Categorical edge-based analyses of phylogenomic data reveal conflicting signals for difficult relationships in the avian tree

Ning Wang^{1,3,4*}, Edward L. Braun², Bin Liang^{1,4}, Joel Cracraft³, Stephen A. Smith¹

¹Department of Ecology & Evolutionary Biology, University of Michigan, 1105 N University Ave, Ann Arbor, MI 48109-1048, USA

²Department of Biology, University of Florida, Gainesville, FL 32607, USA

³Department of Ornithology, American Museum of Natural History, New York, NY 10024, USA

⁴College of Life Sciences, Inner Mongolia University, Hohhot, 010070, China

***Correspondence**

Ning Wang

Current address: College of Life Sciences, Inner Mongolia University, Hohhot, 010070, China

Email: ningwang83@hotmail.com

ABSTRACT:

Phylogenetic analyses of large-scale datasets sometimes fail to yield a satisfactory resolution of the relationships among taxa for a number of nodes in the tree of life. This has even been true for genome-scale datasets, where the failure to resolve relationships is unlikely to reflect limitations in the amount of data. Gene tree conflicts are particularly notable in studies focused on these contentious nodes in the tree of life, and taxon sampling, different analytical methods, and/or data-type effects are thought to further confound analyses. Observed conflicts among gene trees arise from both biological processes and artefactual sources of noise in analyses. Although many efforts have been made to incorporate biological conflicts, few studies have curated individual genes for their efficiency in phylogenomic studies. Here, we conduct an edge-based analysis of Neoavian evolution, examining the phylogenetic efficacy of two recent phylogenomic bird datasets and three datatypes (ultraconserved elements [UCEs], introns, and coding regions). We assess the potential causes for biases in signal-resolution for three difficult nodes: the earliest divergence of Neoaves, the position of the enigmatic Hoatzin (*Opisthocomus hoazin*), and the position of owls (Strigidae). We observed extensive conflict among genes for all data types and datasets even after we removed potentially problematic loci. Edge-based analyses increased congruence and examined the impact of data type, GC content variation (GC_{CV}), and outlier genes on analyses. These factors had different impact on each of nodes we examined. First, outlier gene signals appeared to drive different patterns of support for the relationships among the earliest diverging Neoaves. Second, the position of Hoatzin was highly variable, but we found that data type was correlated with the signals that support different placements of the Hoatzin. However, the resolution with the most support in our analyses was Hoatzin + shorebirds. Finally, GC_{CV} , rather than data type (i.e., coding vs non-coding) *per se*, was correlated with an owl + Accipitriformes signal. Eliminating high GC_{CV} loci increased the signal for an owl + mousebird relationship. Difficult edges (i.e., characterized by deep coalescence and high gene-tree estimation error) are hard to recover with all methods (including concatenation, multispecies coalescent, and edge-based analyses), whereas “easy” edges (e.g., flamingos + grebes) can be recovered without ambiguity. Thus, the nature of the edges, rather than the methods, is the limiting factor. Categorical edge-based analyses can reveal the nature of each edge and provide a way to highlight especially problematic branches that warrant further examination in future phylogenomic studies. We suggest that edge-based analyses provide a tool that can increase our

48 understanding about the parts of the avian tree that remain unclear, even with large-scale data. In
 49 fact, our results emphasize that the conflicts associated with edges that remain contentious in the
 50 bird tree may be even greater than appreciated based on previous studies.

51
 52 *keywords:* edge-based analyses; Neoavian resolution; GC content variation; outlier signal

53

1. Introduction

The increasing availability of genomic datasets is increasingly facilitating the resolution of phylogenetic relationships. However, recent studies have highlighted that as datasets have grown, so has conflict. This may involve gene-tree discordances within a single study (e.g., Wang et al. 2019) or the resolution of difficult and controversial hypotheses found across multiple studies (e.g., Wickett et al. 2014; Puttick et al. 2018). Conflicting gene trees can be the result of biological processes such as hybridization, gene duplication and loss, and incomplete lineage sorting (ILS) or they can be the result of data-set assembly error (Gatesy and Springer 2017) or other artificial noise arising from, for example, orthology inference, sequence alignment, model specification, and insufficient informative variation (Springer and Gatesy 2019). In part, as a result of these conflicts, several areas across the tree of life have consistently been difficult to resolve (Wang et al. 2012; Tarver et al. 2016; Shen et al. 2017; Wang et al. 2019) despite the use of genome-scale datasets (Foley et al. 2016) and increased taxon sampling (Prum et al. 2015; Wang et al. 2017). Additionally, different analytical methods (e.g., concatenation-based or coalescent-based methods), and/or the inclusion or exclusion of multiple data-types (protein coding or noncoding regions, Reddy et al. 2017) can result in conflicting resolutions.

Data limitations such as insufficient sampling of taxa or gene regions can be overcome with increasing effort and worldwide collaboration, as well as advances in DNA sequencing techniques that can obtain DNA from ancient specimens preserved in museums that complement lack of fresh specimen (e.g., Chen et al. 2018). On the other hand, resolution of methodological issues such as inaccurate evolutionary models may be more difficult to address. In particular, difficult-to-resolve edges on the tree of life may be the result of low signal within very short internodes. Concatenation can overcome low signal by combining the signal from multiple genes but also assumes that all genes share a single evolutionary history, yet population genetic and molecular evolutionary processes can result in highly heterogeneous gene trees and violate this assumption (Kolaczkowski and Thornton 2004; Kubatko and Degnan 2007; Edwards 2009). The most commonly used summary coalescent methods (e.g., ASTRAL III, Zhang et al. 2018; MP-EST, Liu et al. 2010), on the other hand, accommodate gene tree conflicts but assume accurately estimated gene trees for neutrally evolving gene regions that are free of intralocus recombination and only exhibit conflict due to ILS (Springer and Gatesy 2016). Typical phylogenomic datasets violate many of these assumptions. If there are numerous gene trees that conflict for reasons

other than ILS (whether this reflect other biological processes or estimation error) the assumptions of coalescent methods are violated and the methods would not be a consistent estimator of the species tree (although many methods will still be relatively robust to these errors). Also, when analyzing large genomic datasets, many genes will have experienced episodic periods of strong selection, or linkage and hitchhiking can also cause violation of neutral evolution for most genomic regions (e.g., Bendall et al. 2016). Some genes may even show within-gene conflict, supporting multiple trees (Kimball et al. 2013; Mendes et al. 2019). These observations highlight the importance of examining individual genes in phylogenetic analyses. Furthermore, while biological processes may drive some conflict, errors in alignment, assembly, and gene-tree estimation may be common, especially as dataset size increases, with larger datasets being more difficult to curate (Gatesy and Springer 2017; Liu et al. 2017).

With ~10,500 living species, birds represent the most species-rich group of tetrapod vertebrates. Despite their importance, several controversial relationships in Neoaves have not been settled, even with genome-scale data (e.g., Jarvis et al. 2014; Prum et al. 2015; Reddy et al. 2017). Challenges to resolving these nodes are not unexpected considering that neoavian birds likely experienced a very rapid radiation (e.g., Ericson et al. 2006; Jarvis et al. 2014; Prum et al. 2015; Houde et al. 2020). Although genome-scale datasets have facilitated the recovery of many relationships (e.g., passerines + parrots, flamingos + grebes, “core” landbirds, “core” waterbirds, sunbittern + tropicbirds, sandgrouse + mesites; Ericson et al. 2006; Hackett et al. 2008; Suh et al. 2011; Wang et al. 2012; McCormack et al. 2013; Jarvis et al. 2014; Prum et al. 2015; Suh et al. 2015), the Neoaves are still well-known for multiple, extremely recalcitrant nodes (Jarvis et al. 2014; Prum et al. 2015). Specifically, the earliest divergences of Neoaves remain one of the most difficult phylogenetic problems in vertebrates, which has even been suggested to represent a hard polytomy (Suh 2016, but see Reddy et al. 2017). Within Neoaves, the positions of Hoatzin (*Ophisthocomus hoazin*) and owls (Strigiformes) remain especially problematic as well as being significant for the analysis of avian development, life history, and morphological traits. The Hoatzin is a bizarre and enigmatic bird species. It converts plant cellulose to simple sugars using microbial foregut fermentation (Domínguez-Bello et al. 1994) and has a modified skeleton (sternum) to accommodate a large crop. Also, Hoatzin chicks have claws on two of their wing digits (Hughes and Baker 1999). These characteristics have confounded its phylogenetic position for decades. Owls are nocturnal predators with raptorial features, but their relationships have

been obscure. Within Neoaves there are at least two distinct raptorial lineages. The first is the falcons (Falconidae), which have now been shown to be related to parrots and passerines, whereas the second, Accipitriformes, includes the hawks, eagles, and vultures. Presently, it remains unclear whether owls represent a third raptorial lineage or whether they are sister to hawks, eagles, and vultures. This uncertainty impacts our understanding of the evolution of raptorial features.

Recent avian studies have suggested that phylogenetic reconstruction is biased by the source of the data (e.g., coding regions [hereafter exons], introns, ultraconserved elements [UCE], Reddy et al. 2017), which we refer to as data-type effects. For example, in Jarvis et al. (2014), concatenated analyses of exons, introns, and UCEs were largely discordant among each other. The resolution of two of the above three difficult relationships (specifically, the position of owls and the topology for the earliest split of Neoaves) has been shown to vary among different data types (Reddy et al. 2017; Braun et al. 2019). The third difficult relationship – the position of the Hoatzin in the bird tree – has been even more variable across studies (e.g., Jarvis et al. 2014; Reddy et al. 2017) and the existence of a correlation with data type is less clear. Due to biased base composition and high GC content variation (hereafter GC_{CV}), exons received the most criticism for model violation and impact on phylogenetic reconstruction (Jarvis et al. 2014; Reddy et al. 2017). Adjusting model heterogeneity to better fit different data (e.g., Galtier and Gouy 1998; Morgan et al. 2013) might be applicable, but the dramatic heterogeneity among gene trees is still notable even within the same data type (e.g., Jarvis et al. 2014). Nevertheless, were exons to be generally misleading in the relationships they reconstruct, this would have far reaching implications for phylogenomic analyses as transcriptomic and targeted sequencing data have become more common (e.g., Prum et al. 2015; Wang et al. 2019).

Herein we examine conflicting signals within two major avian datasets: Jarvis et al. (2014) and Prum et al. (2015). Specifically, we conducted constraint phylogenetic analyses (Smith et al. 2020, hereafter edge-based analyses) to examine the three confounding questions regarding relationships within the large clade Neoaves (i.e., the earliest divergence within Neoaves, the position of Hoatzin, and the position of owls). We explore the efficacy of large datasets to resolving these relationships by carefully curating individual genes and assessing their contribution to the support for alternative hypotheses. In addition to individual gene evaluations, we focused on dissecting the phylogenetic signal associated with different data types (exon,

intron, and UCE) and we examined other factors, such as variation in GC content (Foster and Hickey 1999) and gene evolutionary rate (Chojnowski et al. 2008) that also might influence phylogenetic estimation. The inclusion of multiple data types facilitates examination of whether phylogenetic signal from protein-coding regions were different from that of non-coding regions. The edge-based analyses (see detailed description below) examine phylogenetic signal by investigating alternative resolutions for individual edges, and similar approaches have been successfully applied to determine support for specific controversial relationships (e.g., Walker et al. 2018; Wang et al. 2019; Smith et al. 2020). Our use of edge-based analysis is not designed to resolve these contentious relationships but explore a method for identifying the nature of difficult relationships, especially around short internodes. Through these analyses, we seek to better characterize the ability of large datasets to address difficult phylogenetic questions in birds and other groups.

2. Materials and methods

2.1 Data acquisition and filtering

We obtained the alignment data including 8251 exons, 2516 introns, and 3769 UCEs from Jarvis et al. (2015). In order to increase the phylogenetic signal from individual genes and eliminate potential systematic error or noise (e.g., incorrect orthologous inference from the original study), we conducted several filtering steps. First, genes having a short alignment (< 500bp) and/or missing taxa (< 42 species after deleting sequences with > 70% gaps in the alignment) were excluded. Second, we reconstructed ML gene trees using IQTREE (v. 1.6.3, Nguyen et al. 2014) with the GTR + Γ model and 1000 ultrafast bootstrap (UFBS) replicates. Trees having outgroups that clustered within ingroups (i.e., non-bird species nested within birds, and/or species from Palaeognathae and Galloanserae nested within Neoaves) were treated as systematic error and excluded from additional analyses. Finally, trees with extremely long branch lengths (> 1 expected substitutions per site) were also excluded due to potential error in ortholog identification. Scripts are available at https://bitbucket.org/ningwang83/avian_conflict_assess. In addition, the original 259 sequence alignments and gene trees from Prum et al. (2015) were downloaded and filtered following a same procedure.

2.2 Examination of gene tree resolution diversity

In general, we focused on three controversial edges on the avian tree of life: the earliest divergence in Neoaves and those encompassing the positions of the Hoatzin and owls (Table 1). To assess the diversity of resolutions in the gene trees, we conducted several analyses. First, we examined the topological distributions of ML gene trees (Fig. 1) in relation to the three focal edges under different support cutoffs (i.e., $\geq 0\%$, or 50%, or 70%, or 95% support from either regular bootstrap [BS] analyses or Ultra Fast Bootstrap [UFBS], Minh et al. 2013, Hoang et al. 2018). Second, we examined Robinson–Foulds (RF, Robinson and Foulds 1981) distances among gene-trees for each data type, as well as between each gene tree and the concatenated TENT ExaML species tree from Jarvis et al. (2014) and the concatenated RAxML tree from Prum et al. (2015). RF distances were calculated using the bp script (<https://github.com/FePhyFoFum/gophy>), which allows non-overlapping taxa in the gene trees. Finally, to quantify the relative strength of the phylogenetic signal for a certain edge, we calculated the ICA (internode certainty all; the degree of certainty for a bipartition in a given set of trees in conjunction with all conflicting bipartitions in the same tree set) values (Salichos et al. 2014) on the species trees (same as above) from the filtered gene trees of each data type respectively. These analyses were conducted using RAxML 8.2.11 (Stamatakis 2014) with the GTRCAT model (Stamatakis 2006) and lossless support to accommodate partial gene trees (Kobert et al. 2016).

2.3 Assessment for constraint topologies for three focal edges

We selected several alternative topologies from previous large-scale studies (Table 1) for each of the three focal edges. For the Jarvis et al. (2014) data, gene trees were reconstructed in IQTREE by giving one constraint edge at a time while letting the remainder of the tree to be optimized with the best ML topology. This analysis was conducted for each locus and every focal edge (see example flow chart in Fig. 2), but genes were filtered out if their gene tree was incompatible with any testing constraint under a 95% UFBS cutoff. For example, any ML gene tree showing Hoatzin + Loons, which is conflict with all the seven test hypotheses (Table 1), would be counted in the edge analysis (Fig. 2B) if Hoatzin + Loons shows UFBS support $< 95\%$ (i.e., if it is unsupported) in the ML tree. However, if the UFBS support for Hoatzin + Loons was $\geq 95\%$, we would remove the gene from analyses due to its strong conflict with all of the test hypotheses. To identify the best hypothesis, we conducted the following analysis:

- (1) We compared $\ln L$ scores among conflicting alternative topologies for a single edge for each gene and identified its supporting (i.e., the best) topology.
- (2) We calculated the difference in $\ln L$ ($\Delta \ln L$) between the best and the second-best resolutions for each gene. At this stage we also identified potential outlier genes (cases where $\Delta \ln L$ fell out of the smooth distribution among other genes) by eye.
- (3) We summed $\Delta \ln L$ for each constraint from supporting or significantly supporting ($\Delta \ln L > 2$) genes ($\sum \Delta \ln L$ or $\sum \text{sig} \Delta \ln L$).
- (4) We identified the overall best topology either supported or significantly supported ($\Delta \ln L > 2$) using two criteria: the largest number of genes ($\sum N_g$, note that this value is simply the number of genes supporting a specific topology) and the highest $\sum \Delta \ln L$ (or $\sum \text{sig} \Delta \ln L$) scores (Fig. 2).

Genes for which the $\ln L$ could not be calculated given specific constraint topologies (this occurs when relevant taxa are missing) were excluded from the calculation. For genes that strongly support one hypothesis versus others (i.e., outlier genes, Fig. 2), we manually examined them for misalignment. We also collected information about the functions of the genes. We did this because positive selection can give rise to convergent evolution, at least in principle, and genes with a specific function might group corresponding species together and mislead phylogenetic inferences (e.g., Stern 2013). For example, hearing gene *Prestin* grouped echolocating taxa (microbats and toothed whales), presumably reflecting convergence (Li et al. 2010). Thus, we identified exons with outlier topologies using BLAST to see if those outlier genes exhibit specific functions that might be relevant to the topology they strongly supported. However, we did not conduct lineage-specific selection analyses for outlier exons as their alignments contained several premature stop codons.

To examine taxon-sampling effect, we also conducted edge-based analyses using genes from the large sampling of Prum et al. (2015), following similar processes and criteria indicated above. In order to limit CPU time, the RAxML 8.2.11 (Stamatakis 2014) with GTRCAT model was used. The ML gene-trees including BS support built with RAxML were directly retrieved from Prum et al. (2015). We also reduced the number of constraint topologies for each focal edge by including only the two published hypotheses that are best supported by Jarvis et al. (2014) data, as well as the one shown in Prum et al. (2015). Results were organized after filtering out loci with gene-tree incompatibilities with any testing constraints under 70% BS cutoff. In

addition, to examine the sampling effect on the hypotheses examined here, we focused on the earliest diverging Neoaves and assessed five additional hypotheses by randomly referring one single lineage (i.e., caprimulgiforms [nightjars], Hoatzin, charadriiforms [represented by plovers], gruiforms [cranes], or cuculiforms [cuckoos], Fig. 3) as sister to other Neoaves and conducted edge-based analyses with Jarvis et al. (2014) data using the same criteria mentioned above.

In order to assess influence of GC_{CV} and tree length (i.e., gene evolutionary rate) on the phylogenetic reconstruction using different data types, we sorted the filtered genes from Jarvis et al. (2014) based on their GC_{CV} and tree length. Specifically, we expanded the method from Reddy et al. (2017) by limiting GC content calculations to variable sites and comparing only within bird species (i.e., excluding nonavian outgroups) to generate the ΔGC_{50} (the GC difference between the inter-quartile GC-rich taxa), as well as the ΔGC_{90} (the GC difference between the 95 and 5 percentile GC-rich taxa) and the ΔGC_{100} (the GC difference between the most and least GC-rich taxa). We ordered genes based on ΔGC_{90} to balance the GC difference and similarity among data types (Fig. 4a) and examined the gene trees in the upper and lower quartiles of the variation for their supporting pattern to each testing edge (Fig. 4b). We conducted similar analyses by sorting genes based on their tree length and compared the edge assessment results between gene sets from the upper and lower quartile of the tree lengths (Fig. 4c and d).

2.4 Coalescent simulation – assessing the performance of edge-based analyses for deep coalescence

We conducted simulation analyses to assess the performance of edge-based analyses in recovering the potential true topology at edges under differing level of ILS. The Jarvis et al. (2014) MP-EST* TENT tree, which was built with a binned MP-EST* method (Mirarab et al. 2014), was selected as the true tree. 500 gene trees were then simulated under the neutral coalescent model using the *treesim.contained coalescent_tree* module in Dendropy 4.2.0 (Sukumaran and Holder 2010). We rescaled the internal branches (hereafter inbrlen) of these gene trees with 0.01 that is estimated by comparing the coalescent branch length with the substitution branch length in the empirical data of Jarvis et al. (2014). The tip branch-lengths (hereafter tipbrlens) were rescaled based on species' tipbrlen distribution of the empirical data of Jarvis et al. (2014). A recent study (Duchêne et al. 2018) found that a proportionally linked

model is the best fit for gene tree branch length differences, meaning the tipbrlens are proportional among species across gene trees. Under this condition, we first obtained the branch-lengths distribution for each species from the empirical ML gene trees of Jarvis et al. (2014). Then, we generated tipbrlens for each simulated gene by selecting the same percentile tipbrlens value from the empirical distribution of each species. This rescaling method ensures a proportional tipbrlens across all gene trees. It is noteworthy that the branch length distributions, as well as the modeling parameters (see below) were estimated from exons, introns and UCEs, respectively, to simulate different types of data. Here, we conducted ASTRAL analyses (ASTRAL III, Zhang et al. 2018) using these rescaled true gene trees (for each datatype) to build a coalescent species tree (hereafter simAstral). Moreover, the average RF distance (ARFD) between the true species tree and the true gene trees was used to reflect the level of overall ILS.

We generated one sequence alignment for each tree with length of 800-3000bp. Seq-Gen (Rambaut and Grass 1997), implemented in Dendropy, was used for alignment-generation under the GTR + Γ model. The base compositional frequency, the alpha parameter for gamma distribution, and the GTR substitution rate were also selected randomly from the density distribution of the Jarvis et al. (2014) empirical data of exons, introns, and UCEs, respectively. Simulations were repeated once for each data type to validate results, so there were six sets of 500 simulated genes in total. Using these datasets, we re-estimated gene trees in IQTREE using GTR + Γ model with and without bipartition constraints (i.e., ML gene trees). In total, 184 bipartition constraints were generated from 28 Jarvis species-trees representing the variety of potential hypotheses as to bird relationships from the different data types, so that all the bipartitions in the true tree were included. The edge-based analyses were conducted following the above-method on the three difficult focal edges that exhibit higher level of ILS as well as some commonly accepted edges that exhibit lower levels of ILS, including the core landbirds (**I**), the core waterbirds (**II**), the tropicbirds + sunbittern (i.e., Phaethontimorphae, **III**), and the nightjar + swifts + hummingbirds (i.e., Caprimulgimorphae, **V**) from the “magnificent seven” superordinal groups (Reddy et al. 2017). At this point, we also conducted ASTRAL analyses (hereafter MLAstral) using the re-estimated ML gene trees. We also calculated the mean gene-tree estimation error (GTEE) by conducting a normalized RF distance between the true and estimated gene trees, averaged across all genes, using Dendropy. Moreover, in order to also compare the performance of concatenation analyses, we concatenated the simulated alignments

of each data type and conducted ML analyses in RAxML 8.2.11 using GTR + Γ model that partitioned by locus, and with node support assessed by 200 bootstrap replicates.

2.5 Comparison of tree topologies within genetic regions

Within-genetic region comparisons were conducted using exons and UCEs from Jarvis et al. (2014), which have high alignment quality without long indels, in addition to the Prum et al. (2015) dataset. We split the genes into 1000bp chunks, omitting the short leftovers. We conducted ML analyses in IQTREE based on GTR + Γ model with uncertainty measured using both SH-aLRT (Guindon et al. 2010) and UFBS (Minh et al. 2013), each with 1000 replicates. Edges with < 80% SH-aLRT or < 95% UFBS support were collapsed. We note that this requirement is extremely conservative but limits false positives. We also re-estimated gene trees using these same criteria for consistency among these tests. We then compared trees from gene region fragments to the ML gene trees based on the entire gene region and noted conflicts.

3. Results and Discussion

3.1 Gene informativeness

To increase phylogenetic signal and reduce noise, we carefully curated data before analyses. For genes from Jarvis et al. (2014), 246 introns and 924 exons were excluded due to having a short alignment (< 500bp) and/or a large number of missing taxa (< 42 species, Table S1). After removing genes with potential systematic error (non-monophyletic outgroups and extremely long terminal branches), we obtained 5062 exons, 1409 introns and 3370 UCEs (Table S1). All data from Prum et al. (2015) passed the filtering processes and were therefore retained.

We examined the extent to which each genomic partition had information to resolve nodes within the tree. We found that each gene in the Jarvis et al. (2014) dataset supported, on average, 24 ($\geq 70\%$ UFBS) or 15 ($\geq 95\%$ UFBS, Minh et al. 2013) of the 47 internal nodes, respectively, whereas the Prum et al. (2015) data supported ($\geq 70\%$ BS, Hillis and Bull 1993) ~110 out of 197 internal nodes on each gene tree, with a maximum of 144 and minimum of 22 (Fig. 1A). Although the low support on many nodes can have many causes, it is likely that there was relatively low phylogenetic signal for individual genes that are subject to high gene-tree estimation error (GTEE, Molloy and Warnow 2018). In fact, this is not unexpected as single genes are unlikely to be informative across such broad relationships. Although introns generally

supported more relationships, the number of supporting nodes appeared to correspond to alignment length (Fig. S1). For the three difficult edges of interest, only 23%–33% exons, 38%–51% introns, and 34%–43% UCEs in the Jarvis et al. (2014) data obtained $\geq 70\%$ UFBS on ML gene trees after excluding genes that were missing owls or Hoatzin. Similarly, when we used the 70% BS cutoff for Prum et al. (2015) data, we only found four loci obtained supports on their ML gene trees for the resolution of Hoatzin, two genes showed supports for the earliest diverging clade of Neoaves, whereas no gene supported the position of owls. We also identified 113 genes in the Prum et al. (2015) dataset that showed non-monophyly of owls (i.e., the two families of owl had distinct positions in the tree). Taken as a whole, these results suggest that most loci have limited information for resolving the contentious edges and therefore a high potential for GTEE from these genome data, regardless of the sources of conflict.

3.2 Diversity of resolutions

Many gene regions across datasets and datatypes did not demonstrate a strong preference for any specific resolution regarding the three focal relationships (Fig. S2). However, to ascertain the nature of the support for alternative resolutions when support was indicated, we conducted several analyses. First, we examined how many different resolutions were supported by different genes. We observed, for example in the Jarvis et al. (2014) exons, there were more than 1500 alternative resolutions for each focal relationship. Most of these were unsupported ($\sim$ five resolutions obtained $\geq 95\%$ UFBS and ~ 40 resolutions obtained $\geq 50\%$ UFBS) or were present in fewer than 10 gene trees (i.e., rare resolutions, Fig. 1B). Even the best-supported topology for each data type was still present in less than 5% of gene trees. The situation was similar for introns and UCEs, and not limited to the Jarvis et al. (2014) dataset (Fig. S3A). The Prum et al. (2015) dataset had 259 loci, and we identified 110 or more topologies for each focal edge, but very few of those topologies had high ($> 70\%$) bootstrap support and most were found in a single gene (Fig. S3B).

The uncertainty of focal relationships has also been confirmed by the extremely low and negative ICA values on the species tree. For example, the TENT_ExaML tree in Jarvis et al. (2014) showed Phoenicopterimorphae (VII: flamingos and grebes) + Columbimorphae (VI: doves, sandgrouse and mesites) as the earliest diverging clade within Neoaves (Fig. 2A). Analyses of the non-coding (intronic and UCE) data in Jarvis et al. (2014) and of indels in Houde

et al. (2019) strongly suggest that one or both of the two groups are sister to all other Neoaves. However, this clustering obtained -0.051 (intron), -0.272 (UCE), and -0.317 (exon) ICA values, indicating substantial conflicting signals among gene trees, and a better supported alternative conflicting relationship as shown in the ML topology distribution (Fig. S3). The very different ICA values are also indicative of a data type effect, with introns showing a very different signal than UCEs or exons. Similar situations were also found for the other two focal edges (i.e., for the Hoatzin and owls) and in both datasets (not shown).

The diversity of resolutions can also be observed by the RF distances among gene trees. Trees based on the exons and UCEs from Jarvis et al. (2014) exhibited more among-gene tree differences (~65, 69% of the edges, Fig. S4) than do trees based on introns (~58, 62% of the edges), implying the potential of high GTEE in these gene trees (Simmons et al. 2016). Similar patterns were also found when comparing gene trees to the species-tree topologies. The RF distances among gene trees for the Prum et al. (2015) data were usually larger than 150 (38% of the edges) and as high as 300 (76% of the edges, Fig. S4), suggesting high topological diversity in these data; the larger number of topologies reflects the denser taxon sampling of the Prum et al. (2015) study.

3.3 Edge analyses

Examining conflicting signals in large-scale datasets is perhaps the foremost challenge in the field of systematics now that we have entered the phylogenomic era. One analytical framework that is commonly used is concatenation (e.g., Jarvis et al. 2014; Prum et al. 2015). Although concatenated analyses have been criticized for their failure to model the multispecies coalescent (Edwards et al. 2016) it is still a useful methodology in many contexts (Bryant and Hahn 2020). However, the utility of concatenated analyses for the examination of conflict is limited because they generally yield a single tree. Low support values in analyses of concatenated data can provide evidence that conflict might exist, but low support can also reflect the absence of any signal in the data (e.g., due to the existence of a polytomy). Fundamentally, there are only two ways to examine conflict using the concatenated approach. First, one can subdivide data based on some criterion, such as coding vs. non-coding data (Jarvis et al. 2014; Reddy et al. 2017) or protein structure (e.g., Pandey and Braun 2020), and conduct concatenated tree searches on the data subsets. Second, one can examine differences in the tree selection criterion (typically the

likelihood) for various topologies at the level of the genes (Shen et al. 2017) or the individual site level (Kimball et al. 2013). Both approaches have limitations; the first approach only allows one to assess data-type effects whereas the second approach can only yield information about a small number of topologies due to the limited availability of computational time.

An alternative way to examine conflicts involves comparisons of gene trees; in fact, comparisons among gene trees represent an obvious way to assess conflict because individual gene trees can have any topology. However, analyses of individual genes are subject to high GTEE, so many bipartitions in estimated gene trees may be spurious. Moreover, there can be conflicts within gene regions due to intralocus recombination and other factors (see below, section 3.6). Ultimately, it is difficult to determine whether the extensive conflict typically observed when estimated gene trees are compared reflects a mixture of meaningful biological signals (regardless of whether those signals are genuine conflicts among gene trees or biases due to some other factor) or simply random resolutions due to GTEE. We have chosen to use a method that lies somewhere between concatenation and simple gene-tree analysis: edge-based analyses of genes. This approach limits the consideration of gene-tree topologies to those containing a specific bipartition. In this study we have limited the bipartitions under consideration to those present either in published analyses that were identified using concatenated analyses (see Table 1) or those present in large numbers of gene trees. We used two complementary criteria to examine the results of these edge-based analyses: $\sum \Delta \ln L$ and $\sum N_g$. Comparison of the results given these criteria can provide evidence for outlier genes; when $\sum \Delta \ln L$ for a specific edge is much larger than $\sum N_g$ it implies that a relatively small number of genes are contributing a large amount of the $\ln L$ differences among alternative resolutions. Overall, this edge-based approach should limit GTEE while still retaining the advantages of gene tree-based analyses and provide information about conflicting signals that might be present in the phylogenetic datasets.

We examined gene-wide support for alternative resolutions of the three focal edges using the filtered Jarvis et al. (2014) data (Table S1). We primarily focused on the dominant alternative resolutions that have been previously published (Table 1). However, we found that the ML gene trees typically exhibited topologies different from published dominant relationships (Figs. S2 and S3). Thus, for each edge, we determined whether to exclude a gene by examining whether its corresponding node on the ML tree is strong conflict with all the testing relationships using 95%

UFBS (Minh et al. 2013) threshold (see above, in section 2.3 of the Methods, for additional details). After examining the compatibility of each ML gene tree with our test hypotheses, we obtained 4194 – 5019 exons, 1134 – 1274 introns, and 3097 – 3295 UCEs for each focal relationship (Table S1).

When we tested the set of five published hypotheses regarding the earliest diverging Neoaves (Fig. 3A and Table 1) we found that all data types had the highest support for placing doves sister to all other Neoaves (Figs 2B and S2). When we expanded the set of hypotheses regarding the earliest diverging Neoaves to include the most prevalent topologies in the gene trees (Fig. 3A, hypotheses 6–10) we found that support was more evenly distributed across the hypotheses (Fig. 3A). For the other two edges, we found that Hoatzin were sister to shorebirds (*S*), and owls were sister to mousebirds (Figs. 3B and 3C). Previous studies (e.g., Hackett et al. 2008; Jarvis et al. 2014) have seldom provided support for the best supported hypotheses in this study. Exceptions include owls + mousebirds (*M*), found in a reanalysis of shallow-filtered Jarvis UCE data (Gilbert et al. 2018), doves sister to all other Neoaves in the UCE indel tree (Houde et al. 2019), and Hoatzin + *S* in the complete UCE data matrix (McCormack et al. 2013). Moreover, the relationships published in the Jarvis et al. (2014) TENT tree (i.e., **VI+VII** as earliest Neoaves, *S*+*CR* [shorebirds + cranes and rails] sister to Hoatzin, and *M*+*W* [mousebirds + woodpeckers and allies] sister to owls) were only weakly supported (Figs. 2B and 3B and 3C).

To control for the impact of taxon sampling on topological preference, we conducted similar analyses using the Prum et al. (2015) data, which include 198 bird species. To reduce the CPU time necessary for analyses of the Prum data we only used the hypotheses from Prum et al. (2015) (i.e., **V** as earliest Neoavian branch, **I** sister to Hoatzin, and *M*+*W* sister to owls) and the top two published hypotheses supported by the Jarvis et al. (2014) data (i.e., **VII** and doves sister to all other Neoaves, *S* and **V** sister to Hoatzin, and *M* and *EHV* sister to owls). After examining the compatibility of the ML gene tree with the testing constraints based on 70% BS (Hillis and Bull 1993), we obtained 231 – 258 loci in the edge-based analyses for each focal edge. The best relationships supported by the Prum data were also supported in our analyses of the Jarvis data (e.g., Hoatzin + *S*; owls + *M*, Table S2), indicating a limited taxon-sampling effect on edge-based analyses. Taken as a whole, edge-based analyses increase congruence between analyses of the Prum and Jarvis data. Moreover, analyses focused on the earliest diverging Neoaves using the Prum data revealed a major impact of genes with a large $\Delta \ln L$. For example, we noted that

gene L111 encoding the FH2 domain containing 1 (FHDC1, whose homolog may play a role in both cell shape and microtubule polymerization in humans), disproportionately contributed $\Delta\ln L$ to support **VII** as the earliest divergence of Neoaves. We did not detect alignment error or lineage specific positive selection for this gene. Moreover, eliminating FHDC1 would not shift the best supported hypothesis given the $\sum\Delta\ln L$ criterion (Table S2). Regardless of the details, the best-supported hypothesis in these edge-based analyses of the Prum data were never the relationships present in the Prum et al. (2015) tree; they were hypotheses found in edge-based analyses of the Jarvis data.

3.4 Complex data type effects revealed by categorical edge-based analyses

The relatively even distribution of support for various hypotheses regarding the earliest diverging Neoaves when we examined 10 alternatives (Fig. 3A) suggests a very high degree of conflict among gene trees. However, this result also suggests a hypothesis-sampling-effect (i.e., the more hypotheses tested, the more widely distributed the support). It also provides evidence for a data type effect, where identified that the best hypotheses varied among datatypes; specifically, exons supported cuckoos sister, introns continued to support doves sister, and UCEs supported either clade **VII** (flamingos + grebes; best supported using the $\sum\Delta\ln L$ criterion) or cuckoos (best supported using the $\sum Ng$ criterion). It is noteworthy that Jarvis et al. (2014) placed **VII** sister to all other Neoaves in their concatenated analysis of UCEs and the $\sum\Delta\ln L$ criterion, as the sum of the differences in log-likelihood, is similar to the criterion used for tree selection in ML analyses. However, the observation that UCEs supported cuckoos sister when we used our alternative criterion, $\sum Ng$ (the number of supporting genes), suggests support for **VII** sister in analyses of UCEs reflects a limited number of UCE loci with large $\Delta\ln L$ values. Indeed, support given the UCE data and the $\sum\Delta\ln L$ criterion was highest for the **VII** sister hypothesis with 16.5% of the total $\Delta\ln L$ whereas the $\sum Ng$ criterion placed **VII** sister in fifth place with 12.6% of loci. There were other differences among data types in the rank order of suboptimal hypotheses. For example, when the $\sum\Delta\ln L$ criterion is used **VI** + **VII** as the earliest Neoavian branch received much higher support in analyses of non-coding data, especially introns, than it did in analyses of coding data (10.9% for introns, 7.4% for UCEs, and 2% for exons). These results agree with the data type effects reported by Reddy et al. (2017), although the differences among data types also appear to be more complex than the results of that study suggested.

Reddy et al. (2017) pointed out that data type had a major impact on the topology for the deepest divergence in Neoaves, with analyses of coding data supporting a **V** sister topology and analyses of non-coding data supporting either **VI+VII** sister (introns) or **VII** sister (UCEs). Both Jarvis et al. (2014) and Reddy et al. (2017) pointed out that avian coding data typically exhibit greater GC_{CV} than do non-coding data. Previous studies (e.g., Foster 2004; Jermini et al. 2004) have demonstrated that high GC_{CV} , when fit to a homogeneous model, can result in biased phylogenetic estimation. Applying models with more heterogeneous assumptions (e.g., Galtier and Gouy 1998; Holland et al. 2013) has the potential to improve estimates of phylogeny, although the available programs that implement these models are challenging to apply to phylogenomic datasets (see additional discussion in Reddy et al. 2017). Since our edge-based analyses provide a sensitive tool to examine conflicting signals in phylogenomic datasets we assessed the impact of GC_{CV} in the Jarvis et al. (2014) data on the conflicts among data types that we observed.

Here, we examined parsimony-informative sites (as in Reddy et al. 2017 and Braun et al. 2019) using three schemes, which we called ΔGC_{50} , ΔGC_{90} , and ΔGC_{100} . These values reflect the difference between taxa with different GC contents (i.e., ΔGC_{90} is the GC difference between the 95 and 5 percentile GC-rich taxa). ΔGC_{50} was similar among data types but ΔGC_{90} and ΔGC_{100} variation was larger for exons than for introns and UCEs, consistent with findings in Reddy et al. (2017) (Fig. 4a). We ordered genes based on ΔGC_{90} to maximize differences while eliminating outlier signals (Fig. 4a) and compared the edge analyses results between genes from either the upper (i.e., high GC_{CV}) or the lower (e.g., low GC_{CV}) quartiles of the variation. This results in about 1266 exons, 352 introns and 842 UCEs in each category. If GC_{CV} has an impact on inference we would expect different patterns of support between GC_{CV} categories within the same data type. Indeed, owls + *EHV* (eagles, hawks and vultures) showed much higher support (both $\sum Ng$ and $\sum \Delta \ln L$) in all high GC_{CV} markers (Figs. 4b and S5), and such signal is strongest in EH and IH. It seems the characteristics of high GC-variation genes tend to correlate with owls + *EHV*. Braun et al. (2019) suggested that the position of owls in concatenated analyses was subject to a data type effect, with analyses of coding data having a tendency to yield the owls + *EHV* topology. We found that support for owls + *EHV* was similar for the EL, IL, and UL datasets (low GC_{CV} exons, introns, and UCEs, respectively) but much higher in the EH, IH, and UH (the high GC_{CV} datasets). However, the signal for owls + *EHV* was still stronger in the high

GC_{CV} exons than in the high GC_{CV} introns and UCEs. Thus, our results imply that an interaction between higher GC_{CV} and data type— not coding vs non-coding data type *per se* — is correlated with the owls + *EHV* signal.

As previously noted, the position of Hoatzin has been highly variable across studies and data type effects have not been noted in prior studies (reviewed by Braun et al. 2019). The dominant signal in our categorical edge-based analyses was Hoatzin + *S* (shorebirds), which is also present in the Jarvis exon c123 tree and the McCormack et al. (2013) UCE tree. In the categorical edge-based analyses, there was higher support for Hoatzin + *S* in high GC_{CV} markers (compare EH vs EL, IH vs IL, and UH vs UL; Fig. 4B), but the support differences are quite trivial. In fact, support for alternative relationships of the Hoatzin show more variations among data types (Figs. 4B and S4) regardless of GC_{CV} and tree length. The most notable difference was a slightly lower signal supporting Hoatzin + *S* and a higher signal for Hoatzin + **V** (Caprimulgimorphae) in the introns (Fig 4B and 4D). Provocatively, Hoatzin + **V** is present in the Jarvis intron tree (albeit with only 70% bootstrap support) but it is not in the Jarvis UCE tree (which included Hoatzin + *S* + *CR*, like the Jarvis total evidence nucleotide tree). These results are not consistent with a simple coding vs noncoding effect, as (1) the signal for Hoatzin + *S* is very similar in exons and UCEs (Fig. 4B) whereas the Hoatzin + **V** relationship was elevated in both noncoding data types, and (2) Hoatzin + **I** (core landbirds) — which was found in the Prum et al. (2015) study that largely used coding sequences — received higher support in introns and UCEs (Figs. 4B and S5) than it did in Jarvis' exons (Fig. 4). Thus, in agreement with many previous studies, this study finds that the position of the Hoatzin is very difficult to resolve. However, our edge-based analyses revealed a previously unappreciated whereas potentially more complex data type effect and further showed that Hoatzin + *S* was the strongest signal in all three data types.

The most prominent data type effect hypothesized by Reddy et al. (2017) was the **VI** + **VII** sister to all other Neoaves topology (their non-coding topology) and **V** sister to all other Neoaves (their coding hypothesis). Using our categorical edge-based analyses, **V** as sister to all other Neoaves did not appear to have better support in exons than in introns or UCEs (Figs 4B and S5). The data subset with the highest support for **V** as the earliest neoavian branch was the high GC_{CV} UCEs (UH in Fig 4B). This support for **V** sister appears to result from one outlier gene (chr1_28895_s) that contributes almost 40% of the $\sum \Delta \ln L$ (Table S3) for this relationship.

Unsurprisingly, given the evidence for an outlier gene in the UH dataset, the total number of genes ($\sum Ng$) that support **V** did not vary much among GC_{CV} categories or data types (Fig. S5). As described above **VI+VII** did appear to be better supported in introns and UCEs, especially in the data subsets with high GC_{CV} and high tree lengths (i.e., IH and UH, Figs 4B and 4D). However, the number of genes ($\sum Ng$) supporting **VI+VII** did not vary much among categories and data types (Fig. S5). Thus, the higher support for **VI+VII** sister appears to reflect a relatively small number of genes that make high ΔlnL contributions; given the much higher proportion of signal for **VI+VII** sister in the IH and UH subsets (Fig. 4B) these genes with a high ΔlnL contribution must be especially common in the high GC_{CV} data subsets.

Evolutionary rate is another major factor in phylogenetic resolution (Chojnowski et al. 2008). Tree length (sum of branch lengths) variation suggests that different overall rates of evolution and different data types exhibited different patterns, with introns having larger tree length/internode lengths and less variation than UCEs and exons (Fig. 4C). Tree length in these datasets is primarily the result of tip lengths as internodes were very short across data types (Fig. 4C) as a consequence of sampling at high taxonomic levels. We sorted genes based on their total tree length and compared support for focal edges with the upper and lower quartile sets. In general, tree lengths did not have an obvious impact on support for different hypotheses for each focal edge. However, we noticed two interesting patterns with respect to the earliest diverging Neoaves (Fig. 4D). First, **VI+VII** as sister to all other Neoaves usually received higher support in categories with high GC_{CV} and high tree lengths, but we noted that **VI+VII** also received high support (i.e., the best) in the intron subset of low tree lengths (IL, Fig. 4D). Within this category, we identified one gene (e.g., 11117.intron, Table S3) contributing ~21% of the $\sum \Delta lnL$ supporting this relationship. We did not detect any alignment issue in this gene so it is possible that the biological characteristics of this gene drove an outlier signal to support this relationship, which may not be directly related to evolutionary rate as the tree length of this gene is around the upper edge of this quartile set. Second, **VII** as sister to all other Neoaves received the highest support from the UCE subset of high tree length (UH, Fig. 4D), consistent with the Jarvis et al. (2014) UCE tree. Again, an outlier gene (chr8_1881_s, that contributed 1/3 (666 ΔlnL) to the $\sum \Delta lnL$, Table S3) was identified, and its support for **VII** as sister may be driven by certain characteristics that may or may not be directly related to its high total tree length.

As found previously, genes that have disproportionally high support for specific relationships (i.e., outlier genes) can bias phylogenetic reconstruction (Kimball et al. 2013; Shen et al. 2017, Walker et al. 2018). We noted several potential outliers within these datasets (Table S3). As indicated above, some outliers can significantly support certain hypotheses. However, removal of most single outlier genes did not change the results significantly nor could we detect any alignment errors with these genes. The diversity of alternative hypotheses noted earlier made it difficult to detect the underlying reasons for outliers biased support. One biological process that might cause biased phylogenetic resolution is selection. However, we could not test selection as most exons contained premature stop codons (e.g., single mutation likely caused by sequencing or assembly error). Both introns and UCEs also contain outliers. Although it was impossible to link gene specific traits to the topology they support without annotation, our categorical edge-based analyses indeed implied that the earliest diverging taxa in Neoaves might be influenced by outlier signal. In contrast to the case for the earliest diverging Neoaves, we did not detect as strong of an impact from outlier genes on the other two focal edges (the positions of Hoatzin and owls).

3.5 Simulations

Most of the gene-tree topologies resulted in a single species/lineage as sister to the remaining lineages (Figs. 1B and S3). To determine if this may be expected given that short internodes should have a higher prevalence of ILS, we conducted coalescent simulations using the *TENT.MP-EST.binned* tree from Jarvis et al. (2014). This tree has a clade **VI+VII** as the earliest diverged Neoaves, the Hoatzin groups with a large clade including **II+III**+ *S*+*CR*, and owls are sister to *EHV*. Simulations were conducted twice, with parameters drawn from the same parameter distribution for each hypothetical data type (see details in method, Table 2). We found the overall level of ILS to be relatively high (~58% ARFD between the simulated true gene tree and the true species tree [*TENT.MP-EST.binned* tree]). The GTEE appears to be high as well, showing ~68% normalized RF distances between the true gene tree and the estimated ML gene tree, which implies low phylogenetic signal per gene for reconstructing this rapid Neoavian radiation. Even though coalescent species-trees reconstructed with ASTRAL III using simulated data did recover > 90% of the true relationships across the tree, the concatenation analyses only recovered ~85% of the true relationships (Table S4). However, the positions of the Hoatzin and

the earliest diverging neoavian lineage differed from the true topologies 61% and 44% of the time, respectively, possibly due to the high GTEE and ILS.

After examining the estimated ML gene-trees, we found that the most common topologies for each focal edge were likely to have a single species/lineages sister to the remaining clades (Fig. S6), implying that ILS may be one source of this preference. There were ~20% estimated ML gene-trees showing non-monophyletic Palaeognathae and Galloanserae. The coalescent branch length on the true species tree (*TENT.MP-EST.binned* tree) is long (2.03) and even longer using indel analyses (2.57, Houde et al. 2020), so the expected number of deep coalescences that separate Galloanserae and Palaeognathae should be ~5% – 9%. Under this condition, the high occurrence of non-monophyletic Palaeognathae and Galloanserae in the simulated ML gene-trees is presumably caused by GTEE. We conducted edge-based analyses with the 184 bipartitions from the 28 Jarvis et al. (2014) species-trees representing the extent of the variation of potential species-tree hypotheses. Among these bipartitions, there were 63 (earliest Neoaves), 45 (Hoatzin sister), and 14 (owl sister) conflicting hypotheses for each focal edge respectively (Table 2). The best supported hypotheses usually differed from the true relationships, which is possibly caused by high GTEE and ILS (Table 2). For example, for the earliest diverging Neoaves, 25 and 28 hypotheses obtained some support from the two sets of simulated exons, respectively. Among them, the true topology (i.e., **VI+VII** as earliest Neoaves) is only supported by one or five genes and ranked as the 14th or 18th among the resolutions. The situation was similar for simulated introns and UCEs (Table 2) and for the other two focal edges: the true topology ranks 10th – 17th among the resolutions of the position of the Hoatzin, and ranks second for the position of owls (Table 2) across all data types.

In contrast to the focal edges that exhibited extremely short internodes and dramatic gene tree conflicts, we also examined less contentious edges with longer internodes (i.e., those with a low level of ILS). For example, clades including the core landbirds (**I**), the core waterbirds (**II**), the Caprimulgimorphae (**V**), and the tropicbirds + sunbittern (**III**), have been consistently recovered by previous studies (e.g., Reddy et al. 2017), and are connected by long internal branches (stems) in the true tree. These received high support across the simulations regardless of data type (Table 2), which suggests that the edge-based analyses were able to recover the correct relationships when there was enough phylogenetic signal and weak ILS. These analyses

suggest that strong ILS and high GTEE remain a difficult problem, and may result in biased resolutions for gene trees.

3.6 Within gene region conflict

To determine whether gene tree conflict was the result of conflicting signal within genomic regions, we further examined phylogenetic signal across regions within each gene. The average gene-region length of the Jarvis et al. (2014) was 1998 bp (501 – 15819 bp) for exons, 2509 bp (1787 – 3527 bp) for UCEs, and 9309 bp (506 – 66,762 bp) for introns. For Prum et al. (2015), the average gene region length was 1524 bp (361 – 2316 bp). We focused on exon and UCE alignments as they were smaller and therefore less likely than introns to suffer from mixed signal. We conducted analyses on each gene region to determine whether 1000 bp segments conflicted with the ML tree calculated for the entire region (see methods for detailed splitting strategy). We found extensive conflicting signal after collapsing edges with < 95% UFBS (Minh et al. 2013) and < 80% SH-aLRT (approximate likelihood-ratio test; Guindon et al. 2010) support. Although this is an extremely conservative measure, and likely to have a high false-negative rate (Guindon et al. 2010), it was chosen to prevent false positives. For loci with at least 2×1000 bp segments, we found that 10% of Jarvis et al. (2014) exons (195 of 1985), 6% (207 of 3367) of UCEs, and 50% (8 of 15) of gene regions in Prum et al. (2015) had conflicts. It is noteworthy that exons in Jarvis et al. (2014) are usually a combination of multiple exonic regions of one gene, with some regions separated by long introns. These regions might exhibit different evolutionary histories due to recombination, a different selection level, and/or for other reasons. Although intragenic conflicts within the Prum et al. (2015) data were much higher they could reflect sampling error due to the small number (i.e., only 15) of long gene regions in the data (the larger number of taxa in the Prum et al. 2015 dataset is also likely to have had an impact).

To determine whether within gene conflict corresponded to obvious data “dirtiness”, we examined alignment patterns within the conflicting gene regions. In general, some loci that exhibited substantial intragenic conflicts can still have alignments that appear largely unambiguous (e.g., neat alignments in locus 9838.exons, Fig. S7), which suggests there may be biological processes underlying these intragenic conflicts. However, other intragenic conflicts might be due to misalignment patterns. For example, the 1424.exon gene region from Jarvis et al. (2014) had an alignment that was obviously “messy” alignments (Fig. S7); analyses of

problematic segments placed Rifleman sister to pelicans, mesites sister to cranes, grebes sister to sandgrouse, and fulmars sister to seriemas.

While it is possible for these intragenic conflicts to be the result of biological processes (Scornavacca and Galtier 2017; Mendes et al. 2019), the most likely issue is data curation given the volume of both datasets (Springer and Gatesy 2018; Braun et al. 2019). All large datasets suffer from some amount of noisy data. However, because of the biological nature of the bird radiation with short intervals among many early speciation events, when the true phylogenetic signal is likely to be weak, the importance of data cleanliness is even more important.

4. Conclusions

The categorical edge-based analyses conducted herein provide a way to examine differences in signal associated with subsets of the data and assess the best-supported resolution of specific nodes. The best-corroborated resolutions of our focal nodes were: 1) owl + *M* (mousebirds); 2) Hoatzin + *S* (shorebirds); and 3) doves are the earliest diverging Neoaves. However, the edge-based analyses revealed substantial conflict; indeed, they provided evidence for a larger amount of conflict than the data-type analyses conducted by Jarvis et al. (2014) or Reddy et al. (2017). First, the owls + *EHV* signal reflects a combination of higher GC_{CV} , and data type rather than data type *per se*. Second, the position of Hoatzin is highly variable but there is a data type effect that was not evident in previous studies. Specifically, support for Hoatzin + *V* was strongest in introns, intermediate in UCEs, and weakest for exons. Finally, data type effects were evident for the earliest diverging Neoaves, as expected based on Reddy et al. (2017), but outlier genes also appeared to play an important role in this part of the tree. Perhaps surprisingly, there was not a single factor (e.g., GC_{CV} or data type) that played an important role in the resolution of all three focal nodes. Thus, simple solutions like focusing on a specific data type or eliminating genes with high GC_{CV} are unlikely to be a panacea for analyses of avian phylogeny (or, by extension, other nodes in the Tree of Life that have proven to be recalcitrant). We believe that our findings will facilitate further exploration of phylogenetic conflicts using an edge-by-edge approach.

Perhaps one of the most surprising results of this study is the fact that the results of some of our edge-based analysis contradict the magnificent seven – the superordinal groups that have been consistently identified on the Neoavian tree by recent large-scale studies (Reddy et al. 2017). The relatively consistent recovery of these groups in large-scale analyses has led to the

view that most or all of these clades are present in the true avian species tree (Suh 2016; Braun et al. 2019). However, edge-based analyses indicate that at least some of those groups might have less support than we currently believe. For example, signal placing either cuckoos or doves sister to all other Neoaves is stronger than many other alternatives (Figs. 3 and 4). Both signals contradict magnificent seven clades; specifically, doves as the earliest divergence contradicts Columbimorphae (group VI; doves + mesites + sandgrouse; Fig. 2A) and placing cuckoos sister to all other Neoaves contradicts the Otidimorphae (group V; cuckoos + turacos + bustards; Fig. 2A). There are several interpretations of this result. First, both magnificent seven clades could be present in the true tree but our edge-based analyses reflect the fact that the branches uniting those clades are especially short and, therefore, “noisy.” Arguably, this is the simplest hypothesis given that previous studies (e.g., Jarvis et al. 2014; Prum et al. 2015) agree that the relevant edges are quite short. Second, the fact that signals contradicting those magnificent seven clades are stronger than signals consistent with those clades could provide evidence that the Columbimorphae and Otidimorphae are present but there is something unusual about the branches uniting those clades. In fact, Houde et al. (2020) examined indel data and they reported that conflicts among indels for Columbimorphae were higher than expected given the relevant edge length in the Jarvis et al. (2014) timetree; they interpreted this as evidence for a transient increase in the effective population size for the common ancestor of that clade. Finally, our edge-based analyses can be viewed as evidence that Columbimorphae (if the doves sister hypothesis is correct) or Otidimorphae (if the doves sister hypothesis is correct) are not present as clades in the true bird tree. Regardless of the best interpretation of the edge-based analyses, it is clear that they provide a way to highlight especially problematic branches that warrant further examination in future phylogenomic studies.

Given the evidence for conflict evident in our edge-based analyses we are left with a fundamental question: what is the best-way forward for resolving the nodes in the bird tree that remain uncertain? Obviously, collecting even more data will be important and efforts like the B10K project (Zhang 2015; Stiller and Zhang 2019) and the OpenWings (Kimball et al. 2019) will provide these larger datasets for birds. However, there is ample evidence that simply expanding the size of phylogenomic datasets will not solve the problem of recalcitrant nodes in avian phylogeny (or the difficult nodes in other parts of the tree of life). Rare genomic changes (Rokas and Holland 2000) may provide an alternative data type that sidesteps the complexities

evident in the analyses we present here. In fact, rare genomic changes can avoid (or at least reduce) problems like intralocus recombination, natural selection, and GTEE (Springer and Gatesy 2019). The availability of genome-scale datasets will only increase availability of rare genomic change data and allow the community to move beyond retroposon insertions, which were used even before the phylogenomic era (e.g., Suh et al. 2011), to types of rare genomic changes like numts (Liang et al. 2018) and microinversions (Braun et al. 2011) that have not been used in many phylogenetic studies. These types of data, which are very different from nucleotide sequence data, have the potential to corroborate (or falsify) the hypothesis that the most common resolution in edge-based analyses represents a correct resolution of the species tree. Regardless of whether the hypothesis with the greatest support (whether judged using $\sum \Delta \ln L$ and $\sum N_g$) is correct, it is clear that the edge-based analyses reveal substantial conflicts within the data. Understanding those conflicts is an important problem in evolutionary biology even if we do not know the true resolution of the tree. We can learn even more if the information about conflicts among genes, like those revealed by our edge-based analyses, can be combined with an accurate estimate of the species tree.

Supplementary material

Supplementary figures and tables can be found online.

Acknowledgements

We would like to thank Benjamin Winger, Joseph Walker, Nathanael Walker-Hale and Diego Alvarado-Serrano for their constructive comments on the earliest version of the paper. We appreciate other comments from members of the Smith lab. This work was supported by grants from the US National Science Foundation (DBI-1458466 and DEB-1207915 to S.A.S., DEB-1241066 and DEB-1146423 to J.C., and DEB-1655683 to E.L.B.) and by the Frank M. Chapman Memorial Fund of the American Museum of Natural History.

Author contributions

Study conception and design: N.W., S.A.S.; Acquisition of data: N.W.; Analysis and interpretation of data: N.W., S.A.S., E.L.B., J.C., B.L.; Drafting of manuscript: N.W., E.L.B., J.C., S.A.S.; Image illustration: B.L., N.W., E.L.B., J.C.

Competing interests

The authors have no competing interests declared.

References

- Bendall ML, Stevens SL, Chan LK, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, Froula J. 2016. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME* 10(7):1589.
- Braun EL, Kimball RT, Han KL, Iuhasz-Velez NR, Bonilla AJ, Chojnowski JL, Smith JV, Bowie RC, Braun MJ, Hackett SJ, Harshman J. 2011. Homoplastic microinversions and the avian tree of life. *BMC Evol Biol.* 11(1):141.
- Braun EL, Cracraft J, Houde P. 2019. Resolving the avian tree of life from top to bottom: the promise and potential boundaries of the phylogenomic era. In *Avian Genomics in Ecology and Evolution* (pp. 151-210). Springer, Cham.
- Bryant D, Hahn MW. 2020. The Concatenation Question. Scornavacca, Celine; Delsuc, Frédéric; Galtier, Nicolas. *Phylogenetics in the Genomic Era*, No commercial publisher | Authors open access book, pp.3.4:1–3.4:23. hal-02535651f.
- Chen D, Braun EL, Forthman M, Kimball RT, Zhang Z. 2018. A simple strategy for recovering ultraconserved elements, exons, and introns from low coverage shotgun sequencing of museum specimens: Placement of the partridge genus *Tropicoperdix* within the galliformes. *Mol Phylogenet Evol.* 129:304–314.
- Chojnowski JL, Kimball RT, Braun EL, 2008. Introns outperform exons in analyses of basal avian phylogeny using clathrin heavy chain genes. *Gene* 410(1):89–96.
- Dominguez-Bello MG, Michelangeli F, Ruiz MC, Garcia A, Rodriguez E. 1994. Ecology of the folivorous hoatzin (*Opisthocomus Hoatzin*) on the Venezuelan plains. *Auk* 111(3):643–651.
- Duchêne DA, Tong KJ, Foster CS, Duchene S, Lanfear R, Ho SY. 2018. Linking Branch Lengths Across Loci Provides the Best Fit for Phylogenetic Inference. *bioRxiv* 467449.
- Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63(1):1-19.

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

798 Ericson PG, Anderson CL, Britton T, Elzanowski A, Johansson US, Källersjö M, Ohlson JI, Parsons TJ,
799 Zuccon D, Mayr G. 2006. Diversification of Neoaves: integration of molecular sequence data and
800 fossils. *Biol Lett.* 2(4):543–547.

801 Foster PG. 2004. Modeling compositional heterogeneity. *Syst Biol.* 53:485–495.

802 Foster PG, Hickey DA. 1999. Compositional bias may affect both DNA-based and protein-based
803 phylogenetic reconstructions. *J Mol Evol.* 48:284–290.

804 Galtier N, Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a
805 nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol.*
806 15(7): 871-879.

807 Gatesy J, Springer MS. 2017. Phylogenomic red flags: Homology errors and zombie lineages in the
808 evolutionary diversification of placental mammals. *Proc Natl Acad Sci USA.* 114(45):E9431–
809 E9432.

810 Gilbert PS, Wu J, Simon MW, Sinsheimer JS, Alfaro ME. 2018. Filtering nucleotide sites by
811 phylogenetic signal to noise ratio increases confidence in the Neoaves phylogeny generated from
812 ultraconserved elements. *Mol Phylogenet Evol.* 126:116–128.

813 Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and
814 methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML
815 3.0. *Syst Biol.* 59(3):307–321.

816 Hackett SJ, Kimball RT, Reddy S, Bowie RC, Braun EL, Braun MJ, Chojnowski JL, Cox WA, Han KL,
817 Harshman J, Huddleston CJ. et al. 2008. A phylogenomic study of birds reveals their evolutionary
818 history. *Science* 320(5884):1763–1768.

819 Hillis DM, Bull JJ. 1993. An empirical test of bootstrapping as a method for assessing confidence in
820 phylogenetic analysis. *Syst Biol.* 42(2):182–192.

821 Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018 UFBoot2: Improving the ultrafast
822 bootstrap approximation. *Mol Biol Evol.* 35:518–522.

823 Holland BR, Jarvis PD, Sumner JG. 2013. Low-parameter phylogenetic inference under the general
824 Markov model. *Syst Biol.* 62:78-92.

825 Hughes JM, Baker AJ. 1999. Phylogenetic relationships of the enigmatic hoatzin (*Opisthocomus Hoatzin*)
826 resolved using mitochondrial and nuclear gene sequences. *Mol Biol Evol.* 16(9):1300–1307.

827 Houde P, Braun EL, Narula N, Minjares U, Mirarab S. 2019. Phylogenetic signal of indels and the
828 neoavian radiation. *Diversity* 11(7):108.

829 Houde P, Braun EL, Zhou L. 2020. Deep-time demographic inference suggests ecological release as
830 driver of Neoavian adaptive radiation. *Diversity* 12:164.

831 Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, Suh
832 A. et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds.
833 *Science* 346(6215):1320–1331.

834 Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, Suh
835 A. 2015. Phylogenomic analyses data of the avian phylogenomics project. *GigaScience* 4(1):4.

836 Jermini LS, Ho SY, Ababneh F, Robinson J, Larkum AW. 2004. The biasing effect of compositional
837 heterogeneity on phylogenetic estimates may be underestimated. *Syst Biol.* 53(4):638–643.

838 Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. 2012. The global diversity of birds in space and
839 time. *Nature* 491(7424):444.

840 Kimball RT, Wang N, Heimer-McGinn V, Ferguson C, Braun EL. 2013. Identifying localized biases in
841 large datasets: a case study using the avian tree of life. *Mol Phylogenet Evol.* 69(3): 1021–1032.

842 Kimball RT, Oliveros CH, Wang N, White ND, Barker FK, Field DJ, Ksepka DT, Chesser RT, Moyle
843 RG, Braun MJ, Brumfield RT, Faircloth BC, Smith BT, Braun EL. 2019. A phylogenomic
844 supertree of birds. *Diversity* 11(7):109

845 Kobert K, Salichos L, Rokas A, Stamatakis A. 2016. Computing the internode certainty and related
846 measures from partial gene trees. *Mol Biol Evol.* 33(6):1606–1617.

847 Kolaczkowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics
848 when evolution is heterogeneous. *Nature* 431(7011):980.

849 Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under
850 coalescence. *Syst Biol.* 56(1):17–24.

851 Li Y, Liu Z, Shi P, Zhang J. 2010. The hearing gene Prestin unites echolocating bats and whales. *Curr*
852 *Biol.* 20(2):R55–R56

853 Liang B, Wang N, Li N, Kimball RT, Braun EL. 2018. Comparative genomics reveals a burst of
854 homoplasy-free numt insertions. *Mol Biol Evol.* 35(8):2060–2064.

855 Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees
856 under the coalescent model. *BMC Evol Biol.* 10(1):302.

857 Liu L, Zhang J, Rheindt FE, Lei F, Qu Y, Wang Y, Zhang Y, Sullivan C, Nie W, Wang J, Yang F. 2017.
858 Reply to Gatesy and Springer: Claims of homology errors and zombie lineages do not
859 compromise the dating of placental diversification. *Proc Natl Acad Sci USA.* 114(45):E9433–
860 E9434.

861 McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. 2012.
862 Ultraconserved elements are novel phylogenomic markers that resolve placental mammal
863 phylogeny when combined with species-tree analysis. *Genome Res.* 22(4):746–754.

864 McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. 2013. A phylogeny
865 of birds based on over 1,500 loci collected by target enrichment and high-throughput
866 sequencing. *PLoS One* 8(1):e54848.

867 Mendes FK, Livera AP, Hahn MW. 2019. The perils of intralocus recombination for inferences of
868 molecular convergence. *Philos Trans Royal Soc B.* 374(1777):20180244.

869 Minh BQ, Nguyen MA, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol*
870 *Biol Evol.* 30(5):1188–1195.

871 Mirarab S, Bayzid MS, Boussau B, Warnow T. 2014. Statistical binning enables an accurate coalescent-
872 based estimation of the avian tree. *Science* 346(6215).

873 Molloy EK, Warnow T. 2018. To include or not to include: the impact of gene filtering on species tree
874 estimation methods. *Syst Biol.* 67(2):285–303.

875 Morgan CC, Foster PG, Webb AE, Pisani D, McInerney JO, O’Connell MJ. 2013. Heterogeneous models
876 place the root of the placental mammal phylogeny. *Mol Biol Evol.* 30(9):2145–2156.

877 Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2014. IQ-TREE: a fast and effective stochastic
878 algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 32(1):268–274.

879 Pandey A, Braun EL. 2020. Phylogenetic Analyses of Sites in Different Protein Structural Environments
880 Result in Distinct Placements of the Metazoan Root. *Biology* 2020, 9:64.

881 Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2015. A
882 comprehensive phylogeny of birds (Aves) using targeted next-generation DNA
883 sequencing. *Nature* 526(7574):569.

884 Puttick MN, Morris JL, Williams TA, Cox CJ, Edwards D, Kenrick P, Pressel S, Wellman CH, Schneider
885 H, Pisani D, Donoghue PC. 2018. The interrelationships of land plants and the nature of the
886 ancestral embryophyte. *Curr Biol.* 28(5):733–745.

887 Rambaut A, Grass NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence
888 evolution along phylogenetic trees. *Bioinformatics* 13(3):235–238.

889 Reddy S, Kimball RT, Pandey A, Hosner PA, Braun MJ, Hackett SJ, Han KL, Harshman J, Huddleston
890 CJ, Kingston S, Marks BD. 2017. Why do phylogenomic data sets yield conflicting trees? Data
891 type influences the avian tree of life more than taxon sampling. *Syst Biol.* 66(5):857–879.

892 Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Math Biosci.* 53(1-2):131–147.

893 Rokas A, Holland PW. 2000. Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol.*
894 15(11):454–459.

895 Salichos L, Stamatakis A, Rokas A. 2014. Novel information theory-based measures for quantifying
896 incongruence among phylogenetic trees. *Mol Biol Evol.* 31:1261–1271.

897 Scornavacca C, Galtier N. 2017. Incomplete lineage sorting in mammalian phylogenomics. *Syst Biol.*
898 66(1):112–120.

899 Shen XX, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven
900 by a handful of genes. *Nat Ecol Evol.* 1(5): 0126.

901 Simmons M.P., Sloan D.B., Gatesy J. 2016. The effects of subsampling gene trees on coalescent methods
902 applied to ancient divergences. *Mol Phylogenet Evol.* 97:76–89.

903 Smith SA, Walker-Hale N, Walker JF, Brown J. 2020. Phylogenetic Conflicts, Combinability, and Deep
904 Phylogenomics in Plants. *Syst Biol.* 69(3):579–592.

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

906 Springer MS, Gatesy J. 2018. On the importance of homology in the age of phylogenomics. *System*
907 *Biodivers.* 16(3): 210-228.

908 Springer MS, Gatesy J 2019. Retroposon insertions within a multispecies coalescent framework suggest
909 that ratite phylogeny is not in the ‘Anomaly Zone’. *bioRxiv* p.643296.

910 Stamatakis A. 2006, April. Phylogenetic models of rate heterogeneity: a high performance computing
911 perspective. In Proceedings 20th IEEE international parallel & distributed processing symposium
912 Rhodes, Greece (pp. 8-pp). IEEE.

913 Stamatakis A. 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large
914 Phylogenies. *Bioinformatics* 30 (9):1312-1313.

915 Stern DL. 2013. The genetic causes of convergent evolution. *Nat Rev Genet.* 14(11):751-764.

916 Stiller J, Zhang G. 2019. Comparative Phylogenomics, a stepping stone for bird biodiversity studies.
917 *Diversity* 11: 115.

918 Suh A. 2016. The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. *Zool*
919 *Scr.* 45:50–62.

920 Suh A, Paus M, Kiefmann M, Churakov G, Franke FA, Brosius J, Kriegs JO, Schmitz J. 2011. Mesozoic
921 retroposons reveal parrots as the closest living relatives of passerine birds. *Nat Commun.* 2:443.

922 Suh A, Linnéa S, Hans E. 2015. The dynamics of incomplete lineage sorting across the ancient adaptive
923 radiation of neoavian birds. *PLoS Biol.* 13(8):e1002224.

924 Sukumaran J, Holder MT. 2010. DendroPy: A Python library for phylogenetic computing. *Bioinformatics*
925 26:1569–1571.

926 Tarver JE, Dos Reis M, Mirarab S, Moran RJ, Parker S, O’Reilly JE, King BL, O’Connell MJ, Asher RJ,
927 Warnow T, Peterson KJ. 2016. The interrelationships of placental mammals and the limits of
928 phylogenetic inference. *Genome Biol Evol.* 8(2):330–344.

929 Walker JF, Brown JW, Smith SA. 2018. Analyzing contentious relationships and outlier genes in
930 phylogenomics. *Syst Biol.* 67(5):916–924.

931 Wang N, Braun EL, Kimball RT. 2012. Testing hypotheses about the sister group of the Passeriformes
932 using an independent 30-locus data set. *Mol Biol Evol.* 29(2):737–750.

933 Wang N, Kimball RT, Braun EL, Liang B, Zhang Z. 2017. Ancestral range reconstruction of Galliformes:
934 the effects of topology and taxon sampling. *J Biogeogr.* 44(1):122–135.

935 Wang N, Yang Y, Moore MJ, Brockington SF, Walker JF, Brown JW, Liang B, Feng T, Edwards C,
936 Mikenas J, Olivieri J. et al. 2019. Evolution of Portulacineae marked by gene tree conflict and
937 gene family expansion associated with adaptation to harsh environments. *Mol Biol Evol.*
938 36(1):112–126.

939 Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS,
940 Burleigh JG, Gitzendanner MA, Ruhfel BR. et al. 2014. Phylotranscriptomic analysis of the
941 origin and early diversification of land plants. *Proc Natl Acad Sci USA.* 111(45):E4859–E4868.

942 Zhang GJ. 2015. Bird sequencing project takes off. *Nature* 522:34-44.

943 Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: Polynomial Time Species Tree
944 Reconstruction from Partially Resolved Gene Trees. *BMC Bioinformatics* 19(6):153.

945

Table 1. Three focal edges and major testing hypotheses that published before.

Focal Edges	Hypotheses	Major References
Earliest Neoaves	1. VI+VII	Jarvis et al. 2014; Reddy et al. 2017
	2. V	Jarvis et al. 2014, Prum et al. 2015
	3. VII	Jarvis et al. 2014
	4. III+V+VI+VII	Hackett et al. 2008
	5. Doves	McCormack et al. 2013
Hoatzin's sister	1. I+III+V	Reddy et al. 2017
	2. I	Prum et al. 2015
	3. <i>S+CR</i>	Jarvis et al. 2014
	4. V	Jarvis et al. 2014
	5. <i>S</i>	McCormack et al, 2013, Jarvis et al. 2014
	6. <i>TB</i>	Jarvis et al. 2014
	7. II+III+S+CR	Jarvis et al. 2014
Owls' sister	1. <i>M+W</i>	Jarvis et al. 2014; Prum et al. 2015; Reddy et al. 2017
	2. <i>W</i>	Jetz et al. 2012; McCormack et al. 2013; Jarvis et al. 2014
	3. <i>EHV</i>	Jarvis et al. 2014
	4. <i>M</i>	Hackett et al. 2008

NOTE.— the Greek numbers (bolded) correspond to the “magnificent seven” of Reddy et al. (2017), including **I**: core Landbirds; **II**: core Waterbirds; **III**: Sunbittern and Tropicbirds; **IV**: Cuckoos, Bustards, and Turacos; **V**: Nightjars, Swifts, Hummingbirds, and allies; **VI**: Doves, Mesites, and Sandgrouse; **VII**: Flamingos and Grebes. Other italic acronyms including *S*: Shorebirds; *CR*: Cranes and Rails; *TB*: Turacos and Bustards; *M*: Mousebirds; *W*: Woodpeckers and allies; *EHV*: Eagles, Hawks, and New World Vultures.

Table 2. Simulation results showing the best supported edge

Simulations	earliest Neoaves ¹ (63 [#])	TR	Hoatzin Sister ² (45)	TR	Owls Sister ³ (14)	TR	I (33)	II (4)	III (16)	V (27)
exon1	Doves + Turacos	14	Sunbittern + Turacos	12	Owls + Mousebirds	2	5432*	19746	8497	5581
exon2	Cuckoos + Sandgrouse	18	Sunbittern + Turacos	17	Owls + Mousebirds	2	6438	15763	7757	5416
intron1	Cuckoos + Doves	8	Sunbittern + Turacos	10	Owls + Mousebirds	2	4639	21835	7672	6130
intron2	Cuckoos + Sandgrouse	14	Sunbittern + Turacos	13	Owls + Mousebirds	2	3805	14934	7122	6537
UCE1	Cuckoos + Sandgrouse	14	Cranes + Flamingos + Grebes	13	Owls + Mousebirds	2	4329	18031	7621	5733
UCE2	Doves + Turacos	9	Hoatzin + Nightjars	17	Owls + Mousebirds	2	5906	20108	7616	6508

NOTE.— TR: the ranking of the true topology of each focal edge in the edge-based analyses. [#]: The number in the parenthesis represents the total number of conflicting constraints for each testing edge. ¹: showing the best hypothesis that is conflict with (VI + VII) + other Neoaves. ²: showing the best hypothesis that is conflict with Hoatzin + II + III + S + CR. ³: showing the best hypothesis that is conflict with Owls + EHV. Control edges include I, II, III, and V, which were all best supported by the edge-based analyses. *Bolded numbers: The difference of $\sum \Delta \ln L$ between the best hypotheses (i.e., I, II, III, and V respectively) and conflicting second best hypotheses for each edge. These large numbers indicate that the control edges can all be recovered with strong support. Clade keys can be found in Table 1 and Figure 2.

Figure Legends

Figure 1. A: Node support distribution among ML gene trees. B: Topology distribution of the ML gene trees for the three focal edges using Jarvis et al. (2014) exon data. Analyses for Introns and UCEs can be found in Figure S3. unsupported: relationships shown have less than 50% UFBS (ultrafast bootstrap support, we use lower support cutoff to show potential topology distributions). The Top 20 supported topologies are shown. rare resolutions: relationships shown in less than 10 gene trees. Hoatzin missing or Owl missing: Hoatzin or Owl is missing in the gene trees.

Figure 2. A: Keys (**I-VII**) for each major clade in the hypotheses being tested. B: flow chart on how to conduct edge-based analyses by using the earliest diverging Neoaves as an example.

Figure 3. The three focal edges and the testing hypotheses in the edges-based analyses using exons (upper), introns (middle) and UCEs (lower). A: The earliest diverging clade in Neoaves, including the five dominant hypotheses published before (1–5), and the five additional hypotheses prevalent observed in the gene trees for bias assessment (6–10). B: the sister of Hoatzin, including seven previously identified hypotheses. C: the sister group of Owls, including four commonly found hypotheses. $\sum \Delta \ln L$: the sum of log likelihood score differences; $\sum Ng$: the sum of number of genes. Hypotheses keys can be found in Figure 2.

Figure 4. A: The GC content variation among exons, introns, or UCEs, including ΔGC_{50} (the GC difference between the inter-quartile GC-rich taxa), ΔGC_{90} (the GC difference between the 95 and 5 percentile GC-rich taxa), and ΔGC_{100} (the GC difference between most GC-rich taxon and the least GC-rich taxon). B: Comparison of hypotheses supporting pattern between gene sets from the lower and upper quartile of ΔGC_{90} among data types. C: Tree length (left) and internode length (right) variation among different types of data. D: Comparison of hypotheses supporting pattern between gene sets from the lower and upper quartile of tree length variation. For x axis, EL/EH: exons with lower or higher ΔGC_{90} /tree length; IL/IH: introns with lower or higher ΔGC_{90} /tree length; UL/UH: UCEs with lower or higher ΔGC_{90} /tree length.

Figure 1

Figure 2

Figure 3

Figure 4