Article Type: Discoveries

Whole-genome phylogenomics of the tinamous (Aves: Tinamidae): comparing gene tree estimation error between BUSCOs and UCEs illuminates rapid divergence with introgression

Authors:
Lukas J. Musher[1], Therese A. Catanach[1], Thomas Valqui[2], Robb T. Brumfield[3], Alexandre Aleixo[4], Kevin P. Johnson[5], Jason D. Weckstein[1,6]

Author affiliations:
[1] The Academy of Natural Sciences of Drexel University, Department of Ornithology Philadelphia, PA, 19103, USA.
[2] Facultad de Ciencias Forestales, Universidad Nacional Agraria La Molina, Lima, Peru and CORBIDI-Centro de Ornitología y Biodiversidad, Lima, Peru.
[3] Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge, LA, 70803, USA.
[4] Instituto Tecnológico Vale - ITV, Belém, Brazil.
[5] Illinois Natural History Survey, Prairie Research Institute, University of Illinois Urbana-Champaign, Champaign, IL, USA.
[6] Department of Biodiversity, Earth, and Environmental Sciences, Drexel University, Philadelphia, PA, 19103, USA.

Abstract:

Incomplete lineage sorting (ILS) and introgression increase genealogical discordance across the genome, which complicates phylogenetic inference. In such cases, identifying orthologs that result in gene trees with low estimation error is crucial because phylogenomic methods rely on accurate gene histories. We sequenced whole genomes for the tinamous (Aves: Tinamidae) to dissect the sources of gene and species-tree discordance and reconstruct their interrelationships. We compared results based on four ortholog sets: (1) coding genes (BUSCOs), (2) ultraconserved elements (UCEs) with short flanking regions, (3) UCEs with intermediate flanks, and (4) UCEs with long flanks. We hypothesized that orthologs with more phylogenetically informative sites would result in more accurate species trees because the resulting gene trees contain lower error. Consistent with our hypothesis, we found that long UCEs had the most informative sites and lowest rates of error. However, despite having many informative sites, BUSCO gene trees contained high error compared to long UCEs. Unlike UCEs, BUSCO gene sequences showed a positive association between the proportion of parsimony informative sites and gene tree error. Thus, BUSCO and UCE datasets have different underlying properties of molecular evolution, and these differences should be considered when selecting loci for phylogenomic analysis. Still, species trees from different datasets were mostly congruent. Only one clade, with a history of ILS and introgression, exhibited substantial species-tree discordance across the different data sets. Overall, we present the most complete phylogeny for tinamous to date, identify a new species, and provide a case study for species-level phylogenomic analysis using whole-genomes.

Introduction

Although the growth of high-throughput sequencing approaches over the past decade has greatly improved our understanding of evolutionary relationships, reconstructing the tree of life remains an enduring challenge. Analyses that utilize alternative datasets, methodological frameworks, or substitution models to answer the same phylogenetic questions often yield conflicting results, which is surprising given that phylogenomic datasets often contain hundreds of thousands or even millions of phylogenetically informative characters (Dunn et al. 2008; Jarvis et al. 2014; Prum et al. 2015; Reddy et al. 2017; Simion et al. 2017; Franz et al. 2019; Schultz et al. 2023). In many cases, alternative phylogenomic topologies resulting from different methods or data are equally well-supported. Thus, two key questions for molecular biologists are, (1) why do phylogenomic datasets yield discordant topologies? and (2) how should one choose among conflicting well-supported topologies?

One way in which discordant phylogenies can result from the same or similar datasets is from the use of different methods for reconstructing species trees. One frequently employed method is to concatenate all alignments into a single supermatrix and treat the resulting phylogeny as a genome-wide average (henceforth, concatenation). Another common approach is to estimate a gene tree for each molecular marker independently and summarize their histories to estimate the species tree based on expectations from a multi-species coalescent (MSC) model (Zhang and Mirarab 2022). Unlike the concatenation approach, MSC methods are designed to account for expected variation in gene histories due to incomplete lineage sorting (ILS). ILS causes some gene trees to differ from the species tree due to retained ancestral variation during speciation, resulting in alleles coalescing prior to speciation events in ways that result in incongruence between gene and species trees (Maddison 1997).

Each of these two methods, concatenation and MSC, has theoretical advantages, but can result in erroneous species trees under certain conditions. For example, there is an expectation that concatenation may be the best method when ILS is weak but is likely to be statistically inconsistent when ILS is strong (Mendes and Hahn 2018; Bryant and Hahn 2020). This is because when ILS is strong, gene genealogies are expected to differ more from the species tree (i.e., there is increased heterogeneity) compared to situations when ILS is weak, thus reducing the probability that concatenation will find the true tree. ILS is more likely to occur when successive divergence times are short and effective population sizes are large, such as during rapid or adaptive radiations (Maddison 1997; Mclean et al. 2019; Lescroart et al. 2023; Tan et al. 2023). Thus, discordance among concatenation and MSC methods may be in part driven by the presence of rapid diversification that leads to ILS and high gene tree heterogeneity that biases concatenation results.

When applying MSC methods, identifying orthologous markers that lead to accurate, unbiased gene tree estimation is crucial to properly infer phylogeny (Meiklejohn et al. 2016; Springer and Gatesy 2016; Tea et al. 2021). MSC methods that utilize gene trees as input (gene tree summarization methods) may fail if gene tree heterogeneity is driven by estimation error rather than coalescence events. In other words, the MSC model is intended to incorporate gene tree heterogeneity due to ILS, but erroneous gene trees resulting from low quality, biased, or inconsistent sequence data will increase species tree estimation error (Xi et al. 2015). Thus, dissecting biological and artifactual sources of gene tree discordance to more accurately infer species trees is a major challenge in the age of phylogenomics, where more data does not automatically translate to improved inferences (Meiklejohn et al. 2016; Blom et al. 2017; Smith et al. 2023). Because MSC methods assume that all gene tree discordance is driven by ILS, eliminating erroneous gene trees becomes important (Roch and Warnow 2015; Springer and Gatesy 2016). This point has led to a variety of methods for filtering genomic datasets for phylogenetic efficacy (Doyle et al. 2015; Kuang et al. 2018; B.T. Smith et al. 2018; S.A. Smith et al. 2018; Zhao et al. 2023) but has generally lacked consensus on the types of data that contain low error to begin with.

Given the complexities associated with phylogenetic inference, identifying datasets of gene orthologs containing minimal bias is a principal goal of systematics research and has implications across diverse fields in molecular biology. One widely employed data type for phylogenomics includes coding sequences obtained via anchored hybrid enrichment, transcriptomics, or low coverage whole-genome sequencing (Lemmon et al. 2012; Prum et al. 2015; Allen et al. 2017; Waterhouse et al. 2017; Johnson 2019; Zhang et al. 2019; Burbrink et al. 2020; Boyd et al. 2022; Van Damme et al. 2022). Another data type includes ultraconserved elements (UCEs), which target conserved genomic regions to identify orthology, but also yield sequences of their more variable flanking regions for phylogenetic reconstruction across a range of timescales (Faircloth et al. 2012; McCormack et al. 2012; Faircloth et al. 2013; Musher and Cracraft 2018; Gueuning et al. 2020; Catanach et al. 2021; Ostrow et al. 2023). Both data types are widely applied in phylogenomics but can sometimes result in different phylogenetic topologies. For instance, coding and non-coding markers have resulted in conflicting topologies for the backbone of Neoaves (McCormack et al. 2013; Jarvis et al. 2014; Prum et al. 2015), and some authors have suggested that this discordance can be attributed to data biases associated with coding genes (e.g., elevated GC content and model misspecification) that also bias downstream species tree inferences (Reddy et al. 2017). Nevertheless, there have been few studies comparing the efficacy of these marker types to one another. Thus, the ability of coding versus primarily non-coding markers to resolve phylogenomic trees remains an open question.

One method for evaluating ortholog quality and gene tree error is through quantification of alignment information content, such as the number of parsimony informative sites (henceforth, PIS), or by comparing levels of gene tree heterogeneity among data types (Fong and Fujita 2011; Harris et al. 2014; Burbrink et al. 2020; Leite et al. 2021; Smith et al. 2023). For example, alignments with few PIS often result in erroneous or poorly resolved gene trees because of a lack of phylogenetic signal in the data (Xi et al. 2015; Meiklejohn et al. 2016). In such cases, we expect datasets with fewer PIS to result in increased gene tree heterogeneity that is driven by gene tree estimation error. For example, alignments of coding genes, or UCEs that include long flanking regions, should contain many PIS, and therefore should result in lower rates of gene tree error than UCEs with short flanking regions that contain relatively few variable sites. Some studies have found that loci with more PIS result in more “clock-like” gene trees, suggesting that more informative loci result in gene trees with branch lengths that represent real biological signal rather than data-driven artifacts (Musher et al. 2019). Other studies have shown a negative correlation between the number of PIS and Robinson-Foulds distance (a measure of phylogenetic dissimilarity; henceforth, RF distance) between gene and species trees (Burbrink et al. 2020). This relationship implies that more informative alignments result in reduced gene tree estimation error. Alternatively, one might also expect that phylogenetic markers that are too variable can result in erroneous gene trees, if phylogenetic signal becomes lost due to sequence saturation (Felsenstein 1978; Brinkmann et al. 2005). Thus, documenting the relationship between alignment information content and RF distance, as well as the variation in overall RF distances among different datasets, should inform future phylogenomic study design and illuminate data-type efficacy.

Whole-genome sequencing offers a way to test the robustness of a phylogenetic topology given multiple types of phylogenetic markers, including both coding and non-coding sequences, from the same underlying data. For example, whole genomes are often used to harvest large numbers of single-copy orthologous protein-coding markers called BUSCO (Benchmarking Universal Single Copy Orthologs) genes (Waterhouse et al. 2017; Alaei Kakhki et al. 2023). Given their conservation of function, BUSCO genes can theoretically tackle both relatively deep and shallow phylogenetic problems (Timilsena et al. 2022; Van Damme et al. 2022; Alaei Kakhki et al. 2023). Similarly, there are multiple pipelines that can easily be used to obtain UCEs and other conserved elements for phylogenomics from whole-genome datasets (Faircloth 2016; Edwards et al. 2017). Thus, sequencing whole genomes enables a robust test of the efficacy of different data types for phylogenomics.

In this study, we apply whole-genome sequencing to reconstruct a species-level phylogeny of the tinamous (Aves: Tinamidae). There has been limited work on tinamou phylogenomics to date. Most past phylogenetic studies utilized either morphological data with relatively few genetic markers (Bertelli et al. 2002; Bertelli and Porzecanski 2004; Valqui 2008; Bertelli 2017; Almeida et al. 2022), or phylogenomic data with large-scale molecular sampling but minimal taxonomic sampling (Cloutier et al. 2019). Thus, the species-level phylogenetic relationships of the tinamous remain uncertain. Here, we build on past studies by sampling thousands of orthologous markers across 45 out of 46 described species and a total of 65 named taxa (monotypic species plus subspecies). Specifically, we compare levels of PIS and gene tree estimation error in two types of orthologous markers: complete single-copy protein-coding genes (BUSCO's) and UCE's. In doing so, our objectives are to (1) examine the effect of alignment information content on gene tree estimation error, (2) use this information to identify high quality (low error) orthologs, (3) identify the drivers of phylogenetic discordance among data types and methods, and (4) use these inferences to robustly reconstruct the evolutionary history of the tinamous.

Results

Assembly metrics, completeness, and ortholog metrics

We successfully assembled 61 of 62 newly sequenced genomes and extracted BUSCO and UCE targets from these assemblies plus 10 publicly available tinamou and 2 publicly available outgroup whole genome assemblies (Table S1). Genome wide sequence coverage varied within and among samples but was generally high; the mean of average coverage across samples was 40.41x (standard deviation = 13.55x). Most genomes were also relatively complete, containing an average of 89.69% (standard deviation = 10.88%) of 8,338 complete single-copy BUSCO genes.

From these whole genomes, we extracted and aligned two types of orthologous markers, BUSCO genes and UCEs. Specifically, we compiled four different ortholog sets, which after filtering for complete occupancy (all 72 samples represented in each alignment) included (1) 2,507 BUSCOs (6,368,028 bp in concatenated alignment), (2) 2,887 UCEs with 100 bp flanking regions (969,279 bp in concatenated alignment), (3) 2,879 UCEs with 300 bp flanking regions (2,227,921 bp in the concatenated alignment), and (4) 2,803 UCEs with 1000 bp flanks (6,902,014 bp in the concatenated alignment). Henceforth, we refer to these datasets as BUSCO, UCE100Flank, UCE300Flank, and UCE1000Flank datasets, respectively.

Phylogenomics of Tinamous

Each of the four ortholog datasets was analyzed twice: first we reconstructed the phylogenetic history of tinamous using all loci concatenated into a single alignment (henceforth, concatenated phylogeny) and then using a gene tree summarization approach implemented in ASTRAL (Mirarab et al. 2014; Zhang et al. 2018) (henceforth, MSC phylogeny). The reconstructed concatenated and MSC phylogenies for tinamous were well-supported and broadly congruent across data types (Figures 1, 2, S1–S6). The phylogenies recovered from all four datasets were entirely congruent except for relationships within a single clade (henceforth, Clade A) of *Crypturellus* containing 13 named taxa across nine recognized species (Figure 3). One genome, downloaded from NCBI (GCA 013389825) was labeled as *C. undulatus*, but did not cluster with other members of that species, instead clustering with either *C. strigulosus* (MSC trees) or *C. erythropus* (concatenated trees). We could not confirm the species identity of this sample because the voucher was unavailable, and metadata indicated it was missing its head. Moreover, another downloaded genome, GCA 013398335, was listed as *Nothoprocta ornata* on NCBI but after including newly sequenced material, this *N. ornata* clustered with *N. pentlandii* samples from Peru. After examining the voucher specimen (CORBIDI 168605), the sample was indeed confirmed as *N. pentlandii oustaleti* and not *N*. *ornata* as indicated in NCBI.

Our results were broadly consistent with recognized taxonomic classifications. For example, we recovered monophyletic subfamilies Nothurinae and Tinaminae and most genera and species were also monophyletic. However, we recovered two genera as non-monophyletic: the monotypic genus *Taoniscus* was embedded within *Nothura* and *Rhynchotus* was embedded within *Nothoprocta*, rendering *Nothura* and *Nothoprocta* paraphyletic. We also recovered *Nothoprocta pentlandii* as polyphyletic; one specimen from Bolivia (voucher LSUMZ 123403) was sister to *N. perdicaria*, whereas the remaining samples from Peru formed a clade that was sister to all remaining *Nothoprocta* (consistent with mitochondrial results in Valqui 2008). All other species that were represented by multiple samples were recovered as monophyletic.

Assessment of ortholog information content and gene tree estimation error

We recovered significant differences in the number and proportion of PIS per alignment as well as levels of gene tree estimation error among the four datasets. First, smilograms revealed key differences in patterns of within-locus variability, wherein BUSCOs showed a bimodal distribution of variable sites (likely reflecting increased variability at third codon positions) but UCE datasets showed an expected pattern of gradually increasing variability moving away from conserved UCE cores (Figure 4). As expected, the UCE100Flank dataset had the fewest PIS, averaging just 45 PIS per alignment (standard deviation = 28.84), the UCE300Flank dataset averaged an intermediate number of PIS with a mean of 207 per alignment (standard deviation = 88.37), and the UCE1000Flank dataset contained the most PIS with a mean of 1,036 per alignment (standard deviation = 250.57). The BUSCO alignments also contained a high number of PIS, though with very high variance, indicating a range of informative and uninformative loci (mean = 923, standard deviation = 690.08 PIS per locus) (Table 1; Figure 5). Kruskall-Wallis tests confirmed that differences in both the number (X^2 = 578, *df* = 3, $P < 0.00001$) and proportion (X^2 = 6408.9, *df* = 3, $P < 0.00001$) of PIS significantly differed among datasets. Wilcoxon rank sum tests also indicated that differences between pairwise comparisons of these datasets were all significantly different ($P < 0.00001$ for all).

To measure gene tree estimation error, we examined gene tree heterogeneity across datasets using Robinson-Foulds (RF) distances between gene and species trees. Although some heterogeneity is expected due to the stochastic nature of the coalescent process, we assumed that increases in the mean and variance of RF distances between datasets were attributable to increased gene tree estimation error. The UCE100Flank dataset had the highest RF distances between gene and species trees, whereas the UCE1000Flank dataset had the lowest mean and variance in RF distances (Figure 5; Table 1). The BUSCOs and UCE300Flank datasets had intermediate RF distances, with similar means and variances (Table 1). Interestingly, RF distances from the BUSCO dataset were highly variable, with estimated gene trees from some loci showing relatively low (akin to the UCE1000Flank dataset) and some showing exceptionally high (akin to the UCE100Flank dataset) RF distances relative to both MSC and concatenated species trees. A Kruskall-Wallis test confirmed that RF-distances (X^2 =

2740.1, *df* = 3, *P* < 0.00001) differed significantly among datasets, and a Wilcoxon rank sum test also revealed that differences between pairwise comparisons of these datasets were significantly different (*P* < 0.00001) except that RF-distances did not significantly differ between the UCE300Flank and BUSCO datasets (*P* = 0.25).

To quantify the relationship between alignment information content and gene tree estimation error and test the hypothesis that increased information content led to reduced gene tree estimation error, we modeled RF distances as a function of the number of parsimony informative sites (PIS) per locus using generalized linear models. These models were overall consistent with our hypothesized pattern of a negative association between both the number and proportion of PIS and RF distance, with lower RF distances in gene trees built from more informative alignments (Figure 6; Table S2). However, there was notable variation in the slope and tightness of fit of these models among datasets. For example, the UCE300Flank dataset, with intermediate numbers of PIS overall, had a tight negative association between RF distance and both the number of PIS and percentage of PIS per locus (Table S2). The same negative associations were revealed for UCE100Flank and UCE1000Flank datasets, but these relationships were much noisier. These differences were better visualized when the three UCE datasets were combined, revealing a strong and tightly fitting negative logarithmic relationship between the number of PIS per alignment and RF distance. Although this relationship was similar for the BUSCO dataset using the number of PIS as the predictor variable, we found the opposite pattern when using the percentage of PIS. Thus, based on multiple assessments, BUSCOs and UCEs behave differently with regards to gene tree to species tree discordance and patterns of nucleotide site variation.

Tests of incomplete lineage sorting and introgression

Finally, because we found conflicting phylogenetic relationships between datasets for a clade of species in the genus *Crypturellus* (Figure 3), we assessed the extent to which this conflict could be explained by ILS due to rapid divergences or introgression between non-sister taxa. To do so, we looked at the relative frequencies of alternative quartets for five short internal branches within Clade A (Figure 7A). The MSC model assumes that all gene tree heterogeneity arises via ILS (i.e., no introgression, no gene tree error, no recombination, etc.), and it makes explicit predictions about the relative frequencies of alternative quartet topologies (unrooted four taxon statements) for all internal branches in a phylogeny. Specifically, the MSC model predicts one majority topology that is consistent with the species tree at a relative frequency >⅓ and two minority topologies of equal relative frequencies <⅓ (Pamilo and Nei 1988). As ILS increases, the relative frequencies of all three alternative quartets approach the ⅓ threshold. Deviations from these expectations (i.e., minority topologies are not equivalent in frequency) suggests that processes such as introgression may be influencing gene tree heterogeneity. All five branches that we examined showed evidence of either ILS or deviations from the MSC model. For branches 1 and 2 (two of the three shortest branches within Clade A) we recovered frequencies for three alternative quartets that closely approached the ⅓ threshold, implying ILS has promoted phylogenetic conflict at these branches. For branches 3–5, we found possible deviations from the MSC model, though this could be an artifact of sample size.

To test whether the deviations from the MSC model observed in Clade A could be attributed to introgression, we applied a phylogenomic network analysis, which revealed multiple nodes involved in reticulation that may explain gene tree discordance in *Crypturellus* Clade A (Figure 7B; electronic supplementary material). Specifically, we found two reticulate nodes. In the first, *C. cinnamomeus* was recovered as reticulate between *C. transfasciatus* (inheritance probability = 0.44) and *C. erythropus* (inheritance probability = 0.56). In the second, *C. undulatus* was reticulate between the basal node of the clade (inheritance probability = 0.76) and *C. erythropus* (inheritance probability = 0.24). Thus, some gene tree heterogeneity within the low-error UCE1000Flank dataset may be attributable to historical introgression between non-sister taxa.

Discussion

Here we used whole-genome sequencing to explore the biological and artifactual sources of phylogenomic conflict and reconstruct the species-level history of a relatively old Neotropical avian group of broad interest in evolutionary biology (Bertelli and Porzecanski 2004; Prum et al. 2015; Altimiras et al. 2017; Sackton et al. 2019; Li et al. 2023), the tinamous. Although we found that alignment information content and gene tree error varied considerably within and among datasets, our results based on different analytical approaches (concatenated and MSC methods) and datasets (UCEs and BUSCOs) were largely congruent. The topology of only a single clade (Clade A) varied among methods and datasets (Figures 1, 2, and 3), indicating that species tree estimation is, in general, robust to gene tree estimation error when biological sources of gene tree heterogeneity are limited. Nevertheless, the concatenated tree shows that Clade A contained multiple successive short internodal branch lengths (Figure 1), which suggests ILS may be a key driver of gene tree heterogeneity in this clade, and thus an important factor driving phylogenetic conflict. Indeed, quartet analysis revealed multiple internal branches with relative quartet frequencies approaching the ⅓ threshold predicted by high ILS (Figures 7A and S7). Cases such as this, where ILS is strong, necessitate the use of datasets without significant gene tree estimation error to accurately infer the species tree because concatenation is expected to fail when ILS is strong (Roch and Warnow 2015; Xi et al. 2015; Springer and Gatesy 2016). To complicate matters, we found evidence of historical introgression in this clade that may have further elevated levels of gene tree discordance (Figure 7B). For this reason, we suggest the MSC tree based on the UCE1000Flank dataset, which had the lowest rates of gene tree estimation error (lowest median and variance in RF distances between gene and species trees) is likely the most reliable (Xi et al. 2015). Nevertheless, in addition to ILS, rapid divergence events can be difficult to reconstruct because the short timeframe also means there is limited phylogenetic signal for those divergence events within individual gene trees (Leaché et al. 2015; Springer and Gatesy 2016; Mclean et al. 2019). Thus, given the lack of agreement across datasets and methods, we would argue that more analysis is needed to confirm the relationships of this clade that diversified rapidly with introgression between non-sister taxa.

The relationship between alignment information content and gene tree error

We examined the effects of data type (coding versus primarily non-coding datasets) and information content (number and proportion of PIS) on gene tree estimation error and found that coding sequences and genes with relatively few PIS contained high rates of gene tree estimation error. Although UCEs conformed to the expected negative association between PIS and gene tree estimation error as measured using RF distance (Figure 6), BUSCOs did not. Instead, we recovered a positive association between the proportion of PIS per alignment and RF distance, possibly because coding genes with more variable sites are evolving faster and thus are more prone to multiple substitutions (i.e., saturation effects), especially at third codon positions which are less constrained. This is also consistent with the bimodal distribution of variable sites in this dataset (Figure 4). Overall, these results show that data type and alignment information content have the potential to compromise phylogenomic inference from gene tree summarization methods that rely on the MSC model, even for datasets that contain relatively large amounts of data (Roch and Warnow 2015; Springer and Gatesy 2016).

We measured rates of gene tree estimation error using RF distances between gene trees and the inferred species tree, under the assumption that increases in RF distance were indicative of higher rates of phylogenetic error. Although gene histories are not expected to be congruent with the species tree in many cases (Tajima 1983; Pamilo and Nei 1988; Maddison 1997), the assumption that RF distances are good indicators of error rates is valid because different datasets contained different means and variances in RF distances among gene trees. This was evident when the three UCE datasets were combined into a single analysis, which showed asymptotic convergence on relatively low mean and variance RF distances after reaching about 500 PIS (Figure 6). The asymptotic shape of this relationship implies that as PIS are added to an alignment, the resulting gene trees converge on a level of heterogeneity that is representative of or approaching real biological signal (i.e., ILS and/or introgression) rather than methodological or data-driven artifacts. Moreover, if all gene tree heterogeneity was due to biological processes such as ILS, one would expect the mean and variance in RF distances to be constant among datasets because these are derived from the same underlying samples. Although some level of gene tree heterogeneity is expected in *all* phylogenomic datasets, even if one assumes the impossible, where empirical gene tree estimation error is absent (Maddison 1997; Gatesy and Springer 2014; Edwards et al. 2016), it is reasonable to assume that differences in the mean and variance in RF distances among datasets are good proxies for differences in gene tree estimation error.

Comparisons of estimation error between data types and the benefits of whole genome phylogenomics

We scrutinized orthologous data harvested from whole genome assemblies to identify the characteristics of phylogenomic datasets that might lead to lower bias. We found that short alignments with relatively few PIS had increased error (measured as difference between gene tree and species tree), a finding consistent with other studies (Meiklejohn et al. 2016; Burbrink et al. 2020). Likewise, we found that long UCEs tend to have low rates of gene tree estimation error and are efficacious markers to use in phylogenomic studies. We also found that the gene trees from the protein-coding dataset derived from BUSCO harvesting and splicing of exons from the tinamou genomes were very noisy. Despite containing large numbers and proportions of PIS (Figure 5) available across many genes, we found high variance in RF distances for these genes and thus conclude that such datasets might be more prone to error than datasets of long UCEs, which had lower gene tree estimation error and behaved more predictably (Figures 4 and 5). As a comparison, the UCE300Flank dataset also had high variance in RF distances (i.e., it contained a mix of low and high error trees), but this variance was well explained by information content. Thus, coding and UCE datasets have different underlying properties of molecular evolution, and these differences should be considered when selecting loci for study and in data analysis. Despite these differences, BUSCO and UCE300Flank datasets converged on the same phylogenetic topology for Clade A, perhaps related to their similar levels of gene tree heterogeneity (Wilcoxon rank sum tests accepted the null hypothesis that RF distances for both datasets did not differ).

Although we derived datasets from whole genome sequences, the UCE100Flank and UCE300Flank datasets are likely to most closely match datasets obtained from typical target capture approaches (Smith et al. 2014; Musher and Cracraft 2018; Tea et al. 2021) and are therefore indicative of rates of error in datasets derived from those protocols. For example, many studies utilize sequences from degraded DNA such as historical museum specimens, which may result in many relatively short UCE alignments. Still, even datasets sequenced from fresh tissues typically only capture about 100-300 bp of flanking region around the conserved UCE core. Thus, we offer multiple recommendations for future phylogenomic studies. First, if available, WGS data are highly preferred to sequence capture datasets because they allow for analysis of a much wider array of data types that can be compared as done herein (Jarvis et al. 2014; Zhang et al. 2019). Such data also allow studies to achieve longer flanking regions around conserved UCE cores that significantly reduce bias during gene tree estimation (Figures 5 and 6), and enable analyses of many other processes of interest, such as evolution of gene families. Moreover, we did not include intron data in the current study, but these data would also be available from WGS and may behave more like UCE data rather than like the nearby exons themselves.

Second, if a very large genome size is cost prohibitive for WGS, then target capture datasets such as UCEs will likely contain high variance in RF distances (i.e., a mix of “good” and “bad” loci), but this variance is likely to be well explained by the number of PIS in each alignment (Figure 6). We therefore suggest that these datasets should be scrutinized and filtered, especially if a rapid radiation is involved (Doyle et al. 2015; S.A. Smith et al. 2018; Mclean et al. 2019; Smith et al. 2023). As the number and proportion of PIS were strongly predictive of gene tree estimation error, we suggest that simply removing alignments with too few PIS may be useful in many cases, as has been suggested elsewhere (Harris et al. 2014; Hosner et al. 2016; Blom et al. 2017; Leite et al. 2021). Given these findings, common phylogenomic approaches that sequence loci with intermediate or low levels of information content such as sequence capture of UCEs may be insufficient to resolve the relationships of taxa that evolved rapidly, such as adaptive radiations, where rates of ILS and short internodal distances caused by rapid diversification are expected to be extreme (Gatesy and Springer 2014; Mclean et al. 2019). In these situations, the number of variable sites may also be too few to resolve gene trees involving short branches.

Finally, although our BUSCO dataset was noisy and included a range of high- and low-error gene trees, we do not suggest that datasets of coding genes are inherently poor quality.

Indeed, protein-coding datasets such as BUSCOs, transcriptomes, and others are likely to remain useful for a range of phylogenomic problems. For example, given their conservation of function, datasets of protein-coding genes will be important for reconstructing phylogenetic relationships at deeper timescales using either nucleotide or amino acid sequences (Dunn et al. 2008; Allen et al. 2017; Zhang et al. 2019; Boyd et al. 2022). Still, given that our BUSCO dataset was noisy compared with our UCE datasets, we suggest exercising caution when applying protein-coding genes to difficult phylogenomic problems (Reddy et al. 2017). Careful fitting of substitution models, data filtering, and perhaps even manual inspection of alignments may reduce bias in these and other datasets of protein-coding sequence (Anisimova and Kosiol 2009; Springer and Gatesy 2016).

However, our results do not suggest that BUSCOs, or protein-coding genes in general, should be filtered by information content or RF distances. Information content was a weak and inconsistent predictor of gene tree estimation error for BUSCOs. Alignments with a fewer number of PIS or a higher proportion of PIS were both associated with increased gene tree error. Likewise, some amount of gene tree heterogeneity is expected due to the coalescent process. Thus, assuming that all high RF gene trees are erroneous and removing them from a dataset could mislead MSC analyses. Thus, more research is needed to understand how best to filter BUSCO datasets, and perhaps protein-coding datasets in general, for phylogenomics using gene tree summarization methods.

Phylogeny and taxonomy of the Tinamous

Our phylogenies based on multiple datasets with varying amounts of gene tree estimation error and information content all resulted in highly similar and well-supported relationships among tinamou species (Figures 1–3, S1–S6). Moreover, concatenation and MSC approaches largely agreed on the overarching relationships at both shallow and deeper timescales. However, for one clade, *Crypturellus* Clade A, we recovered phylogenetic conflict among datasets and species tree approaches. Topologies based on analysis of concatenated data tended to be less supported and less consistent than those estimated from gene trees in ASTRAL, and the UCE100Flank trees showed the lowest support overall. However, MSC trees based on longer UCEs and BUSCOs mostly agreed on the topology of Clade A, only varying in the placement of *C. atrocapillus* and a misidentified downloaded genome (GCA 013389825). The concatenated UCE1000Flank species tree was also most similar to these MSC results, though the placement of *C. kerriae* differed. Therefore, for this subclade, the concatenation results were statistically inconsistent, but the MSC results converged on a nearly identical topology across datasets (other than UCE100Flank, which contained high error) (Bryant and Hahn 2020).

Importantly, our phylogenetic results imply that multiple taxonomic changes are necessary to appropriately classify taxa within Tinamidae. First, *Nothoprocta cinerascens* was recovered as sister to the genus *Rhynchotus*. *Rhynchotus* (Spix 1825) has nomenclatural priority over *Nothoprocta* (Sclater and Salvin 1873), and therefore we suggest moving *N. cinerascens* to the genus *Rhynchotus (Bertelli and Porzecanski 2004; Valqui 2008).* Second, *Taoniscus nanus* was recovered as embedded within the genus *Nothura*, and thus, *Taoniscus* (Gloger 1842) is a junior synonym of *Nothura* (Wagler 1827). We therefore suggest transferring

T. nanus to *Nothura.* Finally, we found evidence for what is likely a new species of *Nothoprocta* (see also Valqui 2008); specimens ascribed to *N. pentlandii* were polyphyletic, with a specimen of the nominate subspecies from Bolivia recovered as sister to *N. perdicaria*, and multiple specimens from Peru recovered as a clade sister to all other *Nothoprocta.* Thus, there appear to be at least two species-level taxa within *N. pentlandii*.

Materials and methods

Sampling

A total of 46 species of tinamous classified into nine genera are currently recognized (Clements et al. 2023). Of these, 22 are monotypic and the remaining species contain two or more subspecies. Altogether there are 175 named taxa (monotypic species plus subspecies) recognized in Tinamidae. We sampled 50 fresh tissues, 12 toe pads from historical museum study skins, and downloaded 10 whole-genome assemblies from the NCBI Assembly Archive, spanning a total of 66 named taxa (subspecies plus monotypic species) and 45 recognized tinamou species. Our sampling included all recognized species except one; a sample initially identified as *Crypturellus boucardi* turned out to be *C. soui meserythrus* (Voucher LSUMZ Birds 180663, Tissue LSUMNS B-53413).

Tinamous belong to the infraclass, Palaeognathae, which includes many extant flightless ratites such as ostriches (Struthionidae) and rheas (Rheidae) along with extinct forms such as Moas (Dinorninithiformes). Recent work has indicated that moas are the sister group to tinamous and that rheas belong to a clade that is sister to tinamous plus moas (Cloutier et al. 2019). Thus, as outgroup taxa for tree rooting, we also downloaded whole genome assemblies for *Rhea pennata* (Greater Rhea) and *Anomalopteryx didiformis* (Little Bush Moa) from the NCBI Assembly Archive. Table S1 lists the 74 samples (72 tinamous plus two outgroups) included in this study.

DNA extraction, library preparation, and whole genome sequencing

For fresh tissues, we extracted genomic DNA (gDNA) using the MagAttract High Molecular Weight kit from Qiagen (Valencia, California). Toe pads extraction of historical study skins was carried out in a dedicated historical DNA extraction laboratory at the Academy of Natural Sciences of Drexel University to reduce contamination risk. Toe pad samples were first washed in a brief bath of EtOH to help remove superficial contaminants, and then soaked in H_2O for three hours to hydrate the desiccated flesh for DNA lysis. Samples were then digested using 180 µL ATL and 30 µL proteinase K for each sample and incubated at 56º C overnight. DNA isolation was then performed using the QiaQuick spin columns and extraction kit from Qiagen (Valencia, CA).

Shotgun sequencing libraries were prepared for each extract using the Hyper library construction kit (Kapa Biosystems). These libraries were sequenced using 150 bp paired-end reads on an S4 lane of an Illumina NovaSeq 6000. These libraries were pooled and tagged with unique dual-end adaptors, and pooling consisted of between 13 and 18 samples per lane aimed to achieve at least 30X coverage of the nuclear genome. Adapters were trimmed and demultiplexed using bcl2fastq v.2.20. We deposited raw reads in the NCBI SRA (Table S1).

Whole genome assembly

We cleaned and then mapped raw reads for each sample (Table S1) to a closely-related scaffold-level assembly from NCBI. Decisions about which downloaded genome to use for read mapping were made based on a previous tinamou phylogeny (Almeida et al. 2022). Specifically, we used fastp (Chen et al. 2018) to clean fastq files, trim adapters and remove low quality reads. Next, we used Burrows-Wheeler Aligner (BWA) version 0.7.17 (Li and Durbin 2009) to map the cleaned reads to reference genomes. Once reads were mapped, we used samtools version 1.6 to convert the resulting sam-files into sorted bam-files (Li et al. 2009) and Analysis of Next Generation Sequencing Data (ANGSD) version 1.2.13 (Korneliussen et al. 2014) to convert bam-files into fasta format. We used '-doFasta 2' to ensure that the consensus nucleotide was written for each polymorphic site.

To assess genome quality, completeness, and redundancy for each assembly, we used Benchmarking Universal Single Copy Orthologs (BUSCO) version 5.3.0 (Simão et al. 2015). BUSCO searches genome assemblies and identifies genes present in single copy using a database of known single-copy orthologs from a clade-specific database of genes. We used the 'aves_odb10' lineage dataset, which utilizes 8,338 genes in the chicken genome. We used the '-- augustus' flag to obtain nucleotide sequences for genes in addition to amino acid sequences. This setting uses augustus version 3.2 (Hoff et al. 2019) to annotate each assembly, and outputs full nucleotide gene sequences for all complete single-copy orthologs in fasta format. This step was necessary to obtain our coding gene dataset for phylogenomic analysis. We also used samtools to estimate mean and standard deviation of sequence coverage for each genome.

Ortholog identification and alignment

We identified and extracted two types of orthologous markers from the WGS scaffolds. First, we utilized nucleotide sequences for the 8,338 single copy orthologs obtained from the BUSCO 'aves_odb10' dataset (henceforth, BUSCO dataset). We used custom scripts to append orthologous genes into the same text file and then used MACSE (Multiple Alignment of Coding SEquences) version 2.06 to align nucleotide sequences (Ranwez et al. 2018). MACSE is a codon-aware alignment algorithm that aligns nucleotide sequences with respect to their protein translation, but allows nucleotide sequences to contain exceptions to this underlying codon structure. This allowed us to accurately align each of the 8,338 gene sequences for all 74 samples.

Second, we harvested ultraconserved elements (UCEs) from each genome assembly. UCEs are highly conserved, typically single-copy, sequences in the genome that are flanked by neutral sites that increase in variability moving away from the conserved core (McCormack et al. 2012). These may overlap with coding or non-coding sequences, but the general pattern of conserved UCE core with variable flanking region should result in very different distribution of variation within a locus than the BUSCO genes, which are purely coding and are most likely to vary at third codon positions due to the redundant nature of the genetic code. To identify and extract UCEs from the genome we followed the pipeline outlined in Phyluce version 1.7.1 (Faircloth 2016) for harvesting and loci from whole genomes. We specifically harvested UCE loci using the Tetrapods-UCE-5kv1 dataset, which includes 5,060 UCEs.

In this pipeline, Phyluce scripts are used to align probe sequences (corresponding to conserved UCE cores) to the genome and then extract those sequences from the genome with a user-specified length of flanking region. Because we were interested in understanding whether and how information content covaried with gene tree estimation error, we harvested UCEs three times, varying the length of the flanking region included with the UCE core each time. We harvested UCEs from each genome that included 100 bp (similar to sequence lengths obtained with target capture from degraded DNA datasets such as historical museum skins), 300 bp (similar to sequence lengths obtained with target capture from standard fresh tissue; e.g., Musher and Cracraft, 2018, Tea et al. 2021), and 1,000 bp (longer than typical of target capture datasets) of flanking region. We then used MAFFT (Katoh and Standley 2013) to align orthologous UCE loci for downstream phylogenomic analysis.

Because missing data can be an additional source of phylogenetic noise, we evaluated each of the four datasets (three UCE and one BUSCO) using only complete (alignments for which all samples are included) alignments. This helped to eliminate gene tree discordance and other sources of error deriving from variation in missing data content.

Phylogenomic analyses

For each of the four datasets, we employed both a concatenation approach using RAxML version 8.2.12 (Stamatakis 2014) and an MSC approach using ASTRAL version 5.7.7 (Mirarab et al. 2014; Zhang et al. 2018). First, we concatenated all alignments using the 'phyluce_align_concatenate_alignments.py' script available in the Phyluce software (Faircloth 2016) for each of the four ortholog datasets. We then used RAxML to reconstruct the phylogeny under a GTR + CAT substitution model which approximates the GTR + GAMMA model and is expected to perform well for large datasets (Stamatakis 2006; Abadi et al. 2019) We then examined the robustness of our phylogenies using the autoMRE option to perform bootstrapping but halt replicates when they converged. Next, for our MSC approach, we estimated gene trees (phylogenies derived from presumed independently sorting loci) for each alignment within each of the four datasets using RAxML. Then, we implemented a quartet-based MSC approach in ASTRAL. ASTRAL approximates the MSC model by estimating the proportions of all possible four-taxon statements (quartets) among all estimated gene trees and then summarizing across the genome. This approach assumes that all gene tree heterogeneity (discordance among estimated gene histories) is due to ILS without significant ancestral introgression or gene tree estimation error. We also pruned our resulting trees to a single clade of interest and compared topologies for this clade using Robinson-Foulds (RF) distances using the 'RF.dist' function in phangorn version 2.11.1 (Schliep 2011).

Assessing ortholog quality and gene tree error

To evaluate general differences among dataset variability, we first estimated the average variability among sites for each alignment within each dataset. To do so, we quantified the average variability for each nucleotide site relative to the center of the alignment using the 'phyluce_align_get_smilogram_from_alignments' script available in the Phyluce package (Faircloth 2016). Then, to assess the relationships between alignment information content, data-type, and gene tree estimation error, we wrote custom scripts in R version 4.3.1 (R Core Team 2023) to estimate the number and proportion (informative sites/alignment length) of parsimony informative sites at each locus, as well as the RF distances (Robinson and Foulds 1981) between each estimated gene tree and the inferred species tree. Parsimony informative sites were defined as variable sites in the alignment where each variant nucleotide is represented by at least two samples. To estimate RF distances, we compared the gene trees from each dataset inferred using RAxML to both the MSC and concatenated trees based on the UCE1000Flank dataset. To test for differences in these measures among datasets, we used Kruskall-Wallace and pairwise Wilcoxon rank sum tests with a Benjamini-Hochberg p-correction implemented in base R (R Core Team 2023). We then ran generalized linear models using information content as the independent and RF distance as the dependent variable for each of the four datasets. Because we found no differences in RF distance when using MSC and concatenated species trees, for these models, we only used RF distances relative to the MSC species tree resulting from our ASTRAL analysis. We compared AIC values for linear and logarithmic fits for each regression and chose the model with the lowest AIC for each dataset. These regression analyses were done once for each dataset, and once with all three UCE datasets combined. Although the latter analysis contains the same UCE locus multiple times, albeit with additional flanking regions, we believe this analysis helps illuminate how the information content of loci from across the spectrum (i.e., very uninformative to very informative) influences rates of gene tree error.

Tests of incomplete lineage sorting and introgression

To examine the effects of ILS and introgression on the phylogenetic discordance of Clade A, we first looked at the alternative quartet topologies for five short internal branches and then used a phylogenomic network approach to test for reticulation (phylogenomic non-bifurcation). For the quartet method, we used the custom scripts for the R package 'MSCquartets' to identify the relative quartet frequencies for five nodes of interest (Rhodes et al. 2021; Allman et al. 2023). Quartets are unrooted four taxon statements for which only three alternative sets of relationships (topologies) exist. Because ILS is expected at relatively short internal branches, we qualitatively evaluated the relative frequencies of alternative topologies at these branches to assess whether ILS might be involved in increasing genealogical heterogeneity, and thus phylogenetic disagreement of Clade A.

To test for putative introgression, we applied a Bayesian approach implemented in PhyloNet 3.8.0 (the 'MCMC_GT' algorithm) (Wen et al. 2016). This method employs reverse-jump Markov Chain Monte Carlo (rjMCMC) to sample the posterior distribution of phylogenetic networks under a multi-species network coalescent (MSNC) model using only rooted gene trees as input data. The MSNC is similar to the MSC, but relaxes the assumption of no introgression by modeling genome evolution as a network rather than bifurcating phylogeny. In doing so, it accounts for both incomplete lineage sorting and reticulation (multiple ancestral lineages that contribute alleles to a single daughter lineage). Such reticulation is typically attributed to historical introgression. We used the gene trees from the UCE1000Flank dataset as these contained reduced rates of gene tree estimation error. We ran the analysis three times, varying the maximum number of reticulate nodes in each run to assess variability in the results. Specifically, we ran MCMC_GT allowing for a maximum of between zero (m=0) and ten (m=10) reticulate nodes. We assigned all samples to the species level, and ran the analysis using species as tips in the network rather than samples (i.e., some species were represented by multiple taxa). We ran the rjMCMC for 5 x10^7 generations with a burn-in of 5x10^6 generations, using the pseudo-likelihood calculation to reduce computation time. Because likelihood scores tended to increase with each successive increase in m-value, and the rjMCMC typically found the maximum allowed in each run, we chose the optimal m-value using a breakpoint analysis. Specifically, using the R package ‘segmented’ (Muggeo et al. 2014), we choose the network with the optimal m-value by fitting a segmented linear model to our likelihoods for each m-value(Muggeo et al. 2014). This allowed us to identify breakpoints in the slope of the regression, where increases in m-value resulted in diminishing gains in likelihood. The m-value at the breakpoint in the segmented model was chosen as the optimal m-value.

Acknowledgements:

We thank Alvaro Hernandez and Chris Wright at the University of Illinois Roy J. Carver Biotechnology Center for assistance with Illumina sequencing. We thank Kimberly K. O. Walden for assistance in depositing Illumina reads to NCBI SRA. We also thank Jorge Doña for providing R code to assist with breakpoint analysis. Dan Lane, Steve Cardiff, and Christopher Witt helped confirm voucher ID’s for multiple samples. We thank B. T. Smith, J. Cracraft, P. Sweet, T. Trombone, S. Katanova (AMNH), J. Bates, S. Hackett, and B. Marks (FMNH), F. Sheldon and D. Dittman (LSUMNS), H.F. James and B. Schmidt (USNM), and N. Rice (ANSP) for loaning material that was crucial for this work. Finally, we thank the members of ANSP Ornithology Department (A. Del Grosso, E. Griffith, K. Kuabara, and J. Merwin) for comments on early versions of the manuscript. This study was funded by NSF grant DEB-1855812 to JDW and KPJ. TAC was partially supported by NSF grant DEB-2203228.

Data Availability:

All raw reads are available on the NCBI Sequence Read Archive (project and sample accession numbers can be found in Table S1). All data and code needed to replicate the analyses found herein can be found at LJM’s github page: https://github.com/lukemusher/Tinamou-phylogenomics.git

Literature Cited

Abadi S, Azouri D, Pupko T, Mayrose I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. *Nat. Commun.* 10:934.

Alaei Kakhki N, Schweizer M, Lutgen D, Bowie RCK, Shirihai H, Suh A, Schielzeth H, Burri R. 2023. A Phylogenomic Assessment of Processes Underpinning Convergent Evolution in Open-Habitat Chats. *Mol. Biol. Evol.* [Internet] 40. Available from:

http://dx.doi.org/10.1093/molbev/msac278

Allen JM, Boyd B, Nguyen N-P, Vachaspati P, Warnow T, Huang DI, Grady PGS, Bell KC, Cronk QCB, Mugisha L, et al. 2017. Phylogenomics from Whole Genome Sequences Using aTRAM. *Syst. Biol.* 66:786–798.

Allman ES, Baños H, Mitchell JD, Rhodes JA. 2023. MSCquartets: analyzing gene tree quartets under the multi-species coalescent. R package version 1.3.1. Available from: https://CRAN.R-project.org/package=MSCquartets

Almeida FC, Porzecanski AL, Cracraft JL, Bertelli S. 2022. The evolution of tinamous (Palaeognathae: Tinamidae) in light of molecular and combined analyses. *Zool. J. Linn. Soc.* 195:106–124.

Altimiras J, Lindgren I, Giraldo-Deck LM, Matthei A, Garitano-Zavala Á. 2017. Aerobic performance in tinamous is limited by their small heart. A novel hypothesis in the evolution of avian flight. *Sci. Rep.* 7:15964.

Anisimova M, Kosiol C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol. Biol. Evol.* 26:255–271.

Bertelli S. 2017. Advances on tinamou phylogeny: an assembled cladistic study of the volant palaeognathous birds. *Cladistics* 33:351–374.

Bertelli S, Giannini NP, Goloboff PA. 2002. A phylogeny of the tinamous (aves: palaeognathiformes) based on integumentary characters. *Syst. Biol.* 51:959–979.

Bertelli S, Porzecanski AL. 2004. Tinamou (tinamidae) systematics: A preliminary combined analysis of morphology and molecules. *Ornitol. Neotrop.* 15 (Suppl):293–299.

Blom MPK, Bragg JG, Potter S, Moritz C. 2017. Accounting for Uncertainty in Gene Tree Estimation: Summary-Coalescent Species Tree Inference in a Challenging Radiation of Australian Lizards. *Syst. Biol.* 66:352–366.

Boyd BM, Nguyen N-P, Allen JM, Waterhouse RM, Vo KB, Sweet AD, Clayton DH, Bush SE, Shapiro MD, Johnson KP. 2022. Long-distance dispersal of pigeons and doves generated new ecological opportunities for host-switching and adaptive radiation by their parasites. *Proc. Biol. Sci.* 289:20220042.

Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst. Biol.* 54:743–757.

Bryant D, Hahn MW. 2020. The Concatenation Question. Pages 3.4: 1–3.4: 23 in Phylogenetics in the Genomic Era (C. Scornavacca, F. Delsuc, and N. Galtier, eds.).

Burbrink FT, Grazziotin FG, Pyron RA, Cundall D, Donnellan S, Irish F, Keogh JS, Kraus F, Murphy RW, Noonan B, et al. 2020. Interrogating Genomic-Scale Data for Squamata (Lizards, Snakes, and Amphisbaenians) Shows no Support for Key Traditional Morphological Relationships. *Syst. Biol.* 69:502–520.

Catanach TA, Halley MR, Allen JM, Johnson JA, Thorstrom R, Palhano S, Thunder CP, Gallardo JC, Weckstein JD. 2021. Systematics and conservation of an endemic radiation of

Accipiter hawks in the Caribbean islands. *The Auk* 138:1–23.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890.

Clements JF, Rasmussen PC, Schulenberg TS, Iliff MJ, Fredericks TA, Gerbracht JA, Lepage D, Spencer A, Billerman SM, Sullivan BL, et al. 2023. The eBird/Clements checklist of Birds of the World: v2023. Available from: Downloaded from https://www.birds.cornell.edu/clementschecklist/download/

Cloutier A, Sackton TB, Grayson P, Clamp M, Baker AJ, Edwards SV. 2019. Whole-genome analyses resolve the phylogeny of flightless birds (Palaeognathae) in the presence of an empirical anomaly zone. *Syst. Biol.* 68:937–955.

Doyle VP, Young RE, Naylor GJP, Brown JM. 2015. Can We Identify Genes with Increased Phylogenetic Reliability? *Syst. Biol.* 64:824–837.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452:745–749.

Edwards SV, Cloutier A, Baker AJ. 2017. Conserved Nonexonic Elements: A Novel Class of Marker for Phylogenomics. *Syst. Biol.* 66:1028–1044.

Edwards SV, Xi Z, Janke A, Faircloth BC, McCormack JE, Glenn TC, Zhong B, Wu S, Lemmon EM, Lemmon AR, et al. 2016. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.* 94:447–462.

Faircloth BC. 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* 32:786–788.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* 61:717–726.

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

Felsenstein J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Syst. Biol.* 27:401–410.

Fong JJ, Fujita MK. 2011. Evaluating phylogenetic informativeness and data-type usage for new protein-coding genes across Vertebrata. *Mol. Phylogenet. Evol.* 61:300–307.

Franz NM, Musher LJ, Brown JW, Yu S, Ludäscher B. 2019. Verbalizing phylogenomic conflict: Representation of node congruence across competing reconstructions of the neoavian explosion. *PLoS Comput. Biol.* 15:e1006493.

Gatesy J, Springer MS. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Mol. Phylogenet. Evol.* 80:231–266.

Gueuning M, Frey JE, Praz C. 2020. Ultraconserved yet informative for species delimitation: Ultraconserved elements resolve long-standing systematic enigma in Central European bees. *Mol. Ecol.* 29:4203–4220.

Harris RB, Carling MD, Lovette IJ. 2014. The influence of sampling design on species tree inference: a new relationship for the New World chickadees (Aves: Poecile). *Evolution* 68:501–513.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-Genome Annotation with BRAKER. *Methods Mol. Biol.* 1962:65–95.

Hosner PA, Faircloth BC, Glenn TC, Braun EL, Kimball RT. 2016. Avoiding Missing Data Biases in Phylogenomic Inference: An Empirical Study in the Landfowl (Aves: Galliformes). *Mol. Biol. Evol.* 33:1110–1125.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346:1320–1331.

Johnson KP. 2019. Putting the genome in insect phylogenomics. *Curr Opin Insect Sci* 36:111–117.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* 15:356.

Kuang T, Tornabene L, Li J, Jiang J, Chakrabarty P, Sparks JS, Naylor GJP, Li C. 2018. Phylogenomic analysis on the exceptionally diverse fish clade Gobioidei (Actinopterygii: Gobiiformes) and data-filtering based on molecular clocklikeness. *Mol. Phylogenet. Evol.* 128:192–202.

Leaché AD, Chavez AS, Jones LN, Grummer JA, Gottscho AD, Linkem CW. 2015. Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome Biol. Evol.* 7:706–719.

Leite RN, Kimball RT, Braun EL, Derryberry EP, Hosner PA, Derryberry GE, Anciães M, McKay JS, Aleixo A, Ribas CC, et al. 2021. Phylogenomics of manakins (Aves: Pipridae) using alternative locus filtering strategies based on informativeness. *Mol. Phylogenet. Evol.* 155:107013.

Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol.* 61:727–744.

Lescroart J, Bonilla-Sánchez A, Napolitano C, Buitrago-Torres DL, Ramírez-Chaves HE, Pulido-Santacruz P, Murphy WJ, Svardal H, Eizirik E. 2023. Extensive Phylogenomic Discordance and the Complex Evolutionary History of the Neotropical Cat Genus Leopardus. *Mol. Biol. Evol.* [Internet] 40. Available from: http://dx.doi.org/10.1093/molbev/msad255

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

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 format and SAMtools. *Bioinformatics* 25:2078–2079.

Li Q, Chen D, Wang S. 2023. Character displacement of egg colors during tinamou speciation. *Evolution* 77:1874–1881.

Maddison WP. 1997. Gene Trees in Species Trees. *Syst. Biol.* 46:523–536.

McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res.* 22:746–754.

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. *PLoS One* 8:e54848.

Mclean BS, Bell KC, Allen JM, Helgen KM, Cook JA. 2019. Impacts of Inference Method and Data set Filtering on Phylogenomic Resolution in a Rapid Radiation of Ground Squirrels (Xerinae: Marmotini). *Syst. Biol.* 68:298–316.

Meiklejohn KA, Faircloth BC, Glenn TC, Kimball RT, Braun EL. 2016. Analysis of a Rapid Evolutionary Radiation Using Ultraconserved Elements: Evidence for a Bias in Some Multispecies Coalescent Methods. *Syst. Biol.* 65:612–627.

Mendes FK, Hahn MW. 2018. Why Concatenation Fails Near the Anomaly Zone. *Syst. Biol.* 67:158–169.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30:i541–i548.

Muggeo VMR, Atkins DC, Gallop RJ, Dimidjian S. 2014. Segmented mixed models with random changepoints: a maximum likelihood approach with application to treatment for depression study. *Stat. Modelling* 14:293–313.

Musher LJ, Cracraft J. 2018. Phylogenomics and species delimitation of a complex radiation of Neotropical suboscine birds (Pachyramphus). *Mol. Phylogenet. Evol.* 118:204–221.

Musher LJ, Ferreira M, Auerbach AL, Cracraft J. 2019. Why is Amazonia a “source”of biodiversity? Climate-mediated dispersal and synchronous speciation across the Andes in an avian group (Tityrinae). *Proc. Roy. Soc. B* [Internet]. Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2018.2343

Ostrow EN, Catanach TA, Bates JM, Aleixo A, Weckstein JD. 2023. Phylogenomic analysis confirms the relationships among toucans, toucan-barbets, and New World barbets but reveals paraphyly of Selenidera toucanets and evidence for mitonuclear discordance. *Ornithology* 140:ukad022.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5:568–583.

Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing.

Nature 526:569–573.

Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Mol. Biol. Evol.* 35:2582–2584.

R Core Team. 2023. R: A language and environment for statistical computing. Foundation for Statistical Computing, Vienna, Austria. Available from: https://www.R-project.org/

Reddy S, Kimball RT, Pandey A, Hosner PA, Braun MJ, Hackett SJ, Han K-L, Harshman J, Huddleston CJ, Kingston S, et al. 2017. Why Do Phylogenomic Data Sets Yield Conflicting Trees? Data Type Influences the Avian Tree of Life more than Taxon Sampling. *Syst. Biol.* 66:857–879.

Rhodes JA, Baños H, Mitchell JD, Allman ES. 2021. MSCquartets 1.0: quartet methods for species trees and networks under the multispecies coalescent model in R. *Bioinformatics* 37:1766–1768.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Math. Biosci.* 53:131–147.

Roch S, Warnow T. 2015. On the Robustness to Gene Tree Estimation Error (or lack thereof) of Coalescent-Based Species Tree Methods. *Syst. Biol.* 64:663–676.

Sackton TB, Grayson P, Cloutier A, Hu Z, Liu JS, Wheeler NE, Gardner PP, Clarke JA, Baker AJ, Clamp M, et al. 2019. Convergent regulatory evolution and loss of flight in paleognathous birds. *Science* 364:74–78.

Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592–593.

Schultz DT, Haddock SHD, Bredeson JV, Green RE, Simakov O, Rokhsar DS. 2023. Ancient gene linkages support ctenophores as sister to other animals. *Nature* 618:110–117.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212.

Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Quéinnec É, Ereskovsky A, et al. 2017. A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Curr. Biol.* 27:958–967.

Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Syst. Biol.* 63:83–95.

Smith BT, Mauck WM, Benz B, Andersen MJ. 2018. Uneven missing data skews phylogenomic relationships within the lories and lorikeets [Internet].

Smith BT, Merwin J, Provost KL, Thom G, Brumfield RT, Ferreira M, Mauck WM, Moyle RG, Wright TF, Joseph L. 2023. Phylogenomic Analysis of the Parrots of the World Distinguishes Artifactual from Biological Sources of Gene Tree Discordance. *Syst. Biol.* 72:228–241.

Smith SA, Brown JW, Walker JF. 2018. So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. *PLoS One* 13:e0197433.

Springer MS, Gatesy J. 2016. The gene tree delusion. *Mol. Phylogenet. Evol.* 94:1–33.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

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

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.

Tan X, Qi J, Liu Z, Fan P, Liu G, Zhang L, Shen Y, Li J, Roos C, Zhou X, et al. 2023. Phylogenomics Reveals High Levels of Incomplete Lineage Sorting at the Ancestral Nodes of the Macaque Radiation. *Mol. Biol. Evol.* [Internet] 40. Available from: http://dx.doi.org/10.1093/molbev/msad229

Tea Y-K, Xu X, DiBattista JD, Lo N, Cowman PF, Ho SYW. 2021. Phylogenomic Analysis of Concatenated Ultraconserved Elements Reveals the Recent Evolutionary Radiation of the Fairy Wrasses (Teleostei: Labridae: Cirrhilabrus). *Syst. Biol.* 71:1–12.

Timilsena PR, Wafula EK, Barrett CF, Ayyampalayam S, McNeal JR, Rentsch JD, McKain MR, Heyduk K, Harkess A, Villegente M, et al. 2022. Phylogenomic resolution of order- and family-level monocot relationships using 602 single-copy nuclear genes and 1375 BUSCO genes. *Front. Plant Sci.* 13:876779.

Valqui T. 2008. Phylogeogaphy of Nothoprocta Tinamous and the The Phylogeny of the Tinamidae.Van Remsen J, editor. Available from: https://search.proquest.com/openview/8d1be590e409938fa5a5afd6ca0b43b1/1?pq-origsite=gscholar&cbl=18750&diss=y

Van Damme K, Cornetti L, Fields PD, Ebert D. 2022. Whole-Genome Phylogenetic Reconstruction as a Powerful Tool to Reveal Homoplasy and Ancient Rapid Radiation in Waterflea Evolution. *Syst. Biol.* 71:777–787.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* [Internet]. Available from: http://dx.doi.org/10.1093/molbev/msx319

Wen D, Yu Y, Nakhleh L. 2016. Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent. *PLoS Genet.* 12:e1006006.

Xi Z, Liu L, Davis CC. 2015. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. *Mol. Phylogenet. Evol.* 92:63–71.

Zhang C, Mirarab S. 2022. Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees. *Mol. Biol. Evol.* [Internet] 39. Available from: http://dx.doi.org/10.1093/molbev/msac215

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19:153.

Zhang F, Ding Y, Zhu C-D, Zhou X, Orr MC, Scheu S, Luan Y-X. 2019. Phylogenomics from low-coverage whole-genome sequencing. *Methods Ecol. Evol.* 10:507–517.

Zhao M, Kurtis SM, White ND, Moncrieff AE, Leite RN, Brumfield RT, Braun EL, Kimball RT. 2023. Exploring Conflicts in Whole Genome Phylogenetics: A Case Study Within Manakins (Aves: Pipridae). *Syst. Biol.* 72:161–178.

Tables and Figures:

Dataset	Mean #PIS	StDev #PIS	Mean %PIS	StDev %PIS	Mean RF	StDev RF
BUSCOs	922.5146	690.084	0.3717782	0.1033898	58.5241	21.11341
UCE100Flank	45.20298	28.84223	0.1347156	0.0828647	106.501	20.10551
UCE300Flank	206.8736	88.36566	0.2648982	0.1048726	59.1615	18.40835
UCE1000Flank	1035.713	250.5732	0.4178852	0.08261385	29.8675	8.468606

Table 1: Means and standard deviations of parsimony informative sites (PIS) and Robinson-Foulds Distances (RF) for each dataset analyzed for the study.

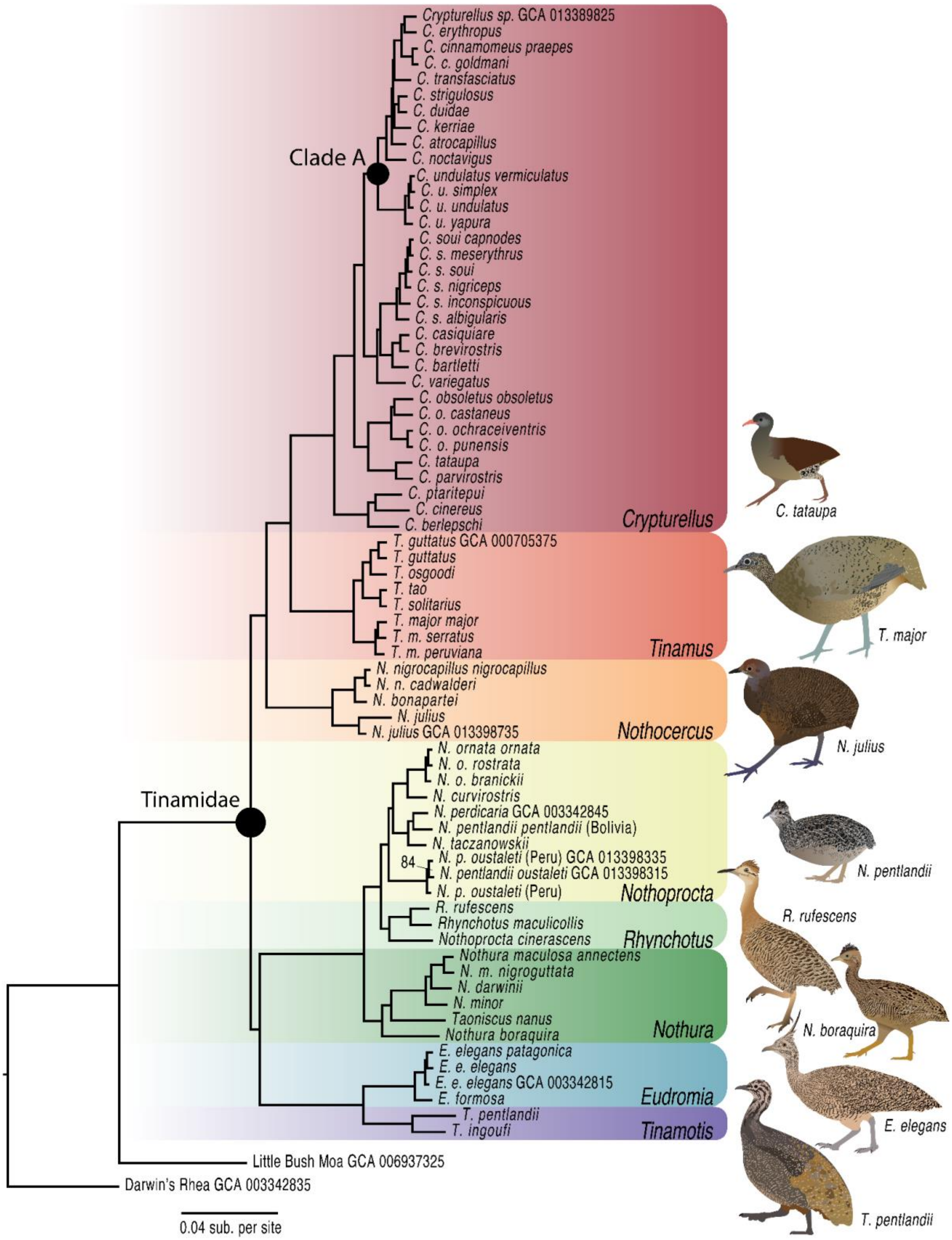

Figure 1: Concatenated Phylogeny of tinamous based on UCEs with 1,000 bp of flanking sequence. All nodes have bootstrap value of 100% except where noted. Note relatively short internodal branch lengths in Clade A. Tinamou illustrations by TAC.

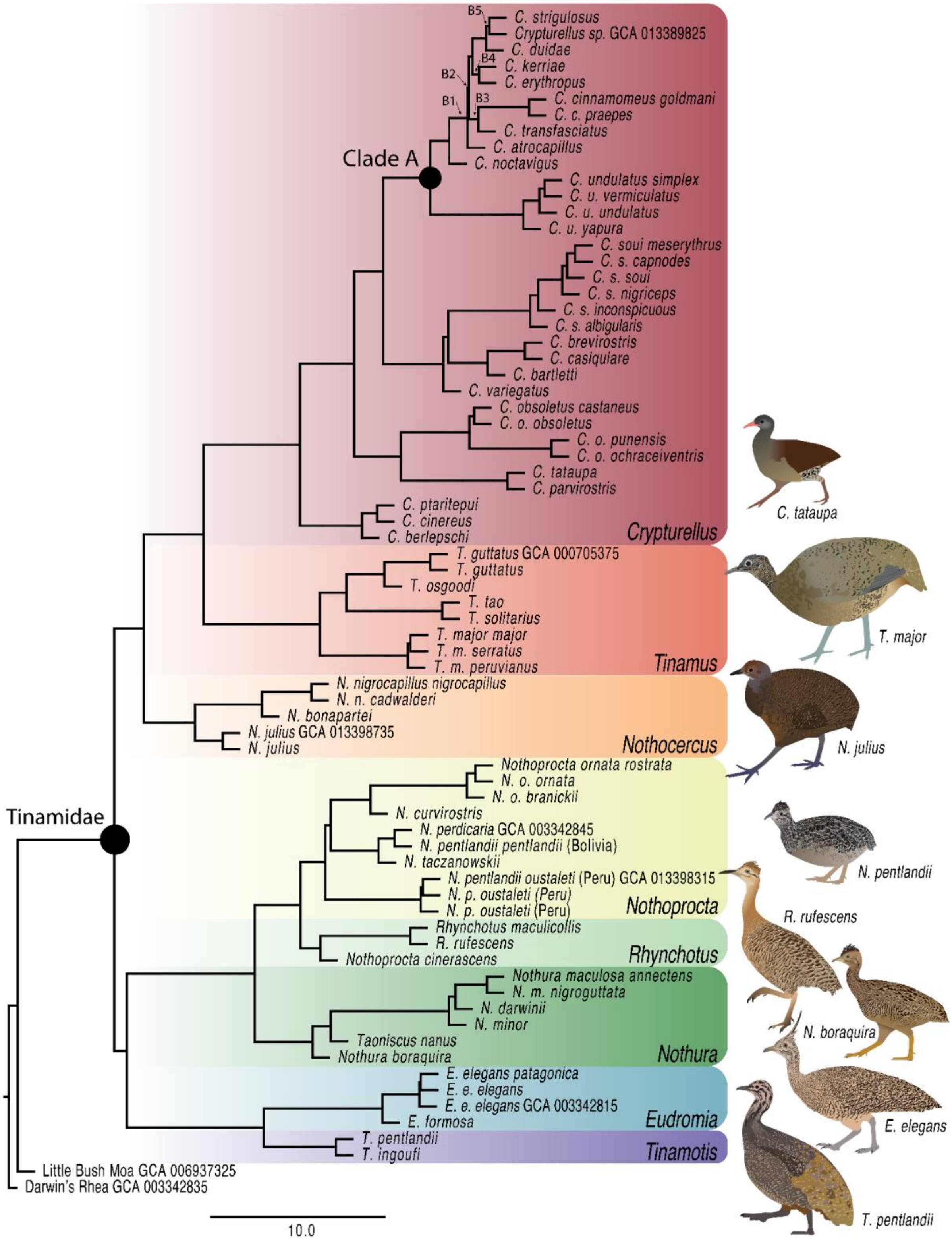

Figure 2: Multi-species Coalescent (ASTRAL) Phylogeny of tinamous based on UCEs with 1,000 bp of flanking sequence. All nodes have posterior probability of 1. Branches labeled B1–B5 denote rapidly diverging branches evaluated using quartet analysis (see Figure 7). Tinamou illustrations by TAC.

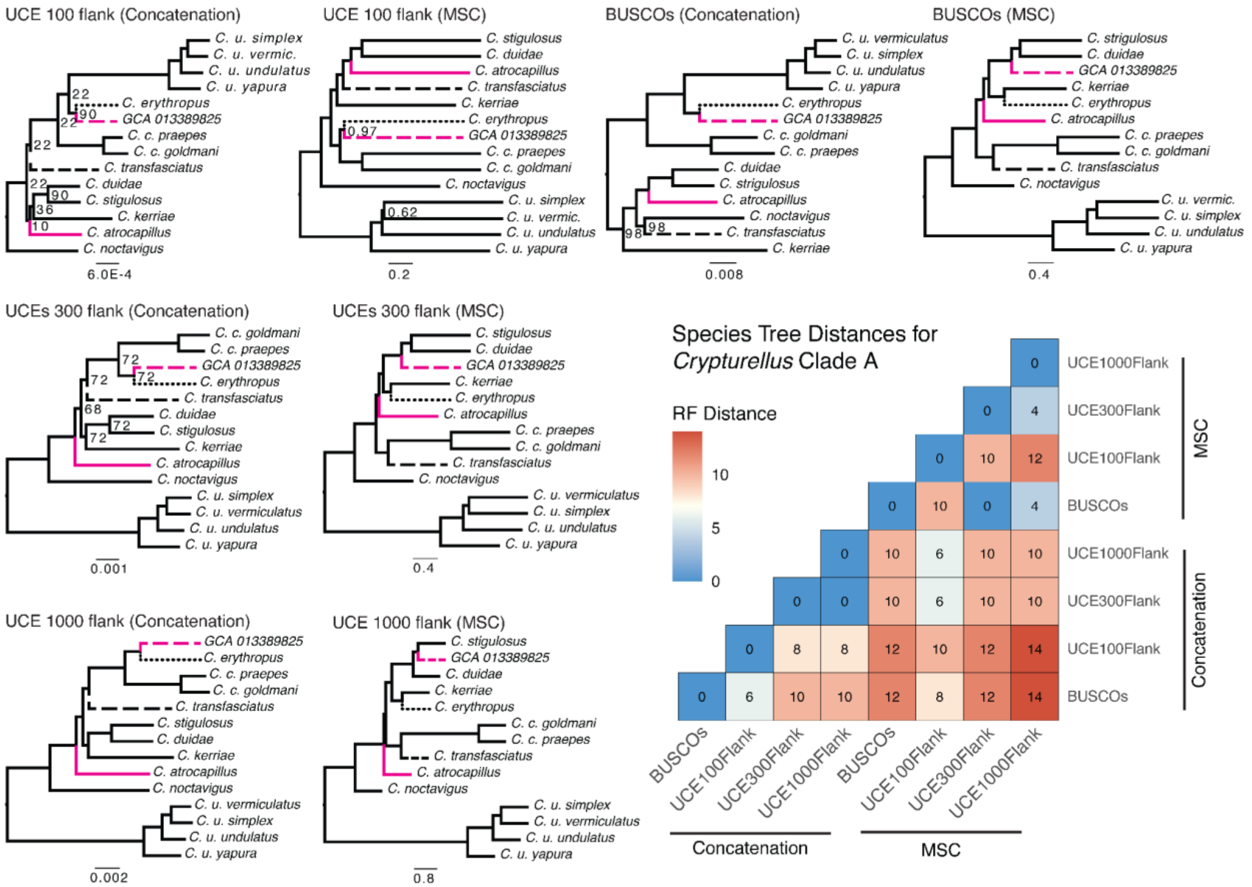

Figure 3: Phylogenetic topologies for a difficult to resolve clade (Clade A) of *Crypturellus* tinamous. The phylogenetic placement of multiple taxa was discordant among datasets and species tree approaches. These include (1) *C. erythropus* (black short-dashed branch), (2) *C. transfasciatus* (black long-dashed branch), (3) *C. atrocapillus* (pink solid branch), and (4) a misidentified genome downloaded from NCBI that was labeled as *C. undulatus*, but shows phylogenetic affinities with *C. strigulosus* and *C. erythropus* (GCA 013389825; pink long-hashed branch). The top row of trees shows topologies for each dataset estimated with RAxML after concatenation, whereas the bottom row of trees shows topologies estimated with ASTRAL. Numbers on the nodes correspond to posterior probabilities (MSC trees only) or bootstrap percentage values (concatenation trees only). Nodes with PP = 1.0 or bootstrap = 100% not shown. Scale bars under each topology indicate substitutions per site for concatenated trees or coalescent units for MSC trees. Bottom right heatmap shows Robinson-Foulds distances between all pairs of trees.

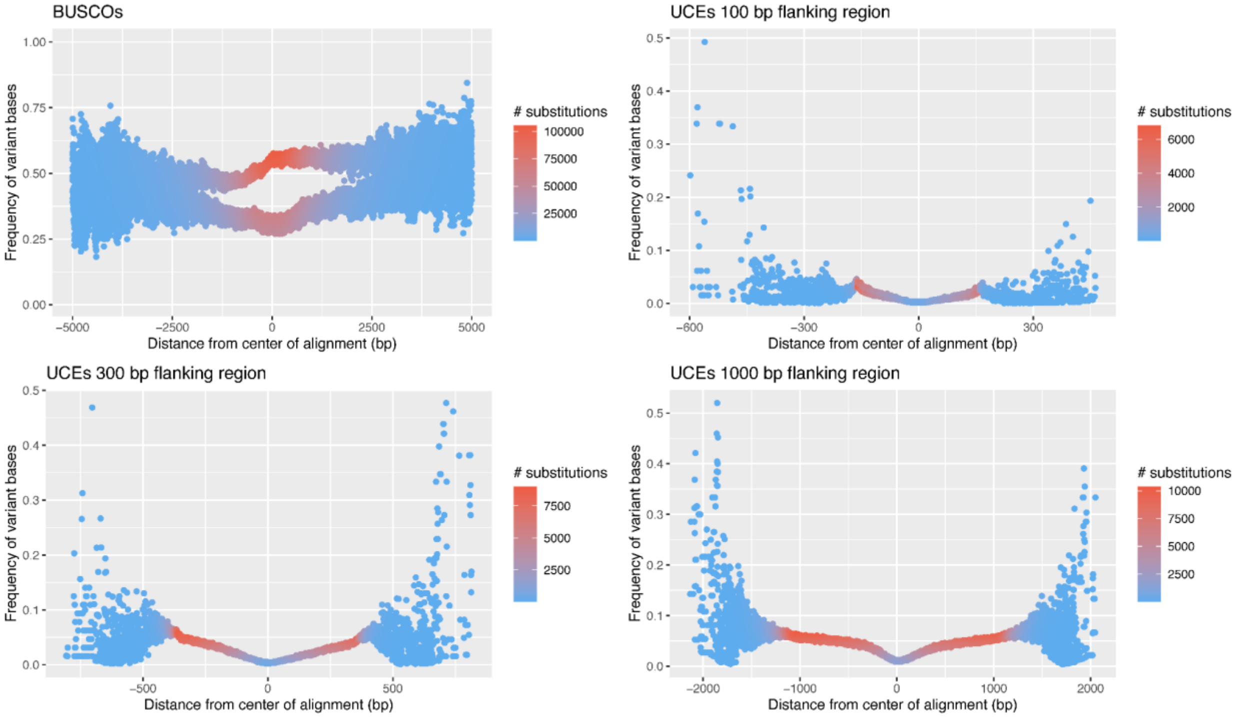

Figure 4: Smilograms for each dataset showing that variation within coding genes is distributed differently than within UCEs. Each point represents a base pair position, defined by its distance from the center, across all alignments in each dataset. UCEs show increasing frequency of variant sites with increasing distance from the UCE core. BUSCOs do not differ by distance from core, but instead show a bimodal distribution of frequency of variant bases likely associated with differences in variability between first and second versus third codon positions. Points are colored by the total number of substitutions at a given site across all samples and alignments. The outer regions of each plot show relatively few total substitutions because relatively few alignments in each dataset are very long.

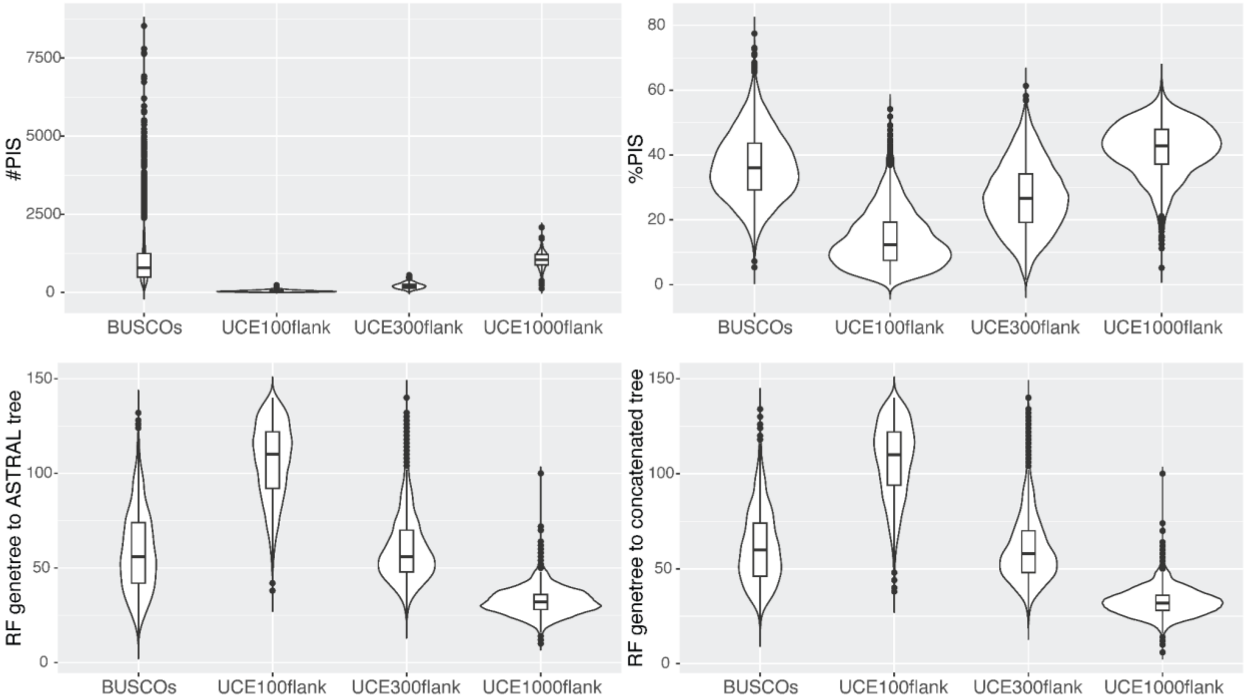

Figure 5: Information content and gene tree heterogeneity for each dataset. The top row of violin plots shows information content for each dataset using either the number of parsimony informative sites (#PIS) or the percentage of parsimony informative sites (%PIS) for each dataset. The bottom row of violin plots shows the gene tree heterogeneity for each dataset measured by the Robinson-Foulds distance between each gene tree and the inferred species tree (assuming the UCE1000Flank dataset). The bottom left plot assumes the MSC tree as the species tree, whereas the bottom right plot assumes the concatenated tree as the species tree. Because there is variation in mean and variance of RF distance among datasets, higher RF distances are indicative of more erroneous gene trees.

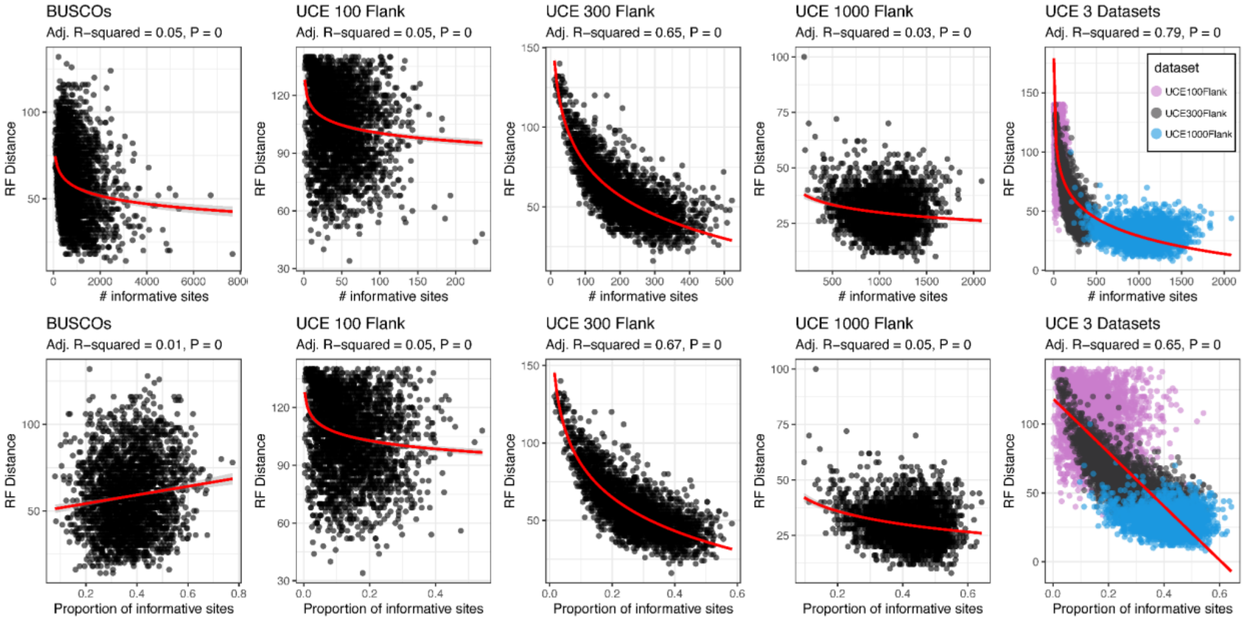

Figure 6: Covariation between information content and gene tree estimation error. Linear regression models exploring the relationship between information content and gene tree heterogeneity. The negative correlation in most instances suggests that alignments with less information content result in increased gene tree estimation error.

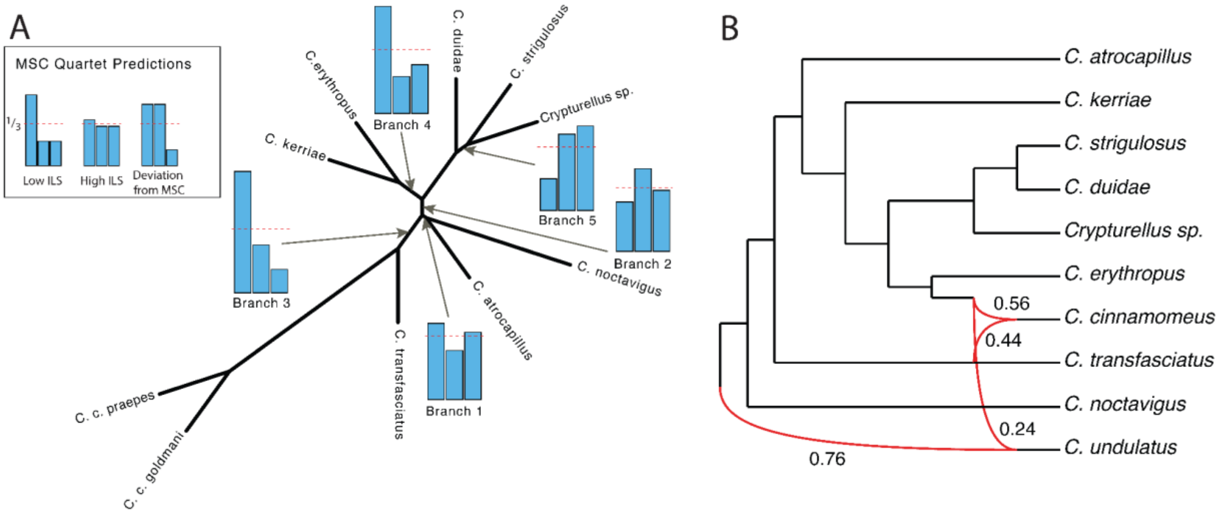

Figure 7: Incomplete lineage sorting and genomic reticulation in the tinamou phylogeny. (A) Relative quartet frequencies for five short internal branches in Clade A based on the UCE1000Flank dataset. Bar graphs depict the relative frequencies of each of three alternative unrooted quartet topologies (for any unrooted four taxon statement there exist only three alternative sets of relationships). The stippled lines indicate the ⅓ threshold for the frequency of gene trees given a multispecies coalescent model, which predicts a single majority quartet topology consistent with the true species tree with a frequency >⅓, and two minority topologies with equivalent relative frequencies < ⅓. As quartet frequencies approach the ⅓ threshold, they indicate stronger ILS. Major deviations from the expectations of the MSC model indicate violations of the model may be present, such as gene tree estimation bias or introgression. (B) Phylogenetic network results for the network with the optimal m-value (m=2) from PhyloNet showing non-bifurcating relationships among a difficult to resolve clade of *Crypturellus* tinamous (Clade A). Red branches demarcate reticulation, multiple ancestors contributing to a single daughter branch. Inheritance probability, or the proportion of the discordant gene trees that can be attributed to introgression, are marked on each reticulate branch. Branches are labeled 1–5 and correspond to labeled branches in Figure 2.