Avoiding Missing Data Biases in Phylogenomic Inference: An Empirical Study in the Landfowl (Aves: Galliformes)

Peter A. Hosner,*1 Brant C. Faircloth,2 Travis C. Glenn,3 Edward L. Braun,1 and Rebecca T. Kimball1

1Department of Biology, University of Florida
2Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge
3Department of Environmental Health Science, University of Georgia

*Corresponding author: E-mail: hosner@ufl.edu.
Associate editor: Claudia Russo

Abstract

Production of massive DNA sequence data sets is transforming phylogenetic inference, but best practices for analyzing such data sets are not well established. One uncertainty is robustness to missing data, particularly in coalescent frameworks. To understand the effects of increasing matrix size and loci at the cost of increasing missing data, we produced a 90 taxon, 2.2 megabase, 4,800 locus sequence matrix of landfowl using target capture of ultraconserved elements. We then compared phylogenies estimated with concatenated maximum likelihood, quartet-based methods executed on concatenated matrices and gene tree reconciliation methods, across five thresholds of missing data. Results of maximum likelihood and quartet analyses were similar, well resolved, and demonstrated increasing support with increasing matrix size and sparseness. Conversely, gene tree reconciliation produced unexpected relationships when we included all informative loci, with certain taxa placed toward the root compared with other approaches. Inspection of these taxa identified a prevalence of short average contigs, which potentially biased gene tree inference and caused erroneous results in gene tree reconciliation. This suggests that the more problematic missing data in gene tree–based analyses are partial sequences rather than entire missing sequences from locus alignments. Limiting gene tree reconciliation to the most informative loci solved this problem, producing well-supported topologies congruent with concatenation and quartet methods. Collectively, our analyses provide a well-resolved phylogeny of landfowl, including strong support for previously problematic relationships such as those among junglefowl (Gallus), and clarify the position of two enigmatic galliform genera (Lerna, Melanoperdix) not sampled in previous molecular phylogenetic studies.

Key words: bias, coalescent, Coturnix, Gallus, Meleagris, ultraconserved elements.

Introduction

A major challenge in phylogenetics is resolving historical relationships when little time has passed between speciation events, because few DNA substitutions accrue during short time intervals (Braun and Kimball 2001; Whitfield and Lockhart 2007; Moyle et al. 2009; Wagner et al. 2013). Those few characters that do support the correct topology may then be subsequently overwritten, obscuring phylogenetic signal (Philippe et al. 2011; Patel et al. 2013). Genome-wide sequence data sets have shown great promise in resolving challenging short internal nodes in phylogenetic trees by providing millions of nucleotides and thousands of unlinked loci suitable for analyses (Faircloth et al. 2012; Lemmon EM and Lemmon AR 2013; McCormack and Faircloth 2013; Jarvis et al. 2014; Misof et al. 2014; Prum et al. 2015). However, analyses of these large, heterogeneous, genome-scale data sets are complicated—especially in light of the discordant phylogenetic signal contained within independent loci (Maddison 1997; Kubatko and Degnan 2007; Edwards 2009; Kimball et al. 2013).

A confounding problem is that evolutionary events leading to series of short internodes potentially produce an area of tree space, known as the anomaly zone, where the most common gene trees conflict with the true species tree (Degnan and Rosenberg 2006, 2009; Kubatko and Degnan 2007; Degnan 2013). Phylogenetic inference of concatenated alignments may be positively misleading in the anomaly zone (Kubatko and Degnan 2007; Roch and Steel 2014), and one common solution to this problem requires analyses in multispecies coalescent frameworks (Edwards et al. 2007; Liu and Edwards 2009). Although the theoretical advantages of coalescent approaches are clear, the strengths and weaknesses of available coalescent-based approaches when applied to empirical data remain murky (Reid et al. 2014; Liu et al. 2015; Roch and Warnow 2015; Tonini et al. 2015; Warnow 2015). Moreover, it is uncertain how often systematic error in estimates of species trees can be attributed to anomalous gene trees instead of other potential sources of error in phylogenetic inference, such as model violations or character limitation (Rannala and Yang 2008; Nabholz et al. 2011; Salichos and Rokas 2013; Betancur-R et al. 2014; Gatesy and Springer 2014; Jarvis et al. 2014).

A specific issue when analyzing phylogenomic data is deciding upon and justifying appropriate thresholds for locus inclusion and the presence of missing data (Philippe 2004; Wiens and Morrill 2011; McCormack et al. 2013; Wagner et al. 2013; Huang and Knowles 2014; Streicher et al. 2016).
There are three root causes for missing data in phylogenomic alignments: 1) Stochasticity inherent in collecting data across thousands of loci, where not all loci are detected in all genomic libraries; 2) variable sequence yield among sample libraries leading to missing data across alignments; and 3) biological processes including insertions, deletions, and other chromosomal changes. Ideally, samples with lower overall yields can be reprepared and resequenced (reducing the first two types of missing data above), but in practice sample availability, expense, and project timelines often limit resequencing, especially when researchers must rely on low molecular weight historical DNA for some taxa (Mundy et al. 1997; Knapp and Hofreiter 2010). To take full advantage of the statistical power that accompanies thousands of unlinked loci and millions of nucleotides, pragmatically, some proportion of missing data must be permitted when alignments grow beyond a few focal taxa. Thus, there is a direct tradeoff between increasing locus number and/or total alignment size, which may aid phylogenetic inference, and increasing proportions of missing data, which may hinder it (Philippe 2004; Wiens and Morrill 2011; Wagner et al. 2013; Huang and Knowles 2014; Streicher et al. 2016).

Effects of missing data on phylogenetic studies have been a major focus of research, including work at phylogenomic scales, but most studies have concentrated on maximum likelihood (ML) and Bayesian inference of concatenated sequence alignments (Philippe 2004; Sanderson et al. 2010; Wiens and Morrill 2011; Roure et al. 2012; Streicher et al. 2016). In general, analyses of concatenated data are robust to large proportions of missing data (Burleigh et al. 2015), as long as the data matrix contains sufficient overlap among taxa to find the true tree (Sanderson et al. 2010). Drawbacks of including sites/loci with missing data in alignments include increased risk of systematic error and increased computation time without concomitant improvement in results (Rannala and Yang 2008; Kumar et al. 2012; Lemmon AM and Lemmon ER 2013; Simmons 2014). However, other issues such as taxon sampling and model choice often appear to be more important than the amount of missing data in analyses of concatenated alignments (Roure et al. 2012).

Multispecies coalescent methods may be more susceptible to the negative effects of missing data than concatenated approaches, although this idea remains largely untested, especially with empirical data (Edwards 2009; Leaché and Rannala 2011; Wiens and Morrill 2011). Coalescent models have the benefit of accommodating differing phylogenetic signals among loci due to incomplete lineage sorting (ILS), and will in theory infer correct species trees in the anomaly zone where concatenation is positively misleading (Liu and Edwards 2009; Roch and Steel 2014, but see also Sun et al. 2014; Warnow 2015). Although theoretically desirable, many multispecies coalescent approaches have severe limitations when applied to genome-scale data. Simultaneous estimation of gene trees and species trees (Edwards et al. 2007; Heled and Drummond 2010) performs well in simulations (Leaché and Rannala 2011) but that approach is computationally intractable for large numbers of taxa and thousands of loci. Gene tree reconciliation analyses are computationally feasible when applied to genome-scale data (Liu and Edwards 2009; Mirarab, Reaz, et al. 2014; Roch and Warnow 2015), but estimating reliable gene genealogies to input to these methods is difficult due to character limitation and other problems (Rosenfeld et al. 2012; Gatesy and Springer 2014; Mirarab, Bayzid, Boussau, et al. 2014; Mirarab, Bayzid, and Warnow 2014; Springer and Gatesy 2016). Thus, it is possible that gene tree reconciliation methods are more sensitive to the ill effects of missing data than concatenation. Another intriguing option are quartet-based analyses that are consistent under ILS, but bypass the problematic stage of gene tree estimation (DeGiorgio and Degnan 2010; Chifman and Kubatko 2014). These approaches have received little testing with simulated or empirical data, so their relative performance when compared with concatenated ML and other coalescent methods are unclear (DeGiorgio et al. 2014; Sun et al. 2014).

Target capture of conserved genomic regions (Faircloth et al. 2012; Lemmon et al. 2012) combined with massively parallel sequencing produce data matrices containing thousands of unlinked loci distributed across the genome suitable for phylogenetic inference. These markers can be generated efficiently, cost-effectively, and are useful across deeper evolutionary scales than restriction enzyme–based reduced-representation libraries (Rubin et al. 2012). One class of conserved genomic regions, ultraconserved elements (UCEs), has been used for phylogenetic reconstruction at a variety of scales in vertebrates, from resolving relationships among major tetrapod lineages (Crawford et al. 2012, 2015; McCormack et al. 2012, 2013; Faircloth et al. 2013; Green et al. 2014; Jarvis et al. 2014; Streicher et al. 2016) to finescale vertebrate phyleogeography (Smith et al. 2014). UCEs feature a conserved core region with low variation (Bejerano et al. 2004) flanked by more variable sites. Conserved core regions are useful as probe targets, whereas flanking regions are variable and useful for inferring historical relationships (Faircloth et al. 2012).

Target capture is efficient, but locus recovery depends on probe design, sample quality, and variables such as genome size and repeat content that affect the library preparation and/or enrichment procedures (Knapp and Hofreiter 2010; Mamanova et al. 2010; Faircloth et al. 2012). Following target capture, entire loci will be missing for some taxa due to stochastic and biological factors. Thus, if the goal of a study is to produce a complete matrix at the locus level, adding taxa will reduce the number of completely sampled loci available for inclusion. An alternative strategy is to allow inclusion of loci that lack sequence data for a subset of taxa, which increases matrix size and locus count substantially, but at the cost of increasing missing data for some taxa (Streicher et al. 2016). Here, we designate a pattern of missing data where entire loci are missing for certain taxa as “type I” missing data (fig. 1). Studies using UCEs have favored complete or mostly complete matrices at the locus level, minimizing type I missing data (Crawford et al. 2012; Faircloth et al. 2013; Smith et al. 2014; Sun et al. 2014). Recently, Streicher et al. (2016) explored the effects of including large proportions of type I missing data in UCE phylogenomic analyses. In concatenated frameworks, they found that support increased substantially with...
increased taxon sampling and type I missing data. In coalescent frameworks, they found slightly greater support when type I missing data were lowest; however, all coalescent results suffered low bootstrap support.

Another characteristic of target capture is that the length of contigs recovered varies substantially among samples. Samples with more efficient sequence capture and greater depth of sequencing coverage typically produce longer contigs, on average, across the data set (Faircloth et al. 2013; McCormack et al. 2013). Because the majority of informative sites are found in flanking regions of UCEs (Faircloth et al. 2012), taxa having shorter average contig lengths (low N50) contain not only more missing data cells but also fewer informative sites relative to taxa having longer average contig lengths. Here, we designate this pattern of missing data where certain taxa have partial sequences for certain loci as "type II" missing data. Type II missing data could lead to errors when inferring individual gene trees for downstream analysis using gene tree reconciliation techniques (fig. 1; Simmons 2014). This potential bias caused by variation in average contig length is not limited to target capture methods or UCEs; variation in contig lengths across whole genome assemblies or other reduced representation genomic sequencing efforts can have similar effects. Type II missing data are rarely considered in large-scale phylogenies, and it is little understood how it may affect empirical phylogenetic inference of groups of interest.

The avian order Galliformes (landfowl) includes the most agriculturally important birds (the chicken [Gallus gallus] and turkey [Meleagris gallopavo], as well as Japanese quail [Coturnix japonica] and guineafowl [Numida meleagris]). The chicken is one of the premier model systems for developmental biology (Le Douarin and Dieterlen-Liévre 2012; Hirst and Marcelle 2015) and several different galliform taxa have been used extensively in behavioral research (e.g., peafowl [Pavo spp.], pheasants [Phasianus] and relatives, and grouse [Tetraoninae]; see also Kimball et al. 2011). Other landfowl are economically important game species (e.g., ring-necked pheasant [Phasianus], several partridge species [Perdix perdix, Alectoris chukar], and northern bobwhite [Colinus virginianus]). As a group, landfowl are disproportionately threatened by habitat destruction and unregulated overharvest—approximately 10% of galliform species are listed as endangered/critically endangered on the IUCN Red List (BirdLife International 2012). Yet, despite their importance in many areas of research, the galliform phylogeny is poorly resolved at many key nodes (Wang et al. 2013; Kimball and Braun 2014). For example, the identity of the sister taxon/group of Ga. gallus, arguably the best studied bird species in the world, still has not been resolved with confidence (Wang et al. 2013; Kimball and Braun 2014; Meiklejohn et al. 2014).

Two factors appear to limit progress toward inferring a robust galliform tree of life. First, galliforms appear to have undergone successive rapid radiations, and previous multilocus studies lacked appropriate resolution and produced conflicting results from concatenated and coalescent approaches (reviewed by Wang et al. 2013, Kimball and Braun 2014). Second, an inordinate number of galliform species are threatened by habitat destruction and overharvest (BirdLife International 2012) limiting availability of fresh tissues for key species of evolutionary interest (i.e., those not bred in captivity). Fortunately, choosing UCEs as phylogenomic markers address both these major problems in reconstructing galliform phylogeny: UCEs have demonstrated sufficient phylogenetic signal to resolve several nodes in a galliform clade that were ambiguous in previous multilocus

---

<table>
<thead>
<tr>
<th>Type I missing data</th>
<th>Type II missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>True tree</td>
<td>True tree</td>
</tr>
<tr>
<td>locus i</td>
<td>locus i</td>
</tr>
<tr>
<td>locus j</td>
<td>locus j</td>
</tr>
<tr>
<td>locus k</td>
<td>locus k</td>
</tr>
<tr>
<td>Inferred concatenated species tree</td>
<td>Inferred concatenated species tree</td>
</tr>
<tr>
<td>True genomic sequence</td>
<td>True genomic sequence</td>
</tr>
<tr>
<td>Data recovered from target capture</td>
<td>Data recovered from target capture</td>
</tr>
<tr>
<td>Inferred species tree</td>
<td>Inferred species tree</td>
</tr>
</tbody>
</table>

---

**Fig. 1.** A hypothetical example of how missing data within loci may bias gene tree estimation when contig lengths are shorter on average for a given taxon across a multilocus data set. In this example, taxon D has shorter contig lengths for loci i, j, and k. In two of the three gene trees, taxon D is mistakenly placed sister to B + C (because of incomplete character sampling), when rooting to outgroup taxon A. The downstream result is that taxon D is placed sister to taxon B + taxon C in gene tree reconciliation analyses, rather than sister to taxon C as in the true tree. In this example, when data are complete (true genomic sequence), or when concatenation is used, the correct tree is inferred.
Avoiding Missing Data Biases

Table 1. Sequence Summary Statistics Comparing Concatenated UCE Matrices of Five Thresholds of Completeness.

<table>
<thead>
<tr>
<th>Alignment</th>
<th>0% Missing Taxa in Loci</th>
<th>&lt;5% Missing Taxa in Loci</th>
<th>&lt;25% Missing Taxa in Loci</th>
<th>&lt;50% Missing Taxa in Loci</th>
<th>Total Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loci</td>
<td>140</td>
<td>1,740</td>
<td>3,361</td>
<td>3,919</td>
<td>4,817</td>
</tr>
<tr>
<td>Informative loci</td>
<td>140</td>
<td>1,739</td>
<td>3,329</td>
<td>3,868</td>
<td>4,638</td>
</tr>
<tr>
<td>Base pair</td>
<td>75,998</td>
<td>838,164</td>
<td>1,416,592</td>
<td>1,573,308</td>
<td>2,208,355</td>
</tr>
<tr>
<td>Variable sites</td>
<td>16,623</td>
<td>171,986</td>
<td>270,380</td>
<td>289,840</td>
<td>363,562</td>
</tr>
<tr>
<td>Informative sites</td>
<td>10,506</td>
<td>105,886</td>
<td>161,224</td>
<td>170,913</td>
<td>179,676</td>
</tr>
<tr>
<td>Average bootstrap %</td>
<td>95.0</td>
<td>98.8</td>
<td>98.6</td>
<td>98.8</td>
<td>98.8</td>
</tr>
<tr>
<td>Partitions</td>
<td>7</td>
<td>28</td>
<td>27</td>
<td>45</td>
<td>34</td>
</tr>
<tr>
<td>% Missing sites</td>
<td>10%</td>
<td>12%</td>
<td>15%</td>
<td>18%</td>
<td>39%</td>
</tr>
<tr>
<td>% Informative sites</td>
<td>14%</td>
<td>13%</td>
<td>12%</td>
<td>11%</td>
<td>8%</td>
</tr>
<tr>
<td>% Variable sites</td>
<td>22%</td>
<td>21%</td>
<td>19%</td>
<td>19%</td>
<td>17%</td>
</tr>
<tr>
<td>Average locus length</td>
<td>536</td>
<td>475</td>
<td>417</td>
<td>397</td>
<td>452</td>
</tr>
</tbody>
</table>

studies (Sun et al. 2014), and target capture shows promise in gathering thousands of loci from historical museum specimens when fresh tissues are unavailable (Knapp and Hofreiter 2010; Sun et al. 2014; McCormack et al. 2015).

The difficulties inherent in estimating galliform phylogeny (reviewed in Wang et al. 2013, Kimball and Braun 2014) also brand it a model system for understanding the consequences of methodological choices in exceptionally challenging phylogenetic inference scenarios. Here, we used UCEs to reconstruct a phylogenetic hypothesis for Galliformes, with a focus on resolving problematic nodes highlighted by previous studies. To have confidence in any resulting galliform phylogeny, it is also important to understand how both type I and type II missing data affect performance of phylogenetic methods—gene tree reconciliation frameworks in particular. We utilized model-based concatenated ML (RAxML; Stamatakis et al. 2014), two quartet methods expected to be consistent in the anomaly zone (supermatrix rooted triples, SMRT-ML; DeGiorgio and Degnan 2010; singular value decomposition scores for species quartets, SVDquartets; Chifman and Kubatko 2014), and two gene tree reconciliation approaches (Accurate Species Tree Algorithm, ASTRAL; Mirarab, Reaz, et al. 2014; Accurate Species Trees from Internode Distances, ASTRID; Vachaspati and Warnow 2015) to estimate the phylogeny of Galliformes. We apply all methodologies to five thresholds of matrix completeness (no type I missing data, 5% type I missing data, 25% type I missing data, 50% type I missing data, and total evidence), to understand how missing data influence concatenated and multispecies coalescent phylogenomic inference, particularly with respect to samples with reduced sequence yield and relatively large proportions of missing data. If results from differing methodologies and thresholds of missing data are qualitatively similar, we would conclude that methodologies are robust and missing data are of little practical consequence. However, if certain analytical techniques or missing data thresholds produce alternate strongly supported results with respect to taxa with large amounts of type I or type II missing data, they are likely biased, demonstrating that missing data are of concern.

Results

Sequence Capture Yields Data for Thousands of UCE Loci
We obtained data for 4,817 UCE loci, of which 4,638 contained at least one parsimony informative site (table 1). The number of UCE loci obtained varied from 1,035 to 4,328 per taxon, and base pairs recovered varied from 345,856 bp (16% of total aligned nucleotides) to 1,625,610 bp (74% of total aligned nucleotides) per taxon (supplementary table S1, Supplementary Material online). The large amount of data available allowed construction of data matrices up to 2.2 Mbp containing more than 180,000 informative sites. As matrix size increased in length, we observed the expected increase in number of variable and informative sites. However, the percentage of variable and informative sites decreased with increased matrix size, demonstrating a pattern of diminishing returns associated with increased missing data in larger alignments. The number of partitions identified by PartitionFinder varied from 7 to 45 depending on alignment, with larger alignments justifying greater numbers of partitions (table 1).

UCE average contig lengths recovered from sequence captures varied from 226 to 386 bp (supplementary table S1, Supplementary Material online). UCE contigs derived from historical DNA (toepads from museum specimens) were significantly shorter (unpaired t-test, P < 0.0001) in contig length (N50 311 bp) than contigs derived from fresh tissue samples (N50 372 bp). UCE enrichments using toepad source material produced similar numbers (unpaired t-test, P = 0.4492) of UCE contigs on average per enrichment (3,301 loci) to fresh material (3,654 loci).

Concatenated Maximum-Likelihood Phylogenetic Inference
ML phylogenetic inference using concatenated data sets produced strongly supported and congruent phylogenetic hypotheses for Galliformes (fig. 2). All well-supported nodes were congruent across all five alignments of varying
Fig. 2. Phylogeny of 90 galliform taxa inferred with ML analysis of 2,208,355 bp from 4,817 concatenated UCE loci. Inset shows multidimensional-scaled visualization of tree space, with each point representing a consensus tree produced with different inferential procedures and different thresholds of missing data (70 iterations, total). Concatenated ML, SMRT-ML, and SVD quartets trees (with the exception of those inferred from the small 0% missing data set) and ASTRAL/ASTRID trees with the most resolved gene trees converge on the same area of multidimensional tree space (lower right); RF SYRMATICUS Material online) lacked the power to resolve nodes among
1114
missing data; supplementary table S1, Supplementary
of missing data included. The 0% missing matrix (no type I
bootstrap support in all ML analyses regardless of the amount
Fig. 2.

Hosner et al. 

returns with the 25% missing matrix was consistent with the
that bootstrap support largely reached a point of diminishing
largely similar to the 25% missing matrices. The observation
node support increased with 1) alignment
larger sequence alignments that included type I missing
data. In general, node support increased with 1) alignment
settings); it received 32–39% bootstrap support in other ma-
5% missing matrix (52–57% bootstrap support, depending on
limited increase in informative characters for the 50% missing
and total evidence data matrices (table 1). The lone exception
to this pattern was support for placement of Pternistis ahan-
tensis (node Q; figs. 2 and 3), which received weak support in
all analyses. We recovered highest support for node Q in the
5% missing matrix (52–57% bootstrap support, depending on
settings); it received 32–39% bootstrap support in other ma-
trees. Only 5 nodes failed to receive 100% bootstrap support
in partitioned analysis of the 25% missing, 50% missing, and
total evidence matrices (table 1). The lone exception
node support increased with 1) alignment
that bootstrap support largely reached a point of diminishing
returns with the 25% missing matrix was consistent with the
limited increase in informative characters for the 50% missing
and total evidence data matrices (table 1). The lone exception
to this pattern was support for placement of Pternistis ahan-
tensis (node Q; figs. 2 and 3), which received weak support in
all analyses. We recovered highest support for node Q in the
5% missing matrix (52–57% bootstrap support, depending on
settings); it received 32–39% bootstrap support in other ma-
trees. Only 5 nodes failed to receive 100% bootstrap support
in partitioned analysis of the 25% missing, 50% missing, and
total evidence matrices for the 91 taxon, rooted, galliform ML
tree (fig. 2).

In our RAxML analyses, neither the bootstrapping algo-
rithm nor the approach used to accommodate among-sites
rate heterogeneity (i.e., fast bootstrapping with GTRCAT vs.
thorough bootstrapping with GTR+I) had a major effect on
support values across the phylogeny (figs. 2 and 3). Likewise,
the support values were similar regardless of whether the analyses were partitioned or unpartitioned. For the small number of nodes with bootstrap support <100%, the bootstrap values we obtained with different settings were within a few percentage points of one another and there were no settings that consistently increased or decreased support across the phylogeny (fig. 3 and supplementary fig. S1, Supplementary Material online).

Quartet-based Phylogenetic Inference Consistent in the Anomaly Zone

SMRT-ML (DeGiorgio and Degnan 2010) and SVDDQuartets (Chifman and Kubatko 2014) produced phylogenies similar to the clade of Old World quail and relatives, except that they failed to recover sufficient bootstrap support to resolve relationships of *Alectoris* and *Anmoperdix* within the clade of Old World quail and relatives (nodes N, O, P; fig. 3).

Two taxa, *Galloperdix spadicea* and *P. perdix*, having large proportions of type I and type II missing data (supplementary table S1, Supplementary Materials online) were placed in unexpected positions with low bootstrap support. These placements conflicted with all ML trees, all SVDDQuartets trees, and all other SMRT-ML trees inferred from larger matrices, which placed *Galloperdix* sister to *Polyplectron* (node H) and recovered *Perdix* monophyly (node X; fig. 2 inset and supplementary figs. S2–S4, Supplementary Material online). Otherwise, SMRT-ML and SVDDQuartets were similar to concatenated ML [Robinson–Foulds (RF) distances generally <5; supplementary fig. 3 and supplementary figs. S2–S4, Supplementary Material online].

Fig. 3. Support values of selected nodes across analyses executed on concatenated matrices. Columns represent all nodes in the phylogeny with <100% bootstrap support for the 0% missing matrix, or where some gene tree reconciliation analyses conflicted with concatenation. Rows represent RAxML across the phylogeny (fig. 3 and supplementary fig. S1, Supplementary Material online).
In general, bootstrap support for SMRT-ML and SVDquartet phylogenies increased with matrix size, with the exception of nodes pertaining to Guttera pucheranii (a low-yield sample with a large proportion of type I and type II missing data; table S1, Supplementary Materials online), which generally decreased in support as matrix size increased (fig. 3 and supplementary fig. S5–S4, Supplementary Material online). This was especially true with SMRT-ML using Quartet MaxCut QMC (QMC; supplementary fig. S3, Supplementary Materials online; Snir and Rao 2012), where Guttera was pulled outside of Numididae with moderate support in the 25% missing, 50% missing, and total evidence matrices. Otherwise, reconstructions of SMRT-ML quartets were qualitatively similar using QMC and Matrix Representation Parsimony (MRP). The 0% missing trees inferred with SMRT-ML and SVDquartets had considerably lower bootstrap support than those inferred with concatenated ML. Otherwise, bootstrap support values inferred with SVDquartets were slightly greater than those inferred with SMRT-ML, and slightly less than those inferred with concatenated ML. Like concatenated ML analyses, SMRT-ML and SVDquartets support values reached diminishing returns at > 25% missing data.

Gene Tree Inference and Gene Tree Reconciliation

Inference of 4,817 loci recovered in 4 or more galliform taxa, 4,613 contained at least one informative site and were used for downstream gene tree reconciliation in ASTRAL and ASTRID. For nine taxa (Crax globulosa, Gu. pucheranii, Melanoperdix niger, Tetraophasis obscurus, Tragopan satyra, Symmatics reevesii, Crossoptilon auritum, Lophura diardi, and Lophura leucolophus), ASTRAL phylogenies inferred using all gene trees (for all data completeness thresholds) consistently conflicted with those inferred using concatenated ML, SMRT-ML, and SVDquartets (RF distances 25–54), often with strong support (fig. 2–5). In each instance, trees inferred using ASTRAL placed these nine taxa toward the root when compared with their placement in concatenated ML and SMRT-ML topologies, creating a more pectinate tree shape (fig. 5 and supplementary fig. S5, Supplementary Material online). A review of N50s recovered for these nine taxa (supplementary table S1, Supplementary Material online) indicated that each had large amounts of type II missing data “within” loci (367, 334, 268, 303, 357, 348, 321, 350, and 367 bp N50, respectively) when compared with the entire data set (median 374 bp). Inferences using ASTRID were similar (fig. 5 and supplementary fig. S6, Supplementary Material online), except that 13 taxa with low N50s were placed toward the root with respect to those inferred with concatenated ML, SMRT-ML, and SVDquartets. These included the nine taxa affected in ASTRAL plus four additional taxa (Arborophila rufogularis, Centrocercus urophasianus, Oreoptyx pictus, Pternistis leucoscepus; 266, 331, 348, and 342 bp N50 respectively) that were placed toward the root with respect to those inferred with concatenated ML, SMRT-ML, and SVDquartets.

In contrast, when we analyzed only more informative gene trees (i.e., those inferred from either the 25% or 5% most informative UCE loci) using ASTRAL, resulting topologies, including relationships for the nine problematic taxa, were similar to concatenated ML, SMRT-ML, and SVDquartets analyses with moderate to strong support (figs. 2–4; RF distances generally <10). Indeed, topologies inferred with ASTRAL using only the 5% most variable loci differed from the concatenated tree at only one node with strong bootstrap support: Placement of the peafowl clade (sister to the most recent common ancestor of Polyplectron and Gallus domesticus in concatenated ML, SMRT-ML, and SVDquartets; sister to the most recent common ancestor of Gallus and Coturnix in ASTRAL). The 0% missing data set was an exception to this pattern, and ASTRAL trees inferred from it were still largely unresolved. Unlike ASTRID, ASTRAL inferred unlikely relationships for several taxa with large proportions of type II missing data (e.g., Gu. pucheranii, Or. pictus) even when only the most informative gene trees were used as input. Across all corresponding analyses, bootstrap support values inferred using ASTRID were considerably lower than those inferred with ASTRAL.

Discussion

Effects of Missing Data across Methods

Phylogenetic analysis of concatenated data reinforce the idea that increasing loci and nucleotides at the cost of increased missing data improves support at nodes that are difficult to reconstruct (Wagner et al. 2013; Huang and Knowles 2014; Streicher et al. 2016). However, few nodes in our galliform phylogeny showed increasing support when matrix size increased beyond the 25% missing threshold (an exception was node Z; figs. 2–4). Streicher et al. (2016) demonstrated a similar pattern in iguanian lizards, but in those analyses they observed diminishing returns at a 50% missing threshold. We expect that the point of diminishing returns differs between empirical data sets, and that data exploration by individual researchers is needed to determine appropriate thresholds. Diminishing returns may stem from the fact that as loci are added with increasing proportions of missing taxa/cells, they are less likely to include sequence data relevant to unresolved nodes in a phylogeny. Including taxa with relatively poor locus recovery (high type I missing data, taxa where we recovered approximately 25% of total loci) is also justified, and our results suggest that taxon exclusion to increase data completeness is likely to be unnecessarily cautious. In fact, there is evidence that increased taxon sampling can aid in resolving problems associated with long branch attraction even at the cost of increasing missing data (Wiens and Tiu 2012).

Similar to standard concatenated ML inference, support inferred with quartet methods (SMRT-ML and SVDquartets) improved with increasing numbers of loci. However, when compared with traditional concatenated ML, longer alignments were needed to produce strongly supported trees. For example, in the 0% missing concatenated ML tree, only two nodes had >50% bootstrap support (BS). In comparison, the 0% missing SVDquartets tree contained 5 such nodes and...
the 0% missing SMRT-ML tree contained 19 such nodes.

SMRT-ML and to a lesser extent SVDquartets appear to have reduced power to infer evolutionary relationships when compared with standard ML inference, as expected based on prior studies (DeGiorgio and Degnan 2010; Sun et al. 2014). Given sufficient data, however, these methods produced results similar to standard concatenation (figs. 2, inset, and 3).

---

![Fig. 4. Support values of selected nodes across gene tree reconciliation analyses. Columns represent all nodes in the phylogeny with <100% bootstrap support for the 0% missing matrix, or where some gene tree reconciliation analyses conflicted with concatenation. Rows represent ASTRID and ASTRAL gene tree reconciliation (with trees from all informative loci, 50% most informative loci, 25% most informative loci, and 5% most informative loci) analyses of the 0% missing (loci contain 100% of taxa), 5% missing (loci contain >95% of taxa), 25% missing (loci contain >75% of taxa), 50% missing (loci contain >50% of taxa), and the total evidence (all loci with >4 taxa) matrices.](image-url)
In contrast to results from concatenated ML, SMRT-ML, and SVDquartets analyses, gene tree reconciliation using ASTRAL inferred wildly different relationships among nine taxa (10% of included species). Gene tree reconciliation with ASTRID performed even more variably, with an additional four taxa placed in unexpected positions (14% of included species). These putatively spurious results do not appear to be related to increasing numbers of loci or type I missing data in alignments. Results for 5% missing, 25% missing, 50% missing, and total evidence matrices were all similar (the 0% missing had few resolved nodes, suggesting low power). Rather, the odd behavior using gene tree reconciliation approaches appeared to be related to the information content of the specific UCE loci we included (see below)—ASTRAL and ASTRID results improved markedly when we culled uninformative loci from the data set.

Uninformative Gene Trees, Biased Gene Trees, and Gene Tree Reconciliation

In contrast to analyses including all loci, using only the 25% most informative loci (ASTRAL; fig. 4) or the 5% most informative loci (ASTRAL and ASTRID; fig. 4) produced results largely congruent with those of ML, SMRT-ML, and SVDquartets inference. We observed this behavior in all alignments except the 100 complete matrix, which was still largely unresolved. Thus, inclusion of low information content loci in gene tree reconciliation approaches appears to hinder phylogenetic inference while also increasing computation time. Our gene tree reconciliation results clearly demonstrate that input gene tree resolution has a dramatic effect on inference.

Our results also suggest that the apparent poor performance of gene tree reconciliation is due to partial sequence data for certain taxa/loci (type II missing data) rather than
entire missing loci (type I missing data, which appeared to have little effect). Variation in enrichment efficiency, due to input sample quality or stochastic factors, along with stochastic variation in sequencing coverage, produces shorter contigs (low N50) for some taxa. Taxa with shorter contig lengths have fewer informative sites, such that estimates of gene trees with respect to those taxa will be poor (Simmons 2014) when compared with more data-rich samples, on average, across all loci. Thus, samples having shorter contigs are more likely to act as “unstable” or “rogue” taxa (Thomson and Shaffer 2010; Aberer et al. 2013; Goloboff and Szumik 2015) during individual gene tree inference. In the coalescent gene tree reconciliation analyses, erroneous placements of low N50 taxa in individual gene trees are attributed to ILS, rather than a systematic error introduced by missing sequence data. The net result is that low N50 taxa are systematically biased toward the root of the tree relative to their putatively true position (figs. 1 and 5). Examination of average contig lengths of these unstable taxa suggests that even relatively little missing data can cause this problem. For example, *Cra. globulosa* was unstable although its N50 of 367 bp was only 7 bp shorter than that of the mean N50 for the entire data set.

These results do not reflect theoretical problems of gene tree reconciliation analyses. Rather, they reflect a consequence of data collection procedures. Interestingly, not all taxa having low N50s suffered from this bias, and differences were apparent between reconciliation methods. For example, using ASTRAL, the taxon with the shortest N50 (*A. rufogularis* 266 bp), was always found sister to *Arborophila torquata* with strong support. With ASTRID, *A. rufogularis* was placed toward the root when all gene trees were analyzed, but ASTRID recovered *Arborophila* monophyly when only more informative gene trees were input. Clearly, N50 and reconciliation method are not the only factors affecting putatively erroneous placement of taxa: The relative branch lengths are also important. The branch joining *Arborophila* taxa to other Galliformes is relatively long, with ample time for substitutions to accrue; thus gene trees were reconstructed fairly reliably despite a low N50 in *A. rufogularis*. In general, both number of misplaced taxa and overall support values suggest that ASTRID is more sensitive to type II missing data than ASTRAL.

Limiting the set of input trees for gene tree reconciliation analyses to those based on more information-rich loci, from which gene trees can be estimated more reliably, provides a potential solution to the problem of biased gene trees without requiring the removal of all low N50 taxa from analyses. In our analyses, eliminating relatively uninformative gene trees greatly improved the apparent performance of gene tree reconciliation. In light of this observation, revisiting previous studies that employed gene tree reconciliation on low-variation markers (McCormack et al. 2013; Streicher et al. 2016) may be warranted to determine if selection of more informative gene trees can improve support. Another suitable approach would be to limit missing data within loci by trimming contigs to the shortest sequences. However, this approach would work poorly with our data set: Most UCE variation occurs in flanking regions at the 5′ and 3′ ends of sequences, and trimming to the shortest contigs leaves only the largely invariant core UCE region for analysis. Even in the most variable UCE loci that contain hundreds of informative sites across Galliformes, such extreme truncation would leave only a handful of informative sites, producing only poorly resolved gene trees for downstream gene tree reconciliation.

Our results indicating that gene tree reconciliation analyses perform better when uninformative gene trees are excluded raises a key question: How should researchers decide which gene trees to include? As a first step, our study chose arbitrary thresholds and compared performances. Although results from these arbitrary thresholds were qualitatively similar, there were slight differences in inferred topologies and estimated support. For example, the 25% most informative loci and the 5% most informative loci differed in their placement of *Pavo* and its relatives, one of the most challenging nodes to resolve in galliforms (Sun et al. 2014). Thus, user decisions regarding locus inclusion can influence results (Betancur-R et al. 2014). Nonetheless, other analyses have suggested that UCE flanking regions have excellent phylogenetic informativeness (Gilbert et al. 2015). Although the best metric for phylogenetic informativeness and the appropriate thresholds for inclusion loci in gene tree reconciliation analyses remain unclear, we found that for UCE loci, a commonly used metric (Townsend 2007; López-Giráldez and Townsend 2011) was strongly correlated with the number of parsimony informative sites (supplementary fig. S7, Supplementary Material online). This prompted us to use the simpler metric. Ideally, a multilocus approach using methods developed to identify and prune rogue taxa in supermatrices (Thomson and Shaffer 2010; Aberer et al. 2013; Goloboff and Szumik 2015) could objectively identify loci and/or taxa to be pruned from data sets.

An alternative and perhaps more robust solution would be to reduce dependence on gene tree reconciliation for phylogenomic inference when there are concerns about gene tree reliability. Methods like SMRT-ML and SVDquartets are executed on concatenated sequences, avoiding problems with type II missing data, biased or poor estimation of gene trees, or user choice of input gene trees. These methods also scale well to large data sets, and they are likely to scale to very large data matrices if it is possible to sample subsets of quartets without sacrificing accuracy. We recommend further testing of these methods to determine their robustness to the many ways that empirical data may violate their assumptions.

Given the numerous rapid radiations at various depths within Galliformes (Kimball and Braun 2014), one might have expected different results inferred from concatenated versus coalescent approaches. Instead, we found that results from each framework were largely congruent, although in the case of gene tree reconciliation methods this required limiting the input to the more informative gene trees. The observed congruence of estimate of the species tree obtained using concatenated and coalescent methods suggests that the primary cause of gene tree discordance in our Galliformes UCE data set was error in gene tree estimation due to character limitation rather than ILS per se. No doubt some discordance among true gene trees reflects ILS, but
congruence between approaches suggests that ILS is relatively low in the galliform tree (Mirarab, Bayzid, and Warnow 2014), and that our ML estimate of galliform phylogeny was not affected by the anomaly zone. We also note that there may be additional sources of discordance among true gene trees for the galliforms due to other processes, such as lateral gene transfer (hybridization is considered to be common in Galliformes) (Johnsgard 1970, 1988; Dong et al. 2013). However, spurious results obtained from including all gene trees in ASTRAL and ASTRID analyses reinforced the idea that error in gene tree estimation may often be the most important source of error in phylogenetic inference (Patel et al. 2013; Gatesy and Springer 2014). Thus, error in gene tree estimation should be considered when weighing selection of concatenated and coalescent approaches for low-variation markers like UCEs.

There are two clades in the galliform tree where concatenated ML differed from coalescent analyses with strong support. First, concatenated ML differed from gene tree reconciliation (when only the 5% most informative gene trees were considered) with respect to placement of the pea-fowl clade (node G; figs. 2–5). If this difference is attributable to a bias in phylogenetic reconstruction by standard ML analyses of concatenated data, we would predict SMRT-ML and SVDquartets to agree with gene tree reconciliation. Yet, our SMRT-ML and SVDquartets results agree with standard concatenated ML. Curiously, ASTRAL results (when the 25% and 50% most informative gene trees were considered) also agree with concatenated ML, SMRT-ML, and SVDquartets, suggesting that user choice of input trees has a strong effect on node G. Second, standard concatenated ML recovered strong support for placement of *Atmoperdix* and *Alectoris* (nodes O and P), but coalescent analyses, including SMRT-ML and SVDquartets, found little support for any relationships pertaining to these nodes. There are two possible interpretations for this result. Concatenated ML is thought to have greater power than coalescent methods to identify relationships when ILS is low. If this is the case, then the inferred concatenated ML tree may be reliable. However, if there is substantial ILS with respect to nodes O and P, concatenated ML could be positively misleading. The first of these two hypotheses seems more likely given the topology and branch lengths, because the subtree defined by nodes O and P is maximally asymmetric and anomaly zone problems reflect the higher probability of symmetric gene trees given asymmetric species trees (Degnan and Rosenberg 2006; Rosenberg 2013), and the recovery of asymmetric trees in concatenated analyses has been used to argue against the existence of a bias due to the anomaly zone (Harshman et al. 2008; Smith et al. 2013). Further exploration of these nodes using more rapidly evolving markers and/or increased taxon sampling would be ideal to resolve the relationship nodes defined O and P with greater confidence.

### Data Quality from Historical Samples

One objective of this study was to explore the potential of using historical DNA extracted from museum skin toepads (Mundy et al. 1997) as source material for UCE enrichment and sequencing (McCormack et al. 2015). Use of historical material (often referred to as “ancient DNA”) is common in systematics, and its use has the clear benefit of allowing researchers to increase taxon sampling when fresh tissues are unavailable (McCallum et al. 2013; Heupink et al. 2014; Mitchell et al. 2014). However, the fragmented nature of historical DNA can lead to polymerase chain reaction (PCR) contamination, PCR errors (Sefer et al. 2007), and preferred amplification of pseudogenes (Greenwood et al. 1999), which compromises phylogenetic inference and may often go unnoticed for years (Zuccon and Ericson 2010; Moyle et al. 2013, 2015). The challenge of producing extensive character data sets with historical DNA has limited researchers to (often partial) sampling of one or a few loci, frequently just the mitochondrial, which may also limit or mislead phylogenetic inference (Maddison 1997).

Consistent with previous studies that have advocated the use of historical DNA in massively parallel sequencing (Knapp and Hofreiter 2010; Mason et al. 2011; McCormack et al. 2015), our results demonstrate that target capture of historical DNA can be successful with little alteration of protocols, and that inclusion of historical samples gives a clearer picture of galliform phylogeny (Sun et al. 2014). Yet, uncritical use of these samples, which have a lower NS5 than fresh tissues (supplementary table S1, Supplementary Material online), could give rise to spurious results when using gene tree reconciliation frameworks (figs. 3 and 4). Thus, historical samples are an important source for target capture, but they are better utilized minimally rather than as a routine substitute for high-quality source material.

### Toward a Robust Phylogeny of Galliformes

We recovered congruent and strongly supported topologies using numerous phylogenomic inference strategies, each subject to different limitations and biases. Thus, UCE results finally clarify many historical relationships that have remained problematic. For example, relationships within the junglefowl (*Gallus*), which includes the domestic chicken, have been weakly supported and variable in topology (Kimball and Braun 2008, 2014; Wang et al. 2013), although results using whole mitochondrial genomes strongly support placing *Ga. gallus* sister to *Gallus varius* (fig. 6; Meiklejohn et al. 2014). In contrast to those results from mitochondrial data, our concatenated and coalescent results unequivocally support placing *G. gallus* sister to a *Gallus sommerati + Gallus lafayettii* clade (figs. 2–6). Given the contrast to strong supported in mitogenomic studies, this could either indicate cyto-nuclear discordance in this clade of closely related birds, or a rooting problem in the mitogenomic tree.

Relationships among genera of Old World quails and relatives have also been largely unresolved in previous studies (reviewed by Crowe et al. 2006; Wang et al. 2013), although only a few of these genera have been included in more than one or two studies. Concatenated ML analyses inferred a well-supported multilocus hypothesis of this group, whereas...
Avoiding Missing Data Biases

Fig. 6. Recent phylogenetic hypotheses for the junglefowl genus Gallus. Node labels refer to ML bootstrap percentages, with * = 100%. Previous studies recovered differing topologies, often without strong support. All UCE analysis, including ASTAL runs using all gene trees (which performed poorly with respect to many taxa), recovered this topology with strong support.

Materials and Methods

Sequence Capture, Assembly, and Alignment of UCEs

We selected 86 galliform taxa and one outgroup (Anseriformes: Oxyura jamaicensis) for UCE enrichment, including all major clades identified by previous studies. Taxa were selected to 1) cover the breadth of galliform diversity; 2) focus on problematic nodes, particularly surrounding model taxa in the Phasianidae; and 3) identify relationships of two enigmatic, monotypic genera not included in previous molecular phylogenetic studies due to a lack of fresh tissue resources (Me. niger, L. lerwa). We extracted genomic DNA from fresh tissues (most samples) or toepads of museum specimens; sample information can be found in supplementary table S1, Supplementary Material online.

Sequence capture libraries were prepared using an approach modified from Faircloth et al. (2012). We prepared Nextera sequencing libraries using the manufacturer’s protocol (Illumina, Inc., San Diego, CA), but using primers with custom index tags (Faircloth and Glenn 2012). We pooled 8 samples together, and enriched each library pool for 5,060 UCE loci targeted by 5,472 probes (Mycroarray, Ann Arbor, MI; http://www.mycroarray.com/mybaits/mybaits-UCEs.html). Enriched libraries were amplified with 18 PCR cycles, quantified using qPCR (quantitative PCR; Kapa Biosystems), and sequenced in a single Illumina HiSeq 2000 Lane (75 nt paired-end reads; UC Irvine Genomics High-Throughput Facility). DNA extracts from toepads underwent the same library preparation procedures as fresh tissue.

We assembled quality-controlled reads into contigs de novo using Trinity (Grabherr et al. 2011), and added UCE loci of Ga. gallus, M. gallopavo, Cot. japonica, and Col. virginianus extracted from published genome assemblies (International Chicken Genome Sequencing Consortium 2004; Dallool et al. 2010; Kawahara-Miki et al. 2013; Halley et al. 2014). Sequences for UCE loci obtained from four or more taxa ($n = 4,817$) were aligned using MAFFT 7 (Katoh et al. 2002; Katoh and Standley 2013). We trimmed ends of alignments when 35% of cells were missing. Resulting final taxon sampling included 90 taxa and 298 currently recognized galliform species (Gill and Donsker 2015).

Alignment Filtering Matrix Construction

We constructed sets of UCE alignments with five thresholds of taxonomic completeness (table 1) with the PHYLUCE 1.5 pipeline (Faircloth et al. 2012, Faircloth 2015). Alignments
Concatenated Maximum Likelihood Phylogenetic Inference

For each concatenated alignment (0% missing, 5% missing, 25% missing, 50% missing, total evidence), we implemented PartitionFinder 1.1 (Lanfear et al. 2012) using the “hcluster” algorithm (using default weighting rate = 1, base = 0, model = 0, alpha = 0) and the GTR+Γ model of sequence evolution with each UCE locus as a data subset. We conducted ML phylogenetic inference in RAxML 8.1.1 (Stamatakis et al. 2008; Stamatakis 2014). For each of the 5 concatenated alignments, we computed an ML tree and 500 bootstrap replicates with 1) rapid nonparametric bootstrapping of the unpartitioned data set using the GTRCAT approximation, 2) thorough nonparametric bootstrapping of the unpartitioned data set using GTR+Γ, and 3) thorough nonparametric bootstrapping of the partitioned data set using GTR+Γ. All phylogenetic analyses were computed at the University of Florida High-Performance Computing Center.

Quartet-based Phylogenetic Inference Consistent in the Anomaly Zone

We estimated phylogenies for each of the five concatenated matrices (0% missing, 5% missing, 25% missing, 50% missing, total evidence) using two quartet-based methods that take advantage of the fact that there are no anomalous trees for a rooted quartet. These approaches are executed on concatenated matrices, and therefore may be less sensitive to missing data than gene tree reconciliation. First, we implemented a modified supermatrix rooted triples approach (SMRT-ML; DeGiorgio and Degnan 2010). All supermatrix-rooted triples (effectively all possible quartets that contain the outgroup Ox. jamaicensis) were inferred with RAxML 8.1.1 under the GTR+Γ model from unpartitioned concatenated matrices using a custom Perl script. Phylogenies were constructed from SMRT-ML quartets were using two methods: 1) We built MRP matrices from quartets with Clann (Creevey and McInerney 2005), which then were executed in PAUP* 4.0b10 (Swofford 2003), and 2) we reconstructed phylogenies from quartets directly with QMC 3.0 (Snir and Rao 2010, 2012). We produced 100 nonparametric SMRT-ML bootstraps for each of the 5 data matrices. Second, we implemented SVDquartets (Chifman and Kubatko 2014), a method that infers quartets based on summaries of SNPs in a concatenated sequence matrix. We invoked SVDquartets in PAUP* 4.1a146, sampling all quartets, and we constructed phylogenies using QMC 3.0 (Avni et al. 2015). We inferred SVDquartets for 100 nonparametric bootstraps for each of the 5 data matrices.

Multispecies Coalescent Inference Using Gene Tree Reconciliation

We inferred phylogenies for each set of five sets of alignments (0% missing, 5% missing, 25% missing, 50% missing, total evidence) under the multispecies coalescent using two gene tree reconciliation algorithms, ASTRAL 4.4.4 (Mirarab, Reaz, et al. 2014) and ASTRID (Vachaspati and Warnow 2015). ASTRAL (Mirarab, Reaz, et al. 2014) takes sets of gene trees (or bootstraps of unrooted gene trees) and computes the phylogeny which agrees with the largest number of quartet trees induced by the gene tree set. ASTRID takes sets of gene trees and computes the phylogeny from internode distance.

ASTRAL and ASTRID use unrooted gene trees and allow for missing taxa, therefore allowing inclusion of loci for which the outgroup taxon is missing. Note that a rewritten version of ASTRAL (ASTRAL II; Mirarab and Warnow 2015) handles missing data differently, and may outperform the standard version. However, like other coalescent-based gene tree reconciliation methods, ASTRAL and ASTRID assume that the input gene trees are estimated without error. To understand the effects of including uninformative loci that may bias gene tree reconciliation, we implemented a series of ASTRAL and ASTRID runs using 1) all gene trees, 2) gene trees from the 50% most parsimony informative loci, 3) gene trees from the 25% most parsimony informative loci, and 4) gene trees from the 5% most parsimony informative loci. To ensure that the number of parsimony informative sites is an accurate indicator of informativeness, we also calculated phylogenetic informativeness using PhyDesign (López-Giráldez and Townsend 2011) for a subset of loci with little missing data (supplementary fig. S7, Supplementary Material online). Because the method of Townsend (2007) was highly correlated with parsimony informativeness (supplementary fig. S7, Supplementary Material online), we selected gene trees for inclusion using only number of parsimony informative sites. To estimate gene trees, we computed 100 thorough bootstraps under GTR+Γ for all 4,638 loci containing 4 or more taxa and at least one informative site. We computed all four thresholds of locus variability (all gene trees, gene trees from the 50% most parsimony informative loci, gene trees from the 25% most parsimony informative loci, and gene trees from the 5% most parsimony informative loci) with each of our five sets of alignments of varying completeness (0% missing, 5% missing, 25% missing, 50% missing, total evidence).

Topological Comparisons

To compare phylogenetic results from different analyses with different thresholds of matrix completeness, we computed pairwise RF distances between majority rule consensus trees in PAUP* 4.0b10 (Swofford 2003), and visualized tree space using multidimensional scaling (Hillis et al. 2005) of the pairwise RF tree distance matrix computed in R 3.1 (R Core Team 2014).

Supplementary Material

Supplementary figures S1–S7 and tables S1–S3 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).
Acknowledgments

We would like to thank institutions, collectors, and collection managers who provided tissues for this study: The Field Museum of Natural History, the Florida Museum of Natural History, the Louisiana State Museum of Natural Science, and the University of Washington Burke Museum. Ricardo Betancur-R and two anonymous reviewers provided constructive comments to improve the manuscript. This work was supported by the National Science Foundation [grant numbers DEB-1118823 to (R.T.K. and E.L.B.) and DEB-1242260 (to B.C.F.)]. Alignments and treefiles are deposited at FigShare at https://figshare.com/projects/5596. Custom Perl scripts are available at GitHub at https://github.com/ebraun68/SMRT-ML DNA Sequence read data are archived on NCBI SRA (PRJNA303085).

References

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC. 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol Lett. 8:783–786.
Faircloth BC. 2015. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics.


McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. 2013. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. Genes 4:236.


Avoiding Missing Data Biases


