster than the development of best practices for assembling, processing, and analyzing large phylogenomic datasets, particularly from low-quality samples. Alignments produced without careful inspection may harbor issues that can have a large impact on downstream analyses (Springer and Gatesy 2018). In the case of this study, the findings presented here have general implications for phylogenomic studies where there is an asymmetry in parsimony informative sites among closely related taxa. The magnitude of biases will likely vary according to clade diversity and age, and the number of loci collected. We found that the bias was most extreme in a diverse and rapid radiation where there was likely limited information, in even complete loci, for teasing apart relationships. Shallow systematic and phylogeographic studies are expected to be most difficult temporal scale for resolving relationships when there is high missing data associated with particular samples. Moving forward, understanding the informational content of a locus, and how that information effects the genealogy, will help avoid inferring dubious phylogenomic relationships.

Identifying Distorted Relationships

As expected, historical samples contained less phylogenetic information than modern ones, but on closer inspection the disparity in information content was more nuanced. We produced both minimally and stringently filtered alignments using bioinformatic pipelines that are typically employed in phylogenomic studies to explore the interaction between sample type and data quality on phylogenetic inference. The unfiltered dataset had less missing data, longer loci, and more variation than the filtered dataset. However, the unfiltered dataset produced trees with more asymmetric branch lengths and more unexpected relationships. A comparison of those loci included in tree-building versus those that we excluded based on gene-wise log-likelihood

scores showed that in all cases, the excluded loci had a higher number of parsimony informative sites despite being similar in locus length and the amount of missing data. For the majority of loci, similar patterns of variable sites and missing data were observed across sample types. Within the modern samples there was a subset of loci with higher variation, and identifying these loci was key to accounting for bias in our data.

Gene and site likelihood scores provide a rapid means of identifying loci that have a large impact on a phylogeny (Shen et al. 2017; Walker et al. 2018). Prior work has applied site and gene likelihoods to look at contentious relationships in higher-level relationships within plants, fungi, and animals (Shen et al. 2017; Walker et al. 2018), and found that the difference between alternative phylogenies could be explained by a few genes or sites. By examining a relativity speciose and recent radiation (Schweizer et al. 2015), we found that gene likelihoods could identify loci biased by missing data. Removing the loci with the highest Δ gene-wise log-likelihoods altered the majority of the questionable relationships and broke-up presumed clumping of samples from historical specimens. Increased filtering of loci reduced support for some nodes and had a limited impact on further breaking up apparent sample-type clustering. Although select loci had a dramatic impact on portions of the tree, most relationships that included modern and historical samples did not appear to be biased by missing data. The disparity in node stability to outlier loci was reflected in the number of loci detected in each of the six clades assessed. The total number of loci with likelihood differences was associated with the sample size in each clade, with the least diverse clade (*Psitteuteles versicolor* and *Parvipsitta*) having < 10 loci with a Δ log-likelihood of ≥ 2 and the most diverse (*Eos*, *Trichoglossus*, *Glossopsitta*, and *P. iris*) having almost 2000.

The more challenging bias in missing data among sample types appears to arise in loci with higher variation instead of those with limited resolution. As the number of parsimony informative or variable sites increases for a given locus, that variation is more likely to be in the modern samples due to the structure of a UCE. For example, longer loci, which are expected to have more variant sites, tended to be detected in the gene likelihood analysis. Our neural net models were not able to predict Δ gene-wise log-likelihood scores, but number of parsimony informative sites was often the most important variable. One explanation for this result was that even though the loci with high Δ gene-wise log-likelihood scores tended to have a higher number of parsimony informative sites, there were also a large number of variable loci with low scores. This pattern indicates that clumping of historical samples occurs because the disparity in phylogenetic signal among sample-types is greater than their phylogenetic distances. Preferentially selecting phylogenetically informative loci is expected to produce more well-supported trees (Gilbert et al. 2018), but our results suggest that this practice can produce less reliable relationships when the data content varies among samples. The controversial relationships in our dataset changed as we removed loci, but support generally remained high across alternative topologies. In the most extreme examples of phylogenetic signal disparity, samples can fall outside of their clade or even the ingroup, as evident in previous phylogenomic studies on birds (Hosner et al. 2016; Moyle et al. 2016). This was the case for seven of our excluded samples, which produced limited data and could not be accurately placed in their genus or higher-level clade.

The majority of the topological instability in our Loriini dataset was within one clade that appears to represent a rapid radiation containing four genera and 37 taxa. Rapid radiations produce short internodes that are notoriously difficult to resolve (e. g., Hackett et al. 2008;

Rothfels et al. 2012; Pyron et al. 2014). Missing data likely exacerbates the challenges of inferring relationships when the phylogenetic signal is low to begin with. Similarly-distributed clades of birds in the Australasian region show exceptional diversification rates (Andersen et al., 2018). Although our Loriini phylogeny is not time-calibrated, we suspect the clade containing *Eos*, *Trichoglossus*, *Glossopsitta*, and *P. iris* exhibits a similar temporal pattern of diversification given its high diversity. For example, the Rainbow Lorikeet (*Trichoglossus haematodus*) is recognized with up to 20 phenotypically distinct subspecies and covers a wide geographic area (Clements et al. 2017). In contrast, another diverse clade containing *Charmosyna*, and *Vini* and *Phigys* (*n*=28) was consistently well-supported and had stable relationships that were not as susceptible to biased loci despite 60% of the samples being from historical specimens. We did not extensively investigate what was driving the stability differences between these clades, but the *Charmosyna* clade has longer internodes. The remaining uncertainty surrounding phylogenetic relationships with *Eos*, *Trichoglossus*, *Glossopsitta*, and *Psitteuteles iris,* will have ramifications for clarifying taxonomy and evolutionary history of the clade.

Taxonomic implications

Our study builds on previous phylogenetic work on the Loriini by further clarifying relationships and adding 58 previously unsampled taxa. We inferred a backbone phylogeny of relationships among genera that was fairly well-resolved with the exception of the clade containing *Trichoglossus*, *Psitteuteles iris*, *Eos*, and *Glossopsitta*. Our analyses corroborated recently proposed taxonomic changes where *Pseudeos cardinalis* was moved into *Pseudeos* from *Chalcopsitta*, and *Parvipsitta* was resurrected to contain *P. pusilla* and *P. porphyrocephala*, which were previously placed in *Glossopsitta* (Schweizer et al. 2015). In all of our trees *P. fuscata* and *P. cardinalis* were sister, and were in turn sister to *Chalcopsitta*. *Parvipsitta pusilla* and *P. porphyrocephala* were sister and not closely related to *G. coccina*. However, we found strong support for *P. pusilla* and *P. porphyrocephala* being sister to *Psitteuteles versicolor*, a novel result. *Psitteuteles versicolor* and *Parvipsitta* could be subsumed under a single genus. Irrespective of this taxonomic decision, the polyphyly of *Psitteuteles* will require that *P. goldei* and *P. iris* be moved into new genera. *Psitteuteles goldei* is sister to the clade containing *Trichoglossus*, *Eos*, and *Glossopsitta*. The taxonomic revision of *P. iris* will depend on how *Trichoglossus* is treated, as *P. iris* is nested within a geographically coherent clade of taxa distributed largely to the west of New Guinea. The clade containing *Charmosyna*, *Phigys*, and *Vini* represents a deep, diverse, and geographically widespread group. The species in these genera are collectively morphologically varied in terms of body size and shape, tail length, plumage, and sexual dimorphism (Forshaw et al. 1989), and based on our analyses, this variation does not neatly sort into groups. Overall, the taxonomic revision of this clade will present challenges with when and where to split or lump taxa into genera.

Relationships within species with multiple subspecies should be further clarified with more detailed multispecies coalescent modeling. Resolving relationships is particularly important within *Trichoglossus*, which harbors many taxa that are separated by short internodes. Within this radiation our analyses inferred a paraphyletic *T. haematodus* and *T. forsteni*, which is still included in *T. haematodus* by some taxonomic checklists (Clements et al. 2017; Dickinson and Remsen 2013). Support for relationships among species within *Lorius* were generally stable, but there were varying levels of support for relationships among the subspecies in the most diverse species in the genus, *Lorius lory*. As with deeper-level relationships within *Charmosyna*,

relationships among subspecies were well-supported. However, there were cases where we sampled multiple individuals per subspecies and these samples were not sister. The loci driving these relationships were not detected in our gene likelihood analyses. In general, our species tree analyses produced less-supported but similar relationships to our tree inferred from concatenated alignments, but additional phylogenetic signal is likely in the heterozygous sites that were dropped. Phasing these nucleotide sites may provide enough information to resolve challenging relationships particularly within *Trichoglossus*.

Funding

This study was funded by National Science Foundation awards to BTS (DEB-1655736) and MJA (DEB-1557051).

Supplementary Material

Data available from the Dryad Digital Repository: (will be provided in a later version).

Acknowledgments

We thank the following institutions and people for providing the material used in this study: AMNH (P. Sweet, T. Trombone, G. Rosen, A. Caragiulo), UWBM (S. Birks, R. Faucett, J. Klicka), USNM (B. Schmidt, H. James, G. Graves), ANWC (R. Palmer, L. Joseph), LSUMZ (D. Dittmann, S. Cardiff, R. Brumfield, F. Sheldon), FMNH (B. Marks, J. Bates, S. Hackett), KU (M. Robbins, R. Moyle). We also thank F. Burbrink for providing the R function Nia_Jax and analysis suggestions. We also thank J. Merwin, K. Provost, L. R. Moreira, B. Faircloth, C. Oliveros, and M. Harvey.

References

Amadon, D. 1943. Birds collected during the Whitney South Sea Expedition. LII, Notes on some non-passerine genera, 3. Am. Mus. Novit. 1237:1-22.

Andersen, M. J., Fatdal, L., Mauck III, W. M., Smith, B. T. 2017. An ornithological survey of Vanuatu on the islands of Éfaté, Malakula, Gaua, and Vanua Lava. Check List 13:755-782.

Andersen MJ, McCullough JM, Mauck WM III, Smith BT, Moyle RG. 2018. A phylogeny of kingfishers reveals an Indomalayan origin and elevated rates of diversification on oceanic islands. J. Biogeogr. 45:269–281.

Arcila, D., Ortí, G., Vari, R., Armbruster, J. W., Stiassny, M. L., Ko, K. D., Sabaj, M. H., Lundberg, J., Revell, L. J., Betancur-R, R. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. Nat. Ecol. Evol. 1:0020.

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

Borowiec, M.L. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ 4:e1660.

Briggs, A.W., Stenzel, U., Johnson, P.L., Green, R.E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M. and Pääbo, S. 2007. Patterns of damage in genomic DNA sequences from a Neandertal.Proc. Natl. Acad. Sci. U.S.A, 104:14616-14621.

Brown, J. M., Thomson, R. C. 2016. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst. Biol. 66:517-530.

Chakrabarty, P., Faircloth, B. C., Alda, F., Ludt, W. B., Mcmahan, C. D., Near, T. J., Dornburg, A., Albert, J. S., Arroyave, J., Stiassny, M. L., Sorenson, L. 2017. Phylogenomic systematics of ostariophysan fishes: ultraconserved elements support the surprising non-monophyly of characiformes. Syst. Biol. 66:881-895.

Chifman, J., Kubatko, L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics. 30:3317-3324.

Clements, J. F., Schulenberg, T. S., Iliff, M. J., Roberson, D., Fredericks, T. A., Sullivan, B. L., Wood, C. L. 2017. The eBird/Clements checklist of birds of the world: v2017. Downloaded from http://www.birds.cornell.edu/clementschecklist/download/

Dickinson, E. C., Remsen, Jr., J. V. 2013. The Howard and Moore Complete Checklist of the Birds of the World, Volume 1: Non-passerines. Ed. 4.

Enk, J. M., Devault, A. M., Kuch, M., Murgha, Y. E., Rouillard, J. M., Poinar, H. N. 2014. Ancient whole genome enrichment using baits built from modern DNA. Mol. Biol. Evol. 31:1292-1294.

Esselstyn, J. A., Oliveros, C. H., Swanson, M. T., Faircloth, B. C. 2017. Investigating difficult nodes in the placental mammal tree with expanded taxon sampling and thousands of ultraconserved elements. Genome Biol. Evol. 9:2308-2321.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., Glenn, T. C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61:717-726.

Faircloth, B. C. 2013. illumiprocessor: a trimmomatic wrapper for parallel adapter and quality trimming. http://dx.doi.org/10.6079/J9ILL.

Faircloth, B. C. 2015. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics 32*:786-788.

Faircloth, B. C., Branstetter, M. G., White, N. D., Brady, S. G. 2015. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Mol. Ecol. Resour. 15:489-501.

Fortes, G.G., Grandal-d'Anglade, A., Kolbe, B., Fernandes, D., Meleg, I.N., García-Vázquez, A., Pinto-Llona, A.C., Constantin, S., de Torres, T.J., Ortiz, J.E., Frischauf, C. 2016. Ancient DNA reveals differences in behaviour and sociality between brown bears and extinct cave bears. Mol. Ecol. 25:4907-4918.

Fox, J., Weisberg, S. 2019. An R Companion to Applied Regression, Third Edition, Sage.

Gilbert, P. S., Wu, J., Simon, M. W., Sinsheimer, J. S., Alfaro, M. E. 2018. Filtering nucleotide sites by phylogenetic signal to noise ratio increases confidence in the Neoaves phylogeny generated from ultraconserved elements. Mol. Phylogenet. Evol. 126:116-128.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307-321.

Hackett, S. J., Kimball, R. T., Reddy, S., Bowie, R. C., Braun, E. L., Braun, M. J. et al. 2008. A phylogenomic study of birds reveals their evolutionary history. Science. 320:1763-1768.

Harvey, M. G., Aleixo, A., Ribas, C. C., Brumfield, R. T. 2017. Habitat association predicts genetic diversity and population divergence in Amazonian birds. Am. Nat. 190:631-648.

Helgen, K.M., Pinto, C.M., Kays, R., Helgen, L.E., Tsuchiya, M.T., Quinn, A., Wilson, D.E., Maldonado, J.E. 2013. Taxonomic revision of the olingos (Bassaricyon), with description of a new species, the Olinguito. ZooKeys. 324:1-83.

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

Huang, H., Knowles, L. L. 2014. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. Syst. Biol. 65:357-365.

Hung, C. M., Shaner, P. J. L., Zink, R. M., Liu, W. C., Chu, T. C., Huang, W. S., Li, S. H. 2014. Drastic population fluctuations explain the rapid extinction of the passenger pigeon. Proc. Natl. Acad. Sci. U.S.A. 111:10636-10641.

Kehlmaier, C., Barlow, A., Hastings, A.K., Vamberger, M., Paijmans, J.L., Steadman, D.W., Albury, N.A., Franz, R., Hofreiter, M., Fritz, U. 2017. Tropical ancient DNA reveals relationships of the extinct Bahamian giant tortoise *Chelonoidis alburyorum*. *Proc. R. Soc. B. 284*:20162235.

Jiang, W., Chen, S. Y., Wang, H., Li, D. Z., Wiens, J. J. 2014. Should genes with missing data be excluded from phylogenetic analyses? Mol. Phylogenet. Evol. 80:308-318.

Kratter AW, Kirchman JJ, Steadman DW. 2006. Upland Bird Communities on Santo, Vanuatu, Southwest Pacific. Wilson J. Ornithol. 118: 295–308.

Kuhn, M. 2008. Caret package. J. Stat. Softw. 28:1-26.

Lanfear, R., Calcott, B., Kainer, D., Mayer, C., Stamatakis, A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol. 14: 82.

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., Calcott, B. 2016. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34:772-773.

Lemmon, A. R., Brown, J. M., Stanger-Hall, K., Lemmon, E. M. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58:130-145.

Li H., Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25:1754-60.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup. 2009. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 25:2078-2079.

Linck, E. B., Hanna, Z. R., Sellas, A., Dumbacher, J. P. 2017. Evaluating hybridization capture with RAD probes as a tool for museum genomics with historical bird specimens. Ecol. Evol. 7:4755–4767.

Malmström, H., Storå, J., Dalén, L., Holmlund, G., Götherström, A. 2005. Extensive human DNA contamination in extracts from ancient dog bones and teeth. Mol. Biol. Evol. 22:2040-2047.

Mayr, E. 1933. Birds collected during the Whitney South Sea Expedition. 24, Notes on Polynesian flycatchers and a revision of the genus *Clytorhynchus* Elliot. Am. Mus. Novit. 628:1-21.

Mayr, E. 1938. Birds collected during the Whitney South Seas Expedition, XL. Am. Mus. Novit. 522:1-22.

Mayr, E. 1942. Birds collected during the Whitney South Sea Expedition. 48, Notes on the Polynesian species of *Aplonis*. Am. Mus. Novit. 1166:1-8.

McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C. Brumfield, R.T. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol. 66:526-538.

McCormack, J. E., Tsai, W. L., Faircloth, B. C. 2016. Sequence capture of ultraconserved elements from bird museum specimens. Mol. Ecol. Resour. 16:1189-1203.

Mitchell, K. J., Llamas, B., Soubrier, J., Rawlence, N. J., Worthy, T. H., Wood, J., Lee, M.S., Cooper, A. 2014. Ancient DNA reveals elephant birds and kiwi are sister taxa and clarifies ratite bird evolution. Science 344:898-900.

Mitchell, K. J., Scanferla, A., Soibelzon, E., Bonini, R., Ochoa, J., Cooper, A. 2016. Ancient DNA from the extinct South American giant glyptodont *Doedicurus* sp. (Xenarthra: Glyptodontidae) reveals that glyptodonts evolved from Eocene armadillos. Mol. Ecol. 25:3499-3508.

Molloy, E.K., Warnow, T., 2017. To include or not to include: the impact of gene filtering on species tree estimation methods. Syst. Biol. 67:285-303.

Moyle, R.G., Oliveros, C.H., Andersen, M.J., Hosner, P.A., Benz, B.W., Manthey, J.D., Travers, S.L., Brown, R.M., Faircloth, B.C., 2016. Tectonic collision and uplift of Wallacea triggered the global songbird radiation. Nat. Commun. 7:12709.

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., Minh, B. Q. 2014. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268-274.

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., Harris, S. R. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial Genomics. 2.

Paijmans, J.L., Barnett, R., Gilbert, M.T.P., Zepeda-Mendoza, M.L., Reumer, J.W., de Vos, J., Zazula, G., Nagel, D., Baryshnikov, G.F., Leonard, J.A., Rohland, N. 2017. Evolutionary history of saber-toothed cats based on ancient mitogenomics. Cur. Biol. 27: 3330-3336.

Paradis, E., Claude, J., Strimmer, K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289-290.

Philippe, H., Snell, E. A., Bapteste, E., Lopez, P., Holland, P. W., Casane, D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21:1740-1752.

Provost, K. L., Joseph, L., Smith, B. T. 2018. Resolving a phylogenetic hypothesis for parrots: implications from systematics to conservation. Emu. 118:7-21.

Revell, L. J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3:217-223.

Robinson, D. F., Foulds, L. R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131-147.

Roure, B., Baurain, D., Philippe, H. 2012. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30:197-214.

Rothfels, C. J., Larsson, A., Kuo, L. Y., Korall, P., Chiou, W. L., Pryer, K. M. 2012. Overcoming deep roots, fast rates, and short internodes to resolve the ancient rapid radiation of eupolypod II ferns. Syst. Biol. 61:490-509.

Ruane, S., Austin, C. C. 2017. Phylogenomics using formalin‐fixed and 100+ year‐old intractable natural history specimens. Mol. Ecol. Resour. 17:1003-1008.

Pyron, R. A., Hendry, C. R., Chou, V. M., Lemmon, E. M., Lemmon, A. R., Burbrink, F. T. 2014. Effectiveness of phylogenomic data and coalescent species-tree methods for resolving difficult nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia). Mol. Phylogenet. Evol. 81:221-231.

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., Pääbo, S. 2012. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. PloS ONE *7*:e34131.

Schweizer, M., Wright, T. F., Peñalba, J. V., Schirtzinger, E. E., Joseph, L. 2015. Molecular phylogenetics suggests a New Guinean origin and frequent episodes of founder-event speciation in the nectarivorous lories and lorikeets (Aves: Psittaciformes). Mol. Phylogenet. Evol. 90:34-48.

Shavit Grievink, L., Penny, D., Holland, B. R. 2013. Missing data and influential sites: choice of sites for phylogenetic analysis can be as important as taxon sampling and model choice. Genome Biol. Evol. 5:681-687.

Shen, X. X., Hittinger, C. T., Rokas, A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1:126.

Simmons, M. P. 2012. Radical instability and spurious branch support by likelihood when applied to matrices with non-random distributions of missing data. Mol. Phylogenet. Evol. 62:472-484.

Simmons, M. P. 2014. A confounding effect of missing data on character conflict in maximum likelihood and Bayesian MCMC phylogenetic analyses. Mol. Phylogenet. Evol. 80:267-280.

Springer, M.S., Gatesy, J. 2018. On the importance of homology in the age of phylogenomics. Syst. Biodivers. 16:210-228.

Sorenson, M. D., Cooper, A., Paxinos, E. E., Quinn, T. W., James, H. F., Olson, S. L., Fleischer, R. C. 1999. Relationships of the extinct moa-nalos, flightless Hawaiian waterfowl, based on ancient DNA. Proc. Royal Soc. Lond. 266: 2187-2193.

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

Streicher, J. W., Schulte, J. A., Wiens, J. J. 2015. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. Syst. Biol. 65:128-145.

Swofford, D. L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). v. 4. Sinauer Associates, Sunderland, Massachusetts.

Thomas, R. H., Schaffner, W., Wilson, A. C., Pääbo, S. 1989. DNA phylogeny of the extinct marsupial wolf. Nature. 340:465.

Tin, M. M. Y., Rheindt, F. E., Cros, E., Mikheyev, A. S. 2015. Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. Mol. Ecol. Resour. 15:329-336.

Walker, J. F., Brown, J. W., Smith, S. A. 2018. Analyzing contentious relationships and outlier genes in phylogenomics. Syst. Biol. syy043, https://doi.org/10.1093/sysbio/syy043.

Wiens, J. J., Morrill, M. C. 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst. Biol. 60:719-731.

Xi, Z., Liu, L., Davis, C. C. 2015. The impact of missing data on species tree estimation. Mol. Biol. Evol. 33:838-860.

Yao, L., Li, H., Martin, R. D., Moreau, C. S., Malhi, R. S. 2017. Tracing the phylogeographic history of Southeast Asian long-tailed macaques through mitogenomes of museum specimens. Mol. Phylogenet. Evol. 116:227-238.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., Lam, T. T. Y. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8:28-36.

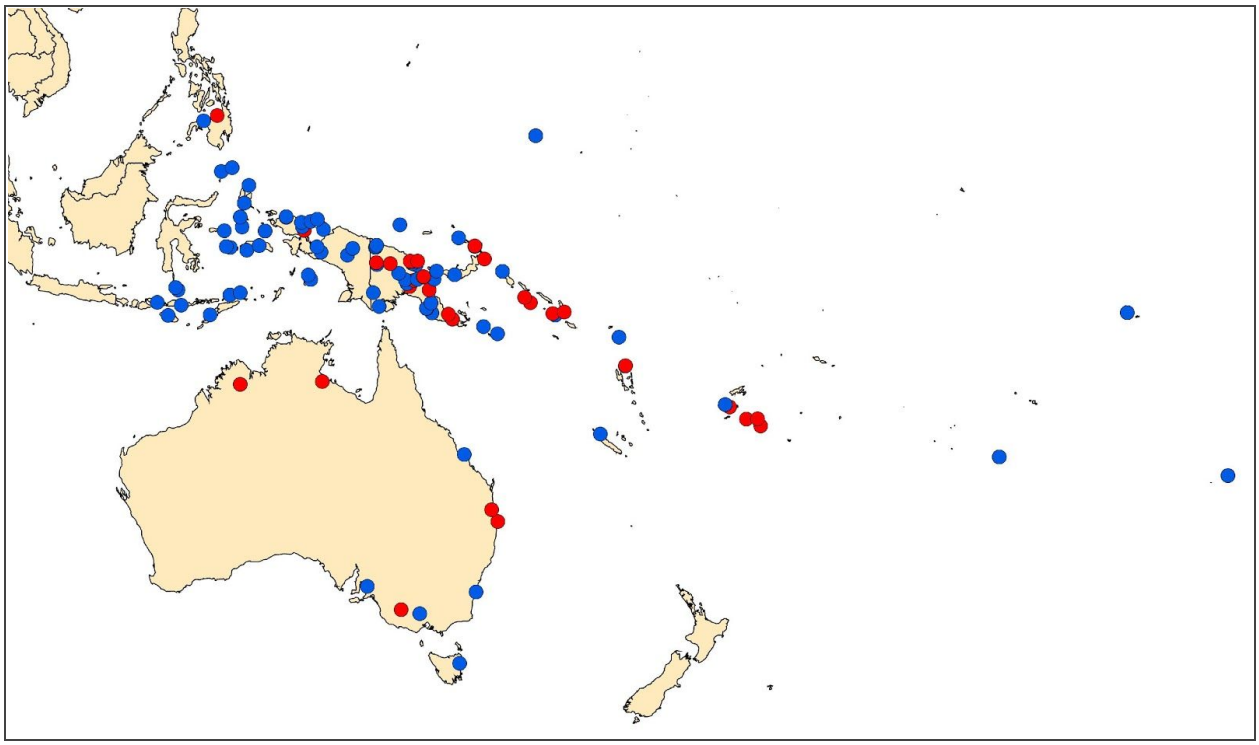

Figure 1. Sampling map of lory and lorikeet taxa used in this study. Colored symbols represent material that came from historical (blue) and modern (red) samples.

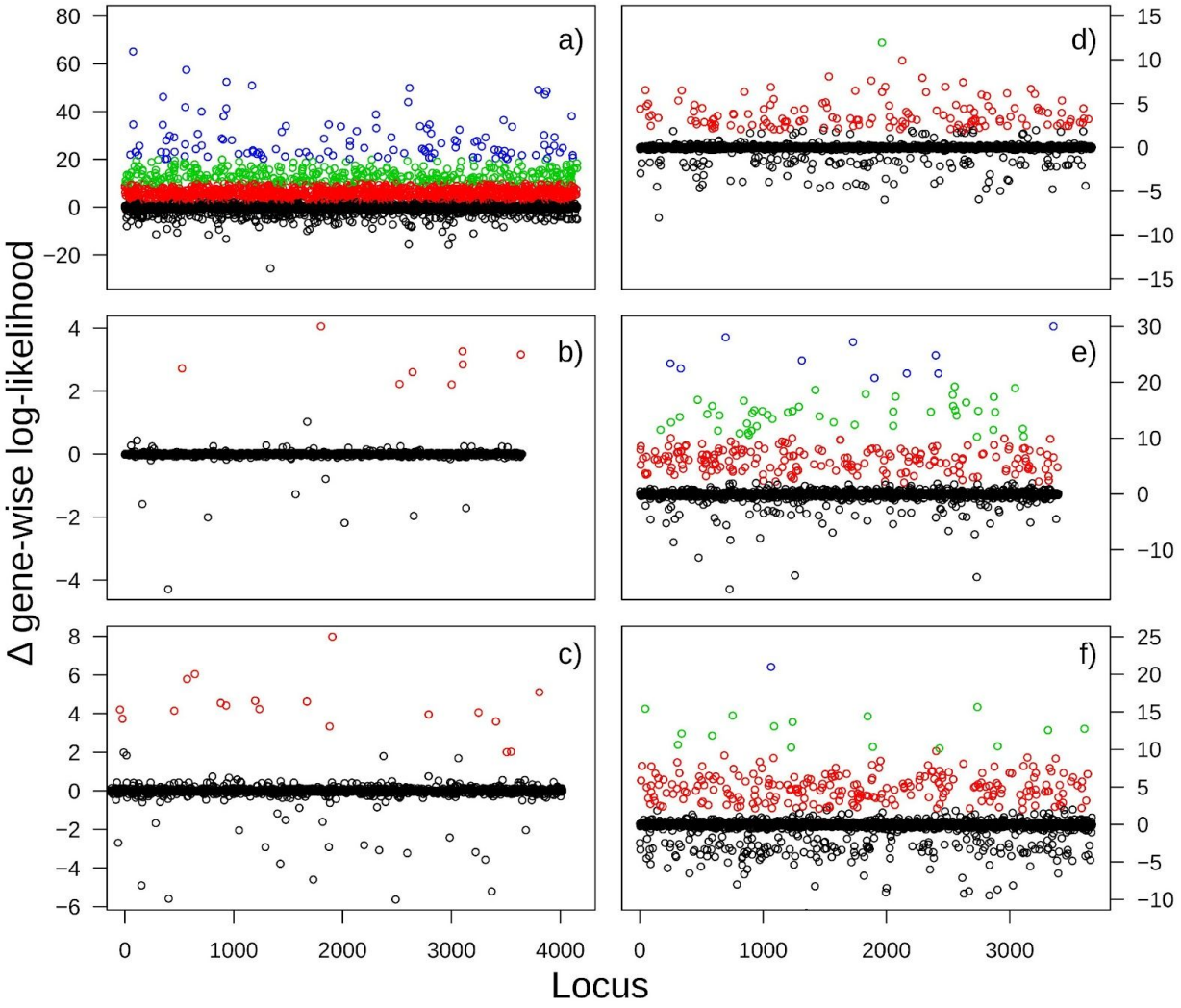

Figure 2. Likelihood plots showing Δ gene-wise log-likelihoods for topologies estimated with and without missing data. The y-axis is the Δ gene-wise log-likelihood and the x-axis represents individual loci in the concatenated alignment. Shown are the results for the six subclades assessed within Loriini using the filtered loci: a) *Eos*, *Trichoglossus*, and *Psitteuteles iris*, b) *Parvipsitta* and *Psitteuteles*, c) *Neospittacus*, d) *Chalcopsitta* and *Pseudeos*, e) *Lorius*, and f) *Charmosyna*. Points are colored according to the magnitude of the Δ gene-wise log-likelihood scores of 20+ (blue), 10–20 (green), 2–10 (red), and < 2 (black).

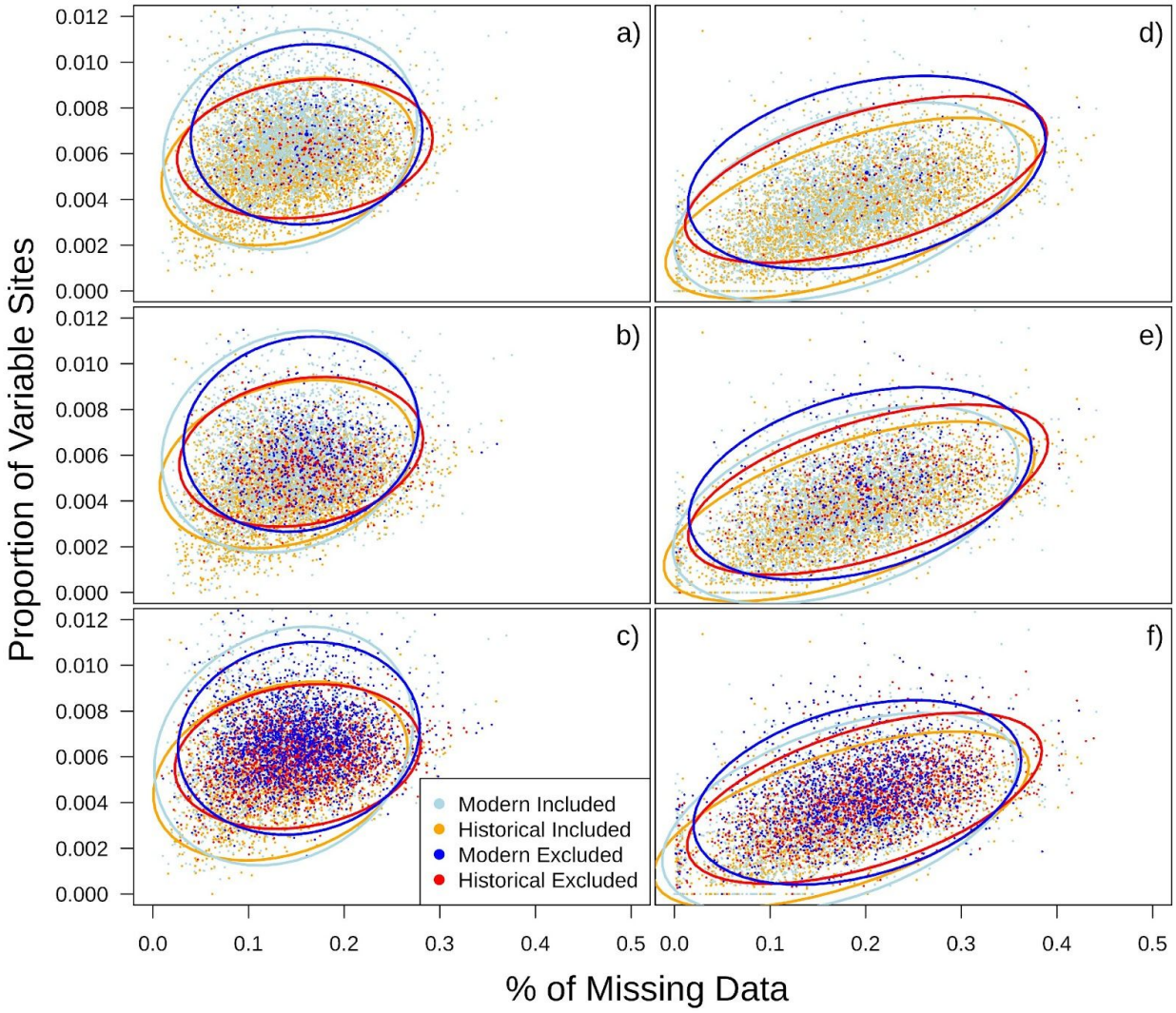

Figure 3. Modern samples have more variable sites than historical samples. The number of variable sites per individual per locus versus the percentage of missing data is plotted with 95% CI of ellipses. Ellipse centers are shown with larger circles. Points and ellipses are colored according to loci obtained from historical and modern samples, and those loci identified by the gene likelihood analyses: modern included (light blue), historical included (orange), modern excluded (dark blue), and historical excluded (red). Included are plots for the unfiltered (a–c) and filtered (d–f) datasets, plots with an increasing number of loci excluded in increments of the Δ gene-wise log-likelihood scores of 20+, 10+, and 2+ (top to bottom).

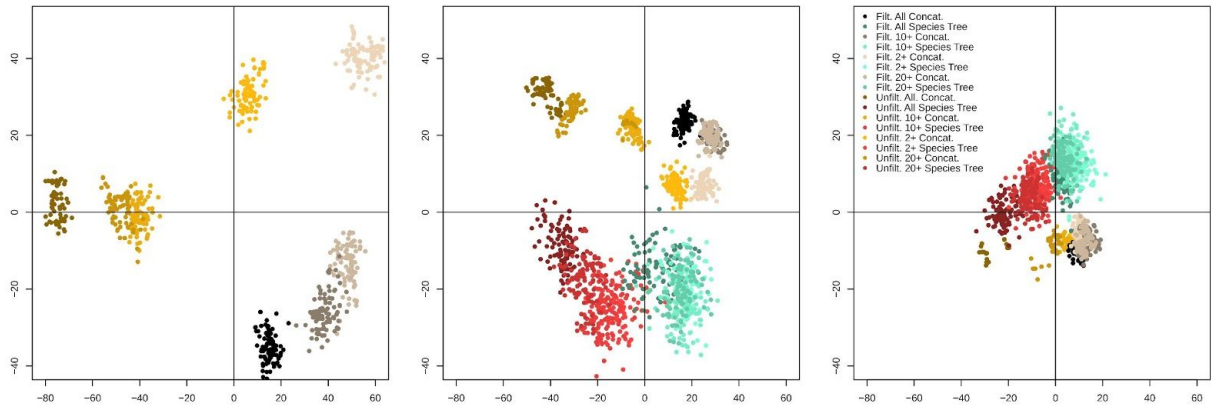

Figure 4. Multidimensional scaling of Robinson-Foulds distances among concatenated and multi-species coalescent topologies. From left to right are trees estimated with all species, only unique taxa, and just species. The alternative topologies in each plot were estimated with all loci and increasing number of loci excluded in increments of the Δ log-likelihood values of 20+, 10+, and 2+. Shown are 100 bootstraps trees estimated with the unfiltered (Unfilt.) and filtered (Filt.) datasets using concatenated (Concat.) and multi-species coalescent (Species Tree) tree building approaches.

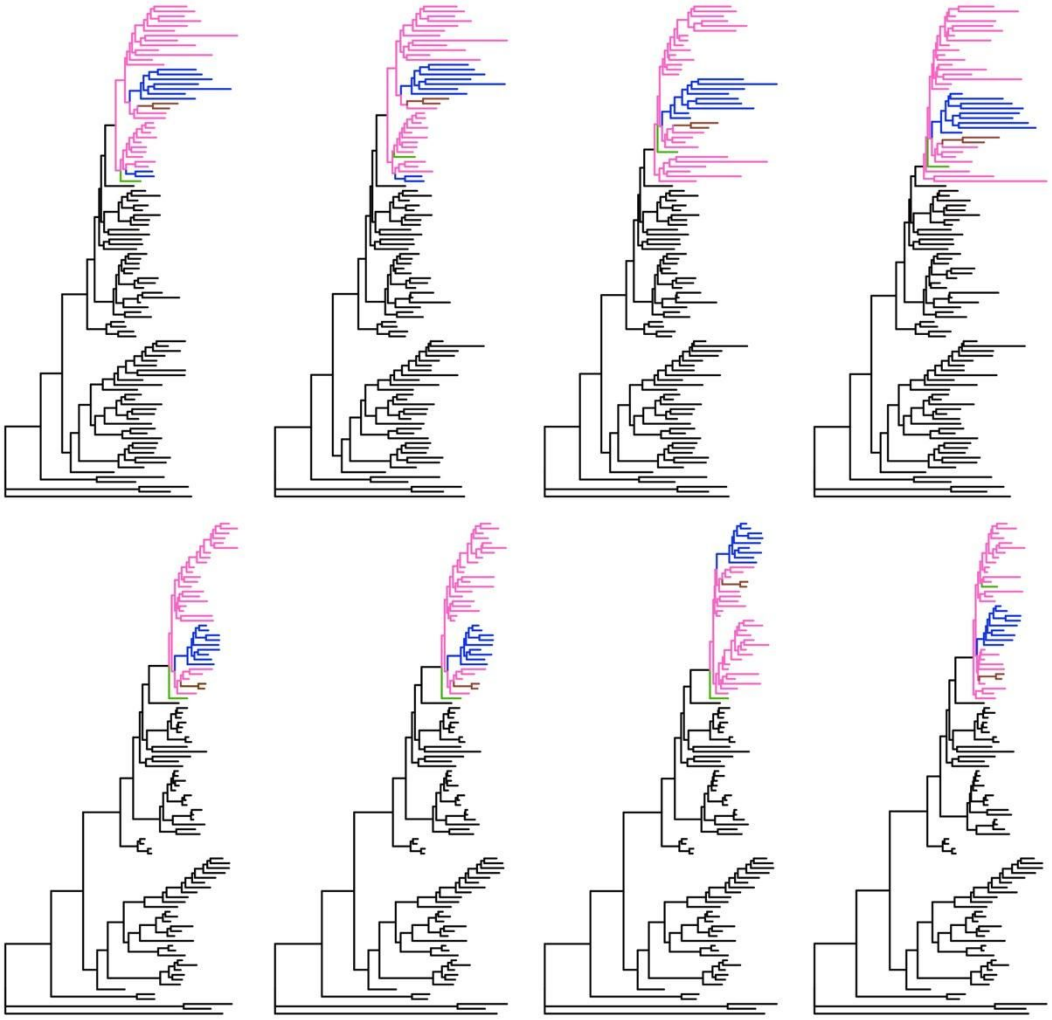

Figure 5. Alternative maximum likelihood trees estimated with the complete datasets and subsets of loci filtered based on Δ log-likelihood values. Shown are trees estimated with the unfiltered (top) and filtered (bottom) datasets with all loci and increasing number of loci excluded in increments of the Δ log-likelihood values of 20+, 10+, and 2+ (from left to right). Colored branches highlight an example of topological instability and variation in branch lengths across topologies. Branches are colored based on clades and tips that correspond to the genera *Eos* (blue), *Trichoglossus* (pink), *Glossopsitta* (green), and *Psitteuteles iris* (brown).

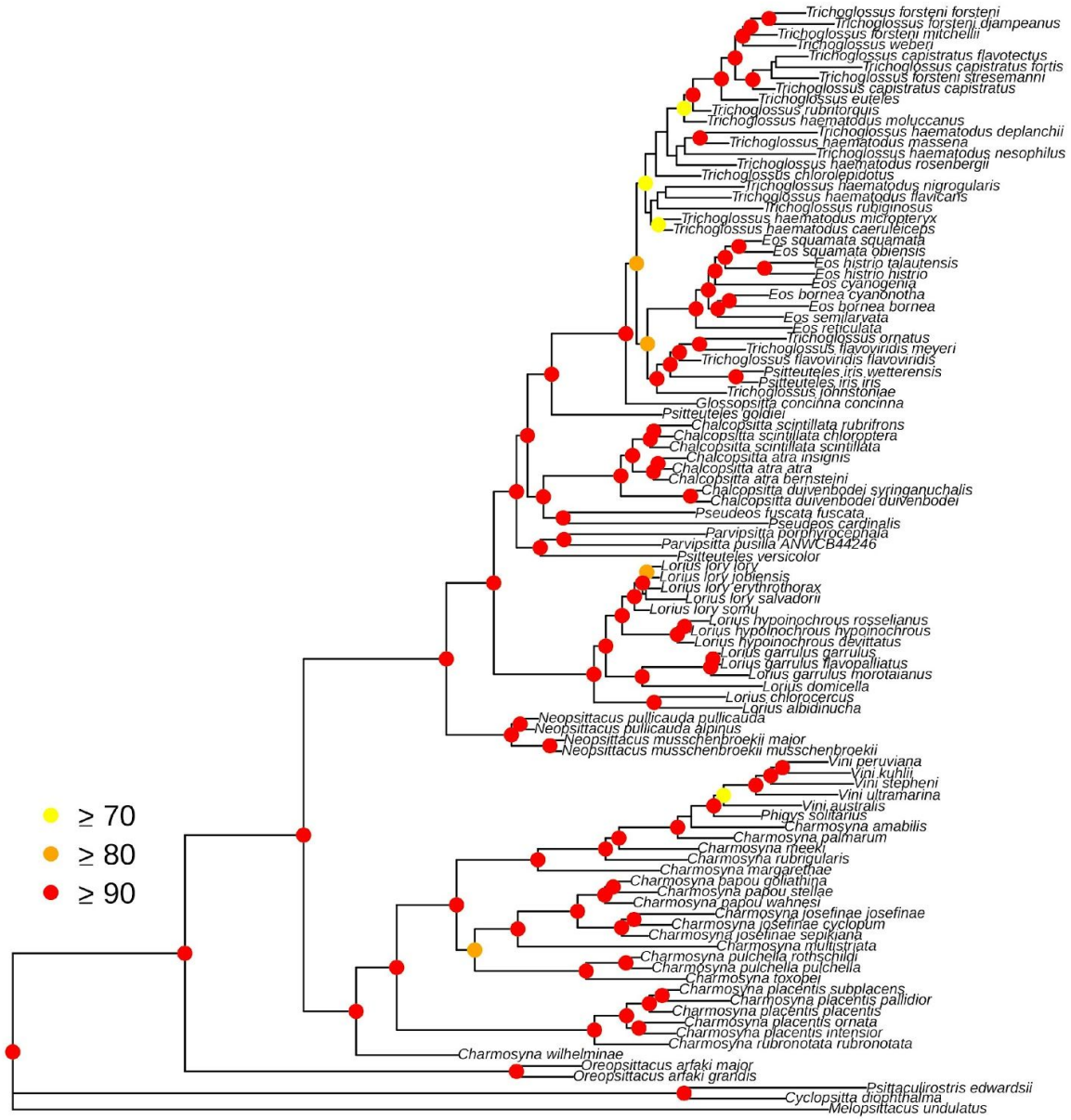

Figure 6. Maximum likelihood tree containing unique taxa in Loriini. The tree was inferred from a concatenated alignment where loci identified with the gene likelihood analysis with Δ gene-wise log-likelihood scores of 20+ were excluded. On each node are shown rapid bootstrap values colored based on support level.

SUPPLEMENTARY TABLE LEGENDS

Supplementary Table 1. Metadata for Loriini samples used in this study. (This table is not included in this submission).

Supplementary Table 2. Neural Net model output that assessed predictors of Δ gene-wise log-likelihood scores. Shown are the variable importance of the five alignment statistics included in the neural net model for each subclade in Loriini. Models were built for both Eos/Trichoglossus/Glossopsitta datasets (see methods) but since the results were similar we only report one for convenience. The per locus statistics were as follows: parsimony informative sites (PIS), alignment length, number of variables sites, percentage of missing data, and GC content. The sample size for each subclade was the total number of alignments and 75%/25% was used to train/test each model. Reported are R^2 and mean square error (MSE).

Unfiltered	**PIS**	**Alignment Length**	**No. Variable sites**	**Missing percent**	**GC content**	**R^2**	**MSE**
Chalcopsitta/Pseudeos	50.76	20.24	1.07	19.83	8.10	0.00	2.01
Charmosyna	37.03	20.21	30.70	10.12	1.94	0.00	7.57
Eos/Trichoglossus/Glossopsitta	48.18	11.28	17.22	17.89	5.42	0.02	30.65
Parvipsitta/Psitteuteles	24.07	37.78	17.11	19.66	1.38	-0.01	0.03
Lorius	46.87	13.26	14.03	15.19	10.66	0.04	6.13
Neopsittacus	9.59	7.09	47.64	21.61	14.07	0.02	0.00
Filtered	**PIS**	**Alignment Length**	**No. Variable sites**	**Missing percent**	**GC content**	**R^2**	**MSE**
Chalcopsitta/Pseudeos	60.09	4.62	15.85	13.04	6.40	0.01	1.39
Charmosyna	37.43	2.24	36.47	9.92	13.94	0.00	4.08
Eos/Trichoglossus/Glossopsitta	45.11	13.81	31.44	5.79	3.85	0.05	5.57
Parvipsitta/Psitteuteles	14.57	28.82	6.27	40.72	9.62	0.00	0.02
Lorius	95.82	3.39	0.40	0.23	0.16	0.08	5.43
Neopsittacus	58.75	5.05	23.31	2.11	10.78	0.13	2.28

Supplementary Table 3. Pairwise comparison of ellipse overlap among loci from modern and historical samples. 95% confidence intervals for the number of variable sites per individual per locus versus percentage of missing data are shown in Figure 3. The percentage of overlap among ellipses is reported in this table modern included, historical included, modern excluded, and historical excluded. Included are results for the unfiltered and filtered datasets, with an increasing number of loci excluded in increments of the Δ gene-wise log-likelihood scores of 20+, 10+, and 2+.

Unfiltered 20+	**Included modern**	**Excluded modern**	**Included historical**	**Excluded historical**
Included modern	--	0.73	0.72	0.59
Excluded modern	0.73	--	0.65	0.59
Included historical	0.72	0.65	--	0.77
Excluded historical	0.59	0.59	0.77	--
Unfiltered 10+	**Included modern**	**Excluded modern**	**Included historical**	**Excluded historical**
Included modern	--	0.81	0.72	0.63
Excluded modern	0.81	--	0.65	0.74
Included historical	0.72	0.65	--	0.80
Excluded historical	0.63	0.74	0.80	--
Unfiltered 2+	**Included modern**	**Excluded modern**	**Included historical**	**Excluded historical**
Included modern	--	0.74	0.70	0.56
Excluded modern	0.74	--	0.63	0.72
Included historical	0.70	0.63	--	0.75
Excluded historical	0.56	0.72	0.75	--
Filtered 20+	**Included modern**	**Excluded modern**	**Included historical**	**Excluded historical**
Included modern	--	0.67	0.80	0.71
Excluded modern	0.67	--	0.60	0.75
Included historical	0.80	0.60	--	0.63
Excluded historical	0.71	0.75	0.63	--
Filtered 10+	**Included modern**	**Excluded modern**	**Included historical**	**Excluded historical**
Included modern	--	0.75	0.80	0.74
Excluded modern	0.75	--	0.66	0.78

Included historical	0.80	0.66	--	0.72
Excluded historical	0.74	0.78	0.72	--
Filtered 2+	**Included modern**	**Excluded modern**	**Included historical**	**Excluded historical**
Included modern	--	0.76	0.79	0.73
Excluded modern	0.76	--	0.66	0.80
Included historical	0.79	0.66	--	0.72
Excluded historical	0.73	0.80	0.72	--

SUPPLEMENTARY FIGURE LEGENDS (These figures are not included in this submission)

Supplementary Figure S1. Boxplots of alignment statistics for the unfiltered (left plots) and filtered (right plots) UCE loci. Shown are the per locus number of parsimony informative sites (PIS), length (bp), and percentage of missing data. Within each plot are pairwise comparisons of the included (coral-colored gradient) and excluded (blue-colored gradient) loci ranging from Δ gene-wise log-likelihood scores of 2+, 10+, and 20+ (from left to right).

Supplementary Figure S2. Likelihood plots showing Δ gene-wise log-likelihoods for topologies estimated with and without missing data. The y-axis is the Δ gene-wise log-likelihood and the x-axis represents individual loci in the concatenated alignment. Shown are the results for the six subclades assessed within Loriini using the filtered loci: a) *Eos*, *Trichoglossus*, and *Psitteuteles iris*, b) *Glossopsitta* and *Psitteuteles*, c) *Neospittacus*, d) *Chalcopsitta* and *Pseudeos*, e) *Lorius*, and f) *Charmosyna*. Points are colored according to the magnitude of the Δ log-likelihoods of 20+ (blue), 10-20 (green), 2-10 (red), and < 2 (black).

Supplementary Figure S3. Maximum likelihood trees containing all samples using the unfiltered loci. The tree was inferred from a concatenated alignment with all loci (a) and an increasing number of loci excluded in increments of the Δ gene-wise log-likelihood scores of 20+ (b), 10+ (c), and 2+ (d). On each node are shown bootstrap values.

Supplementary Figure S4. Maximum likelihood trees with taxonomically unique samples using the unfiltered loci. The tree was inferred from a concatenated alignment with all loci (a) and an increasing number of loci excluded in increments of the Δ gene-wise log-likelihood scores of 20+ (b), 10+ (c), and 2+ (d). On each node are shown bootstrap values.

Supplementary Figure S5. Maximum likelihood trees with only species using the unfiltered loci. The tree was inferred from a concatenated alignment with all loci (a) and an increasing number of loci excluded in increments of the Δ gene-wise log-likelihood scores of 20+ (b), 10+ (c), and 2+ (d). On each node are shown bootstrap values.

Supplementary Figure S6. Species trees with taxonomically unique samples using the unfiltered loci. The tree was inferred from all loci (a) and an increasing number of loci excluded in increments of the Δ gene-wise log-likelihood scores of 20+ (b), 10+ (c), and 2+ (d). On each node are shown bootstrap values.

Supplementary Figure S7. Species trees with only species using the unfiltered loci. The tree was inferred from a concatenated alignment with all loci (a) and an increasing number of loci excluded in increments of the Δ gene-wise log-likelihood scores of 20+ (b), 10+ (c), and 2+ (d). On each node are shown bootstrap values.

Supplementary Figure S8. Maximum likelihood trees containing all samples using the filtered loci. The tree was inferred from a concatenated alignment with all loci (a) and an

increasing number of loci excluded in increments of the Δ gene-wise log-likelihood scores of 20+ (b), 10+ (c), and 2+ (d). On each node are shown bootstrap values.

Supplementary Figure S9. Maximum likelihood trees with taxonomically unique samples using the filtered loci. The tree was inferred from a concatenated alignment with all loci (a) and an increasing number of loci excluded in increments of the Δ gene-wise log-likelihood scores of 20+ (b), 10+ (c), and 2+ (d). On each node are shown bootstrap values.

Supplementary Figure S10. Maximum likelihood trees with only species using the filtered loci. The tree was inferred from a concatenated alignment with all loci (a) and an increasing number of loci excluded in increments of the Δ gene-wise log-likelihood scores of 20+ (b), 10+ (c), and 2+ (d). On each node are shown bootstrap values.

Supplementary Figure S11. Species trees with taxonomically unique samples using the filtered loci. The tree was inferred from all loci (a) and an increasing number of loci excluded in increments of the Δ gene-wise log-likelihood scores of 20+ (b), 10+ (c), and 2+ (d). On each node are shown bootstrap values.

Supplementary Figure S12. Species trees with only species using the filtered loci. The tree was inferred from a concatenated alignment with all loci (a) and an increasing number of loci excluded in increments of the Δ gene-wise log-likelihood scores of 20+ (b), 10+ (c), and 2+ (d). On each node are shown bootstrap values.