Molecular Phylogenetics and Evolution 158 (2021) 107091

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

A phylogenomic supermatrix of Galliformes (Landfowl) reveals biased branch lengths

Rebecca T. Kimball [a,*], Peter A. Hosner [a,b], Edward L. Braun [a]

[a] Department of Biology, University of Florida, Gainesville, FL 32607, USA
[b] Natural History Museum of Denmark and Center for Macroecology, Evolution and Climate, University of Copenhagen, Copenhagen, Denmark

A B S T R A C T

Building taxon-rich phylogenies is foundational for macroevolutionary studies. One approach to improve taxon sampling beyond individual studies is to build supermatricies of publicly available data, incorporating taxa sampled across different studies and utilizing different loci. Most existing supermatrix studies have focused on loci commonly sequenced with Sanger technology ("legacy" markers, such as mitochondrial data and small numbers of nuclear loci). However, incorporating phylogenomic studies into supermatrices allows problem nodes to be targeted and resolved with considerable amounts of data, while improving taxon sampling with legacy data. Here we estimate phylogeny from a galliform supermatrix which includes well-known model and agricultural species such as the chicken and turkey. We assembled a supermatrix comprising 4500 ultra-conserved elements (UCEs) collected as part of recent phylogenomic studies in this group and legacy mitochondrial and nuclear (intron and exon) sequences. Our resulting phylogeny included 88% of extant species and recovered well-accepted relationships with strong support. However, branch lengths, which are particularly important in down-stream macroevolutionary studies, appeared vastly skewed. Taxa represented only by rapidly evolving mitochondrial data had high proportions of missing data and exhibited long terminal branches. Conversely, taxa sampled for slowly evolving UCEs with low proportions of missing data exhibited substantially shorter terminal branches. We explored several branch length re-estimation methods with particular attention to terminal branches and conclude that re-estimation using well-sampled mitochondrial sequences may be a pragmatic approach to obtain trees suitable for macroevolutionary analysis.

1. Introduction

Macroevolutionary studies provide insights into a range of different questions, including the origin of biodiversity, trait evolution, and biogeography (e.g., Pulido-Santacruuz and Weir 2016; McEntee et al. 2018; Ksepka et al. 2020). Ideally, macroevolutionary analyses should be based on phylogenies with complete (or near complete) taxon sampling, strong statistical support for relationships, and branch lengths that accurately reflect the timing of diversification. Although it may not be necessary to sample all taxa for all comparative studies— or even possible, due to extinctions and sampling restrictions, the results of many types of macroevolutionary studies are biased when some or all clades are poorly sampled (e.g., Salisbury and Kim 2001; Wang et al. 2017; Marcondes 2019). This suggests that it may be important to sample both appropriate representatives relative to the question at hand, as well as to include a large number of species throughout the clade.

One approach for building taxon-rich phylogenies is to obtain orthologous loci for all (or almost all) species in the clade. However, such extensive data collection is not always feasible, particularly when some challenging-to-resolve relationships require large-scale data collection. In these cases, there are two *meta*-analytical approaches to build large-scale phylogenies that take advantage of previously collected data: 1) supermatrices (de Quiroz and Gatesy 2007), and 2) supertrees (Bininda-Emonds 2004). Supermatrix methods compile as much data as possible (typically DNA sequence data), relying primarily (or even exclusively) on previously published data from a range of different studies. This results in a heterogeneous data matrix. But, as long as there is reasonable overlap among taxa for some loci (e.g., some loci sampled for many taxa, and or loci sampled for different combinations of taxa), the resulting data matrix can provide reliable estimates of large phylogenetic trees (e.g., Driskell et al. 2004; Goloboff et al. 2009). Alternatively, supertree methods involve combining published trees (that may have been estimated from sequence or other types of information) to yield a much larger tree (Sanderson et al. 1998; Binida-Emonds et al. 2004; Cotton et al. 2009; Warnow 2018). Similar to the case for data in supermatrices, supertree methods can result in a reasonable estimate of phylogenetic relationships as long as trees exhibit sufficient overlap (e.g., Bininda-Emonds et al. 2007; Brown et al. 2017).

While both methods can yield species-rich trees, supermatrix methods result in trees that have a more direct connection to the original

* Corresponding author at: Department of Biology, Gainesville, FL 32607, USA.
E-mail address: rkimball@ufl.edu (R.T. Kimball).

https://doi.org/10.1016/j.ympev.2021.107091
Received 23 June 2020; Received in revised form 16 January 2021; Accepted 27 January 2021
Available online 2 February 2021

underlying data. This direct connection between the data and the estimate of phylogeny in supermatrix methods makes branch length estimation straightforward. Another advantage of the supermatrix approach is its ability to weight information in an intuitive manner and to provide information regarding clade support. In contrast, the relatively indirect connection between the data and the estimate of phylogeny in supertrees makes difficult to estimate branch lengths, although there have been approaches to impose branch lengths on supertrees (e.g., Bininda-Emonds et al. 1999; Webb et al. 2008; Ren et al. 2009; Torices 2010; Kimball et al. 2019). Although methods to upweight more reliable source trees and examine support for supertrees certainly do exist (e.g., Bininda-Emonds, 2004; Burleigh et al. 2006; Kimball et al. 2019), stronger support for relationships emerges naturally in those parts of the tree where more data are available in supermatrix methods. Ultimately, this direct connection between the data and the estimate of phylogeny is the reason that some have advocated the superiority of supermatrix approaches (Gatesy and Baker 2005; Gatesy et al. 2002).

The supermatrix approach suffers from two potential limitations. First, there may be insufficient data to resolve problematic nodes because many supermatrix studies rely on relatively small numbers of loci (e.g., Burleigh et al. 2015; Dufort 2016; Shakya and Sheldon 2017). Second, supermatrix studies often have large amounts of missing data (sometimes $>$ 90%). Although studies have shown missing data is typically not problematic for inferring phylogenetic relationships (e.g., Fulton and Strobeck 2006; Burleigh et al. 2015), it does have the potential to yield trees with biased branch lengths estimates (e.g., Darriba et al. 2016).

Next-generation sequencing (NGS) technologies provide a transformative solution to the problem of insufficient loci in phylogenetics. These technologies facilitate acquisition of massive amounts of sequence data that help resolve problematic nodes (Braun et al. 2019), but the taxon sampling available from NGS approaches lags far behind what is available from Sanger technologies (e.g., compare number of species in Burleigh et al. 2015 with Kimball et al. 2019). Incorporating large-scale phylogenomic data, such as ultraconserved elements (UCE) and transcriptomic data, along with more standard mitochondrial and nuclear data ("legacy" markers) gathered over the past 30 years into a supermatrix is potentially an elegant solution. Phylogenomic datasets provide deep locus sampling to infer difficult relationships with confidence, while 30 years of legacy data leverage broad taxon sampling crucial for macroevolutionary analyses.

Galliformes (chickens, turkeys, peafowl, quail and allies) present a useful test case to examine the performance of "phylogenomic supermatrix" analyses that incorporate NGS data and legacy markers. Phylogenetic relationships among certain galliform genera and taxa have been problematic for many years (reviewed by Wang et al. 2013). However, several phylogenomic studies have utilized sequences of thousands of ultra-conserved elements (UCEs) to estimate phylogenetic relationships with strong support (e.g., Sun et al. 2014; Meiklejohn et al. 2016; Persons et al. 2016). Unfortunately, phylogenomic data is not available for all species. On the other hand, the majority of galliform species have been sampled for legacy markers - i.e., various mitochondrial gene regions (which are available for most species) and a number of nuclear loci (particularly intronic regions, but also some coding regions and 3' untranslated regions). Thus, Galliformes is a model group to explore the feasibility of building phylogenomic supermatricies to estimate both taxon-rich and well-resolved, phylogenies.

A taxon-rich galliform phylogeny has the potential to be especially useful as a tool for comparative studies given our extensive knowledge of galliform biology. Galliforms are agriculturally and therefore economically important, are often used as models for physiological, genetic, and developmental research. They exhibit wide variation in ecological, behavioral and morphological traits. Across species, there is $\sim$ 100-fold variation in mass. While pheasants are highly sexually dimorphic, with males exhibiting bright colors and elaborate plumage, other species are monomorphic, sometimes being dull colored in both sexes. Species also vary in mating system (monogamous, polygynous, lekking), and in modes of parental care (including male-only parental care). Clutch sizes also vary, being one or two eggs in some species to$>$15 eggs in others. Overall, the wide variety of characteristics in this order has made Galliformes a focal group for many comparative studies (Davison 1985; Kimball et al. 2001; Kolm et al. 2007; Nadeau et al. 2007; Krakauer and Kimball 2009; Lislevand and Figuerola 2009; Kimball et al. 2011; Balasubramaniam and Rotenberry 2016; Wang et al. 2017; Hosner et al. 2017; Hosner et al. 2020).

Herein we construct and analyze a phylogenomic supermatrix for Galliformes that samples 88% of extant species. We ask whether combining phylogenomic data (comprising millions of base pairs [bp]) with taxa sampled for one or two loci (that are therefore represent by hundreds of bp) is problematic. We acknowledge that taxa represented by a limited data are primarily represented only by mitochondrial sequences, which are relatively rapidly evolving in animals. Taxa with little missing data will have substantial and proportionately more slowly evolving nuclear data, particularly here where the phylogenomic data largely comprises slowly evolving UCEs. Ultimately, the question of whether supermatrix analyses combining taxa with a large variation in amount of missing data and markers with a large variation in substitution rate are problematic and if so, which specific issues (topology, branch length estimates, or both) is empirical in nature. Thus, we examine the performance of analyses using a phylogenomic supermatrix and explore approaches that may yield realistic branch length estimates for these very large and locus-rich supermatrices.

2. Methods

2.1. Sequences and alignment

We searched GenBank and recent publications to identify loci that have been widely sampled in previous galliform phylogentic studies. To ensure loci included a mixture of taxa (and not just Gallus and other economically important taxa), we retained loci sampled in at least 20 species and sampled broadly across clades. For loci in which more than one region was sampled, we typically aligned each region separately. The exceptions were EEF2, one region for FGB (introns 6 and 7) and HMGN2, where two relatively short introns, and a short intervening coding exon were kept as a single partition (typically these comprised a single amplicon). For mitochondrial data, even when complete mitogenomes were available, we restricted our sampling to the two rRNA regions and the 13 protein coding genes. We refer to these as "legacy" markers as they were largely obtained using traditional Sanger sequencing.

We extracted legacy data from Genbank (Benson et al. 2015) using blastn (Camacho et al. 2009). We selected 83 queries likely to have been sequenced from a sufficient number of galliform taxa and retrieved the chicken (*Gallus gallus*) sequence for each of those regions. Then we masked repetitive sequences (e.g., CR1 transposons; Stumph et al. 1981) using the CENSOR program (Kohany et al. 2006), which is available from https://www.girinst.org/censor. From this we generated two sets of files for each galliform species; one file included sequences that have R.T.K. as an author and a second file with different authors. This was done because our research group, independently or in collaboration with other groups, has conducted many phylogenetic studies focused on Galliformes (see literature cited). Prioritizing the selection of sequences that generated by our research group should reduce the number of distinct individuals represented in sequence. Then we extracted any sequences with a significant ($E < 10^{-20}$) blastn hit to the query and placed the sequences in a file. To supplement the downloaded sequences, we added both some unpublished sequences (using methods detailed in our previous studies; MT587887-MT588075) and extracted homologous regions from published genomes, including from the genome of *Anas platyrhynchos* which we used as an outgroup. This led to a total of 64 independent alignments (Supporting Information Table S1)

sampled for 265 galliform species (Supporting Information Table S2).

UCE sequences were mainly derived from several recent studies reconstructing Galliform phylogeny (Sun et al. 2014, Hosner et al. 2015, Meiklejohn et al. 2016, Persons et al 2016, Hosner et al. 2016a, Hosner et al. 2016b, Hosner et al. 2020). We took assembled contigs from those studies, and rematched contigs to UCE loci in Phyluce 1.5 (Faircloth et al. 2012). We aligned each locus in MAFFT 7 (Katoh et al. 2002, Katoh and Standley 2013). We then trimmed alignment ends when 35% of cells over a 20 bp sliding window were missing. We retained 4577 UCE alignments with a minimum of 20 taxa for downstream analysis.

2.2. Data vetting and concatenation

To identify potential problems with data downloaded from GenBank, we estimated the maximum likelihood tree from each gene or mitochondrial region using RAxML ver. 8.2.10 (Stamatakis 2014) with GTRGAMMA and 10 replicates. We examined each gene tree to ascertain whether any taxa were recovered in highly unexpected positions or exhibited extremely long branches. Through this, we identified seven sequences that we excluded from alignments. In all cases, examination of the alignment indicated that these sequences were identical or nearly identical, over a large part of the sequence, to an unrelated taxon (different genus or even family). Five gene regions from complete mitogenomes were excluded: ND1 from *Acryllium vulturinum* (also noted by Meiklejohn et al. 2014), CYB from *Francolinus pintadeanus*, and ATP6, ATP8 and COII from *Francolinus pondicerianus*. Additionally, two sequences extracted from genomes were also excluded: introns ACAN 1 from *Coturnix japonica* and CHRNG from *Centrocercus minimus*. After excluding these sequences, we then concatenated remaining data of both the legacy and UCE data using SequenceMatrix 1.7.8 (Vaidya et al. 2011). The complete matrix is available as Supporting Information (Kimball_Supermatrix.nex).

2.3. Phylogenetic analyses

To partition data for phylogenetic analyses, and some branch length re-estimation, (see below), we used PartitionFinder 2.1 (Lanfear et al. 2017) using linked branch lengths, AIC_c, and the rclusterf scheme. As input to PartitionFinder, we considered each UCE, intron or UTR alignment as a distinct partition. For coding regions (both nuclear and mitochondrial), we separated into 1st, 2nd, and 3rd codon positions. PartitionFinder identified a scheme of 96 subsets used for all partitioned analyses.

We executed ML searches and bootstrapping in ExaML 3.0.17 (Kozlov et al. 2015). For ML searches, we implemented 20 rapid hill-climbing searches from random start trees. We implemented both unpartitioned and partitioned analyses with the GAMMA rate heterogeneity model. We also executed 100 bootstrap searches on both the unpartitioned and partitioned datasets, each with the GAMMA and PSR rate heterogeneity models (PSR was not used to identify the ML topology as likelihood values are not comparable across replicates with PSR), also using the rapid hill-climbing mode and random start trees, which we report as the proportion of trees supporting a node.

We compared our results with recent, large-scale studies that included many galliforms (Jetz et al. 2012; Burleigh et al. 2015; Stein et al. 2015; Brown et al. 2017). For comparison to Jetz et al. (2012) we downloaded all 10,000 trees with the Hackett et al. (2008) backbone and generated a majority rule consensus tree. For the other three, we used a published ML tree. We reconciled taxonomic names, and excluded any taxa that did not have a clear match between our tree and each of the published trees, as well as outgroup taxa. The Brown et al. (2017) tree included subspecies so we selected the first subspecies as the match to our species. We then converted names in the published trees to match the spelling and name used in our analysis (Supporting Information Table S3) and calculated normalized RF (Robinson and Foulds 1981) distances between each tree and our unpartitioned topology using PAUP 4.0a Build 168 (Swofford 2002) with unrooted trees. We also generated a strict consensus topology for each published topology and our unpartitioned ML tree to allow an assessment of conflicts between trees and assessed the number of polytomies in each consensus tree. We did this by calculating the normalized RF distance between the consensus tree and an unresolved tree, doubling that distance, and converting to a percentage; this yields the percent resolved branches in each consensus tree.

2.4. Branch length estimation

We explored several approaches to adjust branch length heterogeneity that might be due to missing data and variable substitution rates among loci. To do this, we used a fixed topology (both the ExaML total evidence topologies resulting from the unpartitioned as well as the partitioned analysis) and re-estimated branch lengths using RAxML. We did this in several ways. First, we re-estimated branch lengths with no other changes (e.g., data was unpartitioned). Second, we re-estimated branches while partitioning the data using the PartitionFinder results. Third, we kept each non-UCE locus as a separate partition (e.g., no separating into codon positions). Due to limits on the number of partitions allowed for some options (see below), we combined all UCEs into a single partition. Fourth, we used the stolen branch length correction implemented in RAxML (-f k), along with the per-partition branch length estimation (-M). For this method, we partitioned as in the third method (by locus, and combining UCEs), as the -M option is limited to a maximum of 128 partitions in pre-compiled versions of RAxML. This approach is also implemented in ForeSeqs (Darriba et al. 2016). We attempted branch length estimation in ForeSeqs, but this analysis ran out of memory after completing very few calculations, and so only results from RAxML are included.

We used these four branch length re-estimation approaches with two different sets of sequence data: the entire supermatrix, and just the two loci with the least missing data (the mitochondrial ND2 and CYB regions, for which<20% of sites were missing). All taxa were sampled for at least one of these two mitochondrial markers and 76% were sampled for both. We explored all of these approaches using both the GTRGAMMA and GTRCAT models. Note that GTRCAT refers to the among-sites rate heterogeneity model described by Stamatakis (2006); this approach is quite distinct from the CAT mixture model described by Lartillot and Philippe (2004). The CAT rate heterogeneity model has been renamed the PSR (per site rate) model (Stamatakis and Aberer 2013) and the latter nomenclature is used in ExaML. We use the PSR nomenclature throughout this manuscript wherever possible to avoid confusion; however, we used "GTRCAT" when referring to the RAxML model because it is the name of the option in that program. Thus, for each of the two trees (unpartitioned and partitioned ExaML topologies), we had 16 alternative branch length estimates (4 approaches, 2 datasets, and 2 models).

Since the greatest skew in branch lengths was on the terminal branches, we evaluated the impact of our different approaches by examining the sum of the terminal branch lengths to the sum of the internal branch lengths. Methods that reduce the bias in branch lengths due to missing data are expected to have smaller ratios, whereas methods that had no effect (or a negative effect) on branch lengths should have large ratios. Branch lengths (both terminal and internal) estimated from the more rapidly evolving mitochondrial markers are expected to be longer due to the more rapid evolutionary rate of these regions, so comparing overall tree lengths among all analyses would not be valid.

To explore whether branch length issues were ameliorated in ultrametric trees, we compared the ML unpartitioned phylogram with original branch lengths and the same tree with branch lengths re-estimated using mitochondrial data and GTRGAMMA in RAxML. To make the tree ultrametric, we used the chronos function in ape, version 5.4–1 (Paradis and Schliep 2019) with a correlated model of evolution and lambda = 1.

We then used the same process as above to compare the ration of the terminal branch lengths to the internal branch lengths.

2.5. Statistical analyses

We performed Pearson and Spearman (depending on the distribution of the data) correlations in R (ver. 3.6.1; R Core Team 2019) using cor.test (standard correlation analyses). We also performed partial correlations using the ppcor package (Kim 2015).

3. Results

3.1. Data matrix

The data matrix contained 265 ingroup species (88% of the total). In initial analyses, one taxon (*T. cuvieri*, represented only by a small amount of mitochondrial data), did not group with other *Talagella*. However, previous studies have shown this genus to be monophyletic, so we considered this taxon to be misplaced. Given this, we excluded *T. cuvieri* from the supermatrix for all results presented. No other unexpected placements were noted.

This left us with 264 ingroup taxa (plus one outgroup taxa), of the 300 species recognized by the IOC World Bird List v 9.2 (Gill and Donsker 2019; the 300 species includes the putatively extinct *Ophrysia supercilliosa* that we did not sample). Thus, our supermatrix represented 88% of the order (Table 1). We had reasonably even sampling throughout, with representation of at least 85% of each family. Within the largest family, Phasianidae, we represented the three major groupings, though we sampled only 58% of the Arborophilinae (largely due to a lack of sampling in the genus *Arborophila*).

The final data matrix we analyzed included 1,875,451 sites for 264 galliform taxa (Table 2). The complete alignment had 742,485 distinct alignment patterns, and 69.7% missing data. The majority of these sites were due to the UCE data, though over 54 kb were from "legacy" markers. Distribution of sequence data was heterogeneous among species (Fig. 1), and ranged from as few as 178 bp up to 1,754,137 bp (about 10,000x difference). Every species in the supermatrix was represented by at least some mitochondrial data, 223 (84%) had at least some nuclear data and 107 (41%) included UCE data (five of these lacked any other nuclear data).

3.2. Phylogenetic relationships

The resulting phylogeny recovered the higher-level relationships within galliforms (Fig. 2). All five currently recognized families were recovered with strong support. Within the largest family, Phasianidae, we recovered the Arborophilinae (Crowe et al. 2006), and the remaining phasianid taxa separated into two major clades, which have been designated the "erectile clade" (Kimball and Braun 2008) and the "non-erectile clade" (Kimball and Braun 2014). Relationships throughout the tree were largely well-supported, with 80% of nodes having support values of at least 0.9 (and 84% with at least 0.8). Not surprisingly, lower support values were typically (but not exclusively) recovered in the parts of the tree where taxa were represented by small amounts of data. The unpartitioned and partitioned trees were similar in overall structure (see Supporting Information Treefile), and the few differences involved poorly supported nodes.

We identified numerous topological differences between our new results and those of other published studies (Table 3). Stein et al. (2015), which was the only galliform-specific study included, had a quarter of bipartitions that differed from the analysis of our own supermatrix. However, this was still more in line with our topology than the other examined trees, where the proportion of bipartitions that differed varied from 0.33 to 0.42. The consensus trees (Table 3, Supporting Information Figure S1 and Treefile) showed somewhat similar results, with the consensus to the Stein et al. (2015) topology exhibiting the highest degree of resolution. The Jetz et al. (2012) topology had the lowest resolution (Table 3; Supporting Information Figure S1C), likely due to misplacement of *Ptilopachus nahani*, which Jetz et al. (2012) placed in an incorrect family (see also Wang et al. 2017).

Table 1
Taxonomic sampling in supermatrix.

		Number of Species	Represented in Supermatrix	Percent Sampled
Megapodiidae		21	19	90
Cracidae		55	47	85
Numididae		6	6	100
Odontophoridae		34	32	94
Phasianidae		183	160	87
	Arborophilinae	24	14	58
	Non-erectile	96	86	90
	Erectile	63	60	95
Total		299	264	88

Table 2
Different types of data represented in supermatrix.

	Number of Partitions	Maximum Number of bp	% of Data Matrix	Number of Species[3]
Mitochondrial Coding	12[1]	11,389	0.61	265/266
Mitochondrial rRNA	2	2,867	0.15	172
Nuclear Coding	10	8,081	0.43	93
Nuclear Intron[2]	35	28,791	1.54	216
Nuclear UTR	5	3,610	0.19	46
Nuclear UCE	4577	1,820,713	97.08	108

[1] Mitochondrial ATP6 and ATP8 were combined into a single partition.
[2] Intron alignments represented 31 distinct loci.
[3] Includes outgroup

3.3. Branch lengths

The ML phylogram showed a high degree of heterogeneity in branch lengths, particularly among the terminal branches (Fig. 3). Since terminal branch lengths were highly correlated between unpartitioned and partitioned analyses ($r = 0.99$, $p < 0.0001$, $n = 265$; Supporting Information Fig. S2), we focused on the unpartitioned topology for the remainder of analyses. As expected, longer terminal branch lengths were correlated with the amount of missing data (Fig. 4, Spearman's $r = 0.57$, $p < 0.0001$, $n = 265$), suggesting that missing data may represent a major source of bias in the branch length estimates.

Mitochondrial data evolve far more rapidly than nuclear data in vertebrates, particularly when compared to the highly conserved UCEs that formed the greatest amount of data for some species in our dataset. Species whose branch length estimation depends primarily (or completely) on mitochondrial data would be expected to have much longer branches than species where nuclear data dominated, as we observed (Spearman's $r = 0.62$, $p < 0.0001$, $n = 265$). However, species with the highest proportions of missing data were also those where the available data was dominated by mitochondrial data (Spearman's $r = 0.84$, $p < 0.0001$, $n = 265$). A significant relationship between terminal branch length and proportion of mitochondrial data persisted even when using a partial correlation to control for the affect of missing data (Spearman's $r = 0.33$, $p < 0.001$, $n = 265$), suggesting that branch lengths are affected by data type (with the rate of evolution likely having the most important role) as well as missing data.

Re-estimating branch lengths with other methods did affect resulting branch lengths, as expected. When using the entire dataset, partitioning (either with PartitionFinder results or by locus with UCEs combined) did affect the ratio, but only when using the PSR model (Fig. 5). Unexpectedly, the stolen branch actually increased the ratio (so attributed

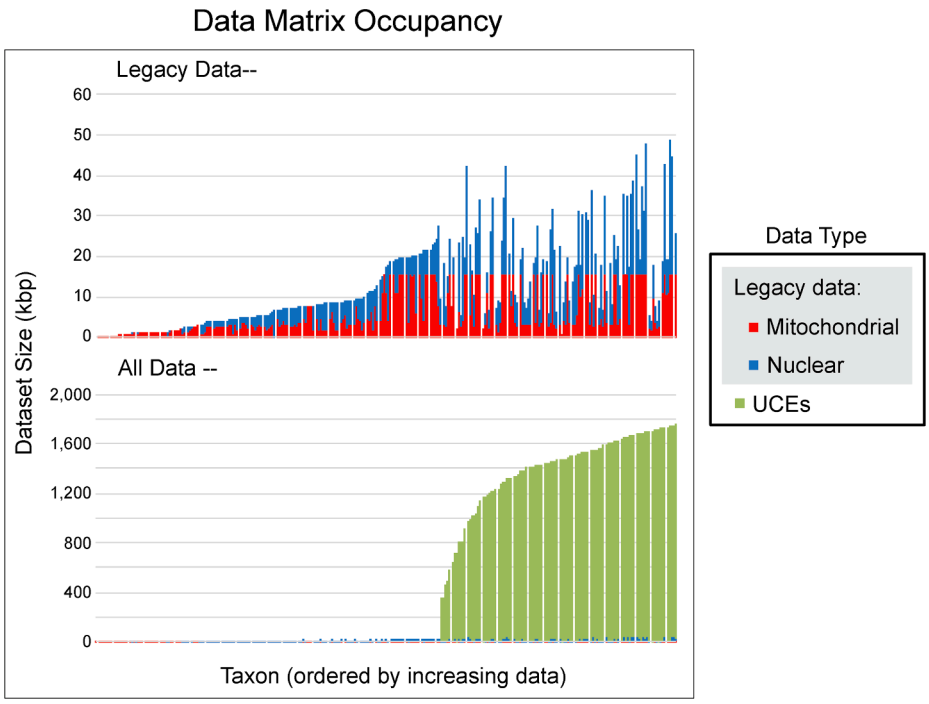

Fig. 1. Data type and amounts for each taxon. Taxa are organized based on the amount of data. The taxon with the smallest amount of data (*Francolinus pictus*; 178 bp) is presented to the left) and the taxon with the most data (*Tetrastes bonasia*; 1,754,137 bp) is presented to the right. The upper panel focuses exclusively on legacy (mitochondrial and some nuclear) data. The lower panel presents information for all three data types, although the large amount of UCE data makes the mitochondrial and legacy nuclear data difficult to visualize.

more of the tree length to the terminals rather than internal branches), though this was less extreme when using the PSR model. In all cases, re-estimating branch lengths using the mitochondrial CYB and ND2 sequences reduced the ratio of terminal to internal branch lengths (Fig. 5). For the mitochondrial data, there was relatively little effect of the rate heterogeneity model (GAMMA versus PSR), whether the sequences used to estimate the branch lengths was unpartitioned or partitioned, or whether or not the branch length correction in RAxML was used (Fig. 5).

Forcing the tree to be ultrametric did not ameliorate the branch length skews. The ratio of the terminals to internals in the ultrametric tree estimated from the phylogram based on all data was>2-fold greater than the ultrametric tree based on the phylogram with mitochondrial adjusted branch lengths (3.17 versus 1.38). This was actually a greater difference than was observed for the phylograms themselves, where the ratio for the original phylogram was<2-fold greater than the mitochondrial phylogram (2.59 versus 1.48; see also Fig. 5).

Overlaying the original phylogram (Fig. 3) to the tree with branch lengths adjusted using mitochondrial data (unpartitioned GAMMA) demonstrates the shifts in terminal branch lengths (Fig. 6). In addition, comparing the change in terminal branch lengths between the original topology and the MT-GAMMA unpartitioned re-estimation shows that the greatest change occurred in taxa with large amounts of missing data (Spearman's r = 0.20, $p = 0.0009$, $n = 265$) and a high proportion of mitochondrial data (Spearman's r = 0.25, $p < 0.0001$, $n = 265$).

4. Discussion

Our supermatrix phylogeny was well-resolved and it demonstrates how combining large amounts of NGS data for a set of backbone taxa with other data types for a broader range of taxa may overcome some of the limitations in taxon-sampling and resolution that have affected previous comparative studies in galliforms (e.g., Wang et al. 2017). However, we also highlighted some of the issues that can arise when using supermatrices, particularly that branch length estimation may be biased. While the impact of missing data on branch length estimation (i.e., the potential for branch length mis-estimation) has been acknowledged (Darriba et al. 2016) we highlight that data type (i.e., loci of varying rates of evolution) is likely to represent another factor that is very important. Thus, controlling only for missing data (e.g., minimizing amounts of missing data) might be insufficient to obtain accurate branch length estimates in matrices like ours that also have a great deal of heterogeneity in the proportion of different data types for each species. Despite these issues, however, we also establish that there are computationally efficient and easily applied methods that can improve branch length estimation.

4.1. Galliform megaphylogenies

There have been previous large-scale phylogenies of both galliforms (Eo et al. 2009; Stein et al. 2015) as well as of all birds, with the latter also sampling many galliforms (e.g., Jetz et al. 2012; Burleigh et al. 2015; Brown et al. 2017; Kimball et al. 2019). These published megaphylogenies have been estimated using both supertrees (e.g., Eo et al. 2009; Brown et al. 2017; Kimball et al. 2019), supermatrices (Burleigh et al 2015; Stein et al. 2015) or a hybrid constrained supermatrix approach (Jetz et al 2012). Although these other megaphylogenies exist, all of these have some limitations that we have overcome. Our supermatrix includes more taxa than most previous megaphylogenies, and about 100x more sites for analysis than other supermatrix studies. While our supermatrix includes fewer galliform species than the complete Jetz et al. (2012) phylogeny (and the Brown et al. 2017 supertree study that included the complete Jetz et al. tree), Jetz et al. 2012 included some taxa not represented by underlying sequence data, but instead placed based on other information. Subsequent studies have demonstrated examples of misplaced galliforms in the Jetz et al. 2012 tree (e.g., Hosner et al. 2015; Persons et al. 2016), and that some misplaced taxa may have biased comparative analyses (e.g., Wang et al. 2017). Our comparison of these phylogenies relative to the one estimated here highlight the number of differences among studies. It is not surprising that the Stein et al. (2015) tree the closest to our supermatrix; the Stein et al. (2015) study focused on galliforms, and thus likely had a more carefully curated set of galliform sequences. Overall, although other megaphylogenies are available, our supermatrix is a major improvement for Galliformes by including a large number of taxa (all supported by underlying sequence data), and the benefit of the large number of sites to resolve problematic galliform relationships.

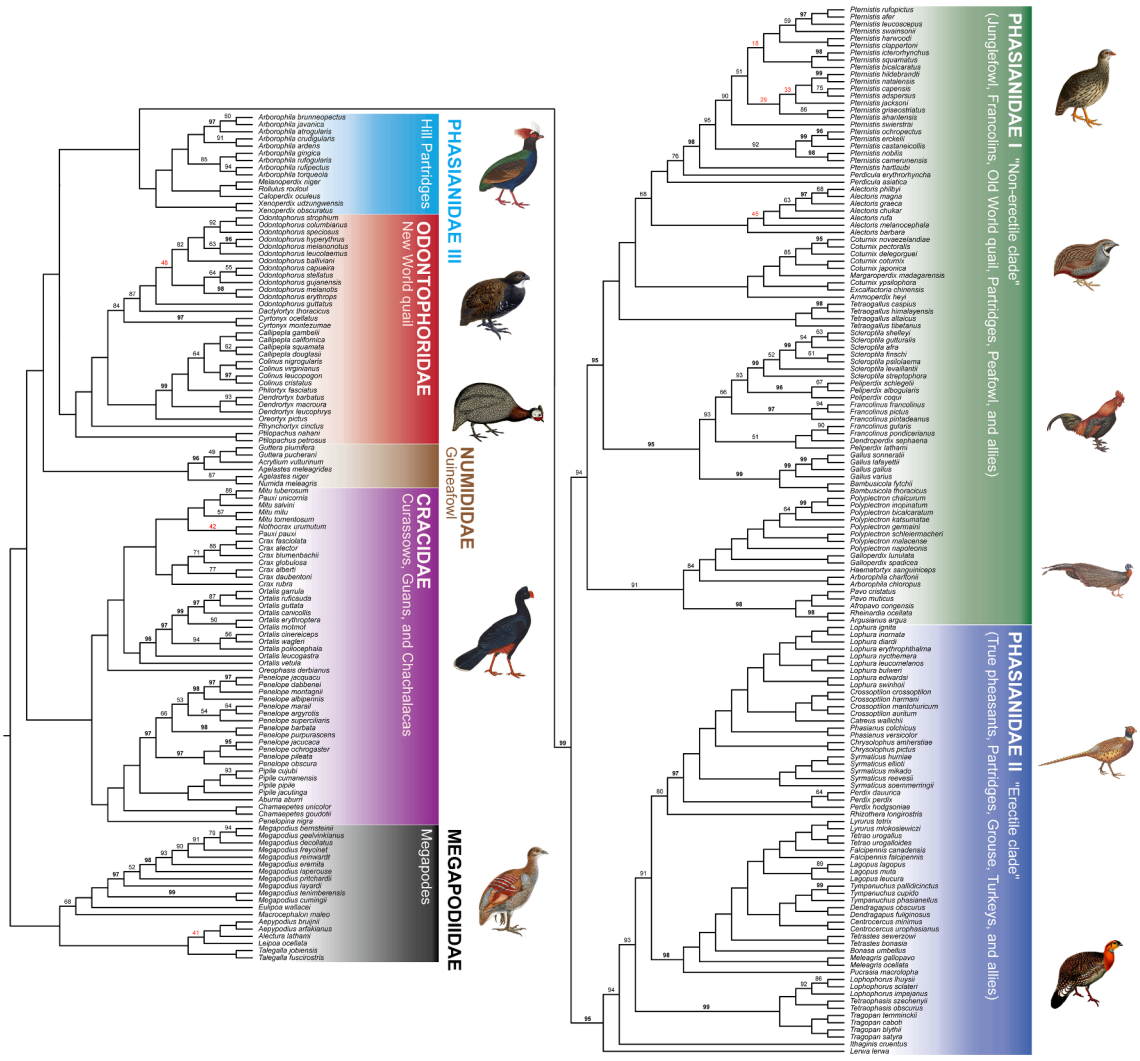

Fig. 2. Cladogram showing relationships among galliform birds. This figure presents the topology recovered in the unpartitioned analysis of the supermatrix; the partitioned analysis is quite similar (Supporting Information Treefile). Bootstrap support is presented above each branch whenever it is < 100% (due to the number of species it will be necessary to zoom in the tree to read species names). Each of the four families outside Phasianidae is assigned a color to the right of the tree, as are the three large clades (Arborophilinae, the "erectile clade," and the "non-erectile clade") within Phasianidae. Illustrations of representative taxa are also presented to the right of the taxa (see Supporting Information for sources of the illustrations).

Table 3
Comparison of unpartitioned ML topology with other large-scale galliform trees.

	# Matching Taxa (% matched)	RF Distance	Normalized RF	% Resolution of consensus tree
Brown et al.	264 (100%)	170	0.33	57%
Burleigh et al.	195 (74%)	160	0.42	56%
Jetz et al.	262 (99%)	206	0.40	42%
Stein et al.	221 (84%)	98	0.23	75%

4.2. Branch length estimation

Comparative methods ideally utilize branch length information that reflects evolutionary time; therefore, the quality of branch length estimates can have a profound impact on comparative studies (e.g., Litsios and Salamin 2012). Thus, obtaining unbiased estimates of branch lengths is important, and can be a limitation of megaphylogenies. Unlike supertree methods, which do not provide branch length estimates in the absence of additional information (Kimball et al. 2019), supermatrices typically involve analyses of sequence data and branch lengths are obtained as part of likelihood analyses. However, concerns have been raised about the impact of missing data on branch length estimation (Darriba et al. 2016). Although not always explicitly examined, earlier studies have also re-estimated branch lengths using well-sampled mitochondrial sequences (e.g., McGowen et al. 2009; Sun et al., 2014), suggesting that concerns about branch length estimation in light of missing data have long been recognized. Since supermatrices typically have substantial missing data, it stands to reason that branch length estimates in supermatrix studies may be biased.

We show that type of data, in addition to the amount of missing data, can also affect branch length estimation (Fig. 4). This suggests that focusing just on minimizing missing data (e.g., removing poorly sampled loci) may not always effectively deal with biased branch lengths. In many supermatrices, as we observed, species with large proportions of missing data are likely to be represented only by mitochondrial information. These are often taxa for which good quality tissues are limited and have been more likely sampled for mitochondrial regions which are more easily obtained from degraded and/or limited amounts of DNA. Thus, for these species, branch lengths may be particularly skewed even though they are often taxa of interest because they are often rare and/or difficult to study.

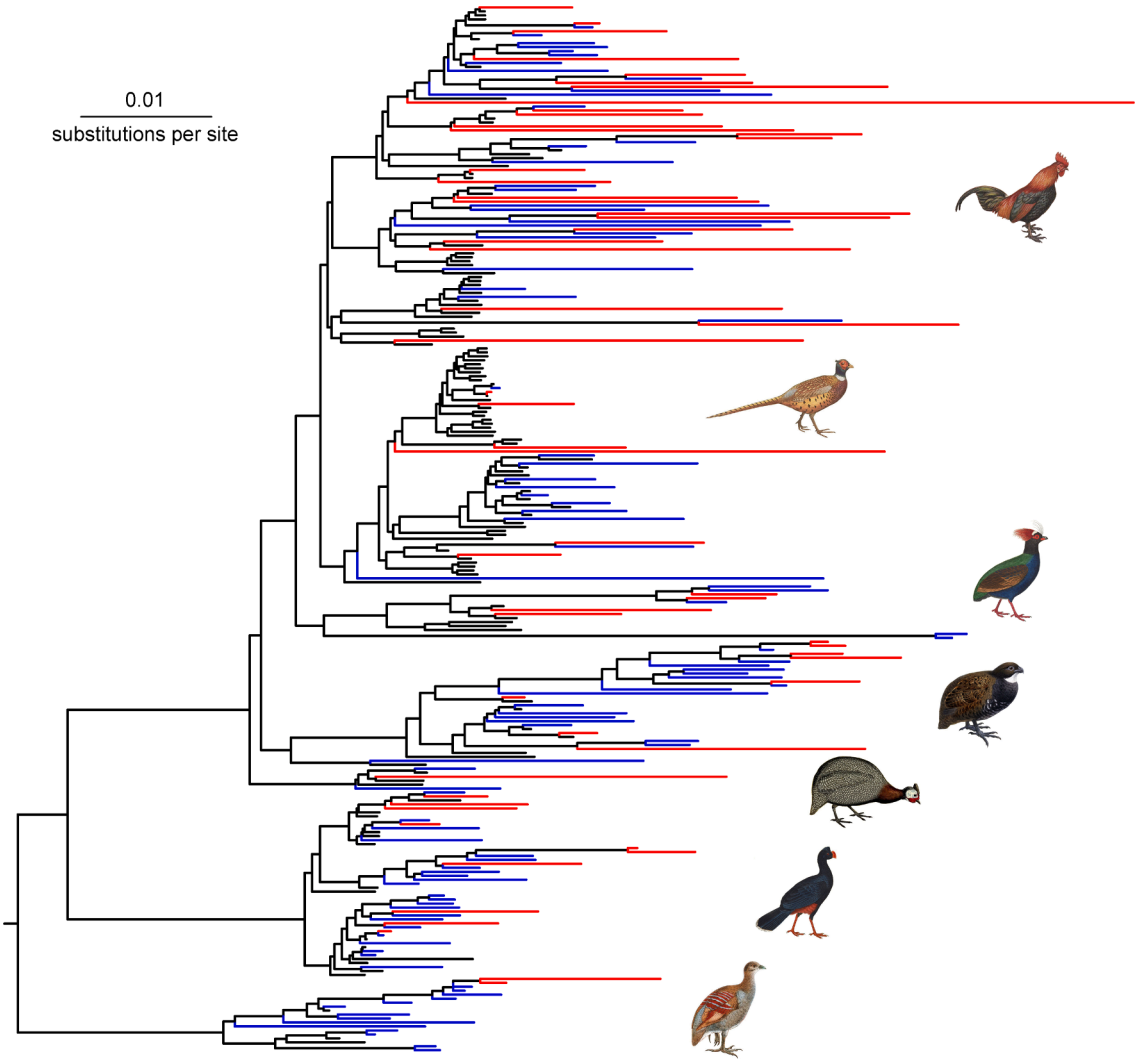

Fig. 3. Phylogram for the unpartitioned analysis with terminal branches labeled based on the type of data and the amount of missing data. Terminal branches are red if > 75% of the available data are mitochondrial and blue if UCE data are absent. Taxon order is identical to the order in Fig. 2, with Megapodiidae at the bottom and Phasianidae I (the non-erectile clade) at the top. To orient readers, we have used seven of the bird illustrations from Fig. 2 to the right of the tree. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Our supermatrix tree exacerbated branch length issues because inclusion of NGS data led to extreme differences in data scope for different species. Since most supermatrices have been constructed using some mitochondrial and small numbers of nuclear markers (e.g., Burleigh et al. 2015; Dufort 2016; Shakya and Sheldon 2017), the skew in branch lengths is likely to be much less problematic than we observed. However, as more NGS data becomes available, future supermatrices are likely to have the high disparity in sequence amounts that we observed. Data structure in such studies will likely be similar to ours, with NGS sampling to tackle problematic nodes and/or to represent the phylogenetic backbone across genera and families, but with some taxa represented by legacy (primarily mitochondrial) data. Thus, effective methods to obtain more realistic branch lengths for macroevolutionary studies will remain important until the taxonomic scope of legacy data has been entirely superseded by phylogenomic data.

When using the entire dataset, implementation of the PSR (GTRCAT in RAxML) model always resulted in a lower ratio of terminal to internal branch lengths. This was particularly true for partitioned analyses, where partitioned PSR analyses showed similar levels of improvement to that obtained by re-estimating branch lengths using mitochondrial data. In addition to exhibiting good performance in reducing the ratio of terminal to internal branch lengths, the PSR model was also computationally efficient. Based on theory, the PSR model is expected to have memory and run-time requirements about 1/4 of GTRGAMMA (assuming a four-category discrete approximation to the Γ distribution; Stamatakis 2006). Our empirical results compared well to theoretical estimates. Using the complete dataset and estimating without partitioning the GTRGAMMA estimation took 3.6x longer (GTRGAMMA = 15,428.4 s; PSR = 4,177.2 s based on the RAxML information file; obviously exact times will depend on specific computer resources, but all analyses here were run on the same system). Implementing partitioning greatly added to the total compute time, and showed greater variability depending on whether partitioning used the PartitionFinder results (GTRGAMMA = 181,369.9 s; PSR = 21,388.3 s) or as we partitioned for the stolen branch analysis (GTRGAMMA = 95,605.0 s; PSR = 48,961.1 s). However, although partitioning may add to the compute time for both models, the improved performance of partitioned PSR branch length re-estimation may outweigh those costs, and still require less compute time than GTRGAMMA.

We also show that re-estimating branch lengths using a small number of well-represented loci in the dataset (e.g., McGowen et al. 2009; Sun et al., 2014), in our case the mitochondrial CYB and ND2 genes, is a simple and practical strategy that effectively reduces terminal branch lengths for taxa with a large amount of missing data or high proportions

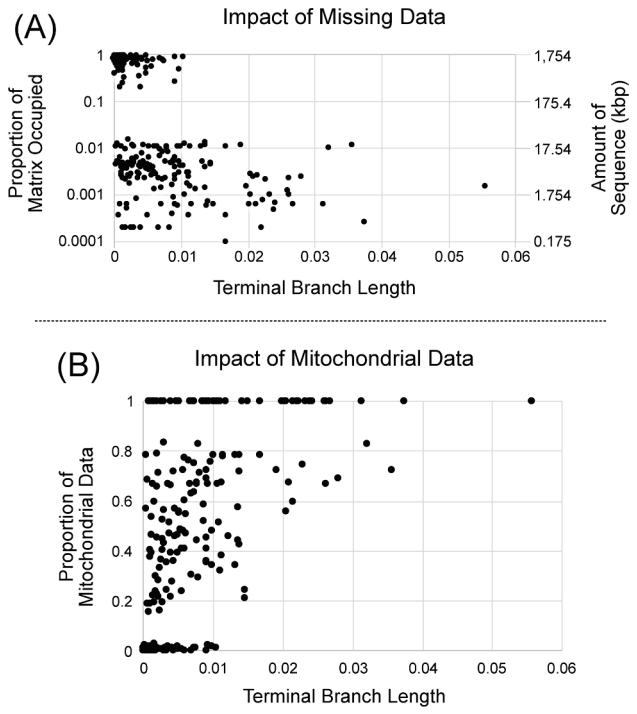

Fig. 4. Relationships with terminal branch lengths. (A) Relationship between the terminal branch length for galliforms and proportion of the matrix occupied (note the log scale for the proportion of matrix occupied). The amount of sequence data in nucleotides (presented in kilobases) is shown to the right of the graph. Note the gap between taxa with UCE data (matrix occupancy > 0.1) and those taxa limited to legacy data (matrix occupancy < 0.1). (B) Relationship between the proportion of mitochondrial data and the terminal branch length. Note that all taxa with UCEs will have a low proportion of mitochondrial data since the upper limit for the mitochondrial data we analyzed is 14.272 kb (this value is limited by the size of the complete mitogenome).

of mitochondrial data. Although the specific loci that are likely to be best represented may vary taxonomically, traditionally phylogenies often relied on a small number of typically overlapping loci that means this strategy may be effective in most or all groups. This also had the advantage of being quick— branch length estimation using unpartitioned mitochondrial genes was far faster than similar analyses using the entire unpartitioned dataset (GTRGAMMA = 10.8 s; PSR = 6.7 s), with both GTRGAMMA and PSR running very efficiently, even when partitioning (e.g., GTRGAMMA = 96.4 s; PSR = 61.0 s).

Divergence times estimated from mitochondrial data do appear to differ from those estimated using nuclear data from the same taxa (e.g., Ksepka and Phillips 2015). These differences are not always consistent: in some cases, mitochondrial sequences lead to more ancient divergences using mitochondrial data, while in others mitochondrial sequences may lead to more recent divergences (e.g., Ksepka and Phillips 2015). Although there may be some bias towards using mitochondrial data for branch length re-estimation, it is likely that the heterogeneous nature of a supermatrix dataset, even when using methods to reduce branch length mis-estimation, might also result in some biases when the complete dataset is used. We have used this approach to estimate branch lengths on supertrees and we obtained divergence time estimates similar to those based on nuclear datasets (Kimball et al. 2019). Thus, we feel that use of just the well-sampled mitochondrial partitions is likely to be a reasonable approach towards estimating branch lengths.

4.3. Conclusions

We have assembled a phylogeny for galliforms that can be used in comparative studies that focus on this group. Many previous comparative studies in this group have used taxon-poor phylogenies that were often biased with extensive sampling in some genera while many genera were missing (e.g., Kimball et al. 2001; Kolm et al. 2007; Nadeau et al. 2007; Krakauer and Kimball 2009; Kimball et al. 2011; Balasubramaniam and Rotenberry 2016; Wang et al. 2017). In addition, galliforms have many nodes that have been problematic to resolve (reviewed in Wang et al. 2013), though the use of UCE data has resulted in stable, highly supported resolution for most of these (e.g., Sun et al. 2014; Meiklejohn et al. 2014; Hosner et al., 2016a,b, 2017).

Funding

Funding was provided by the United States National Science Foundation (grants DEB-1118823 and DEB-1655683 to R.T.K. and E.L.B). P. A.H. has also received support from the Danish National Research Foundation (Center for Macroecology, Evolution, and Climate) and the Villum Foundation (Center for Global Mountain Biodiversity).

CRediT authorship contribution statement

Rebecca T. Kimball: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Funding acquisition. **Peter**

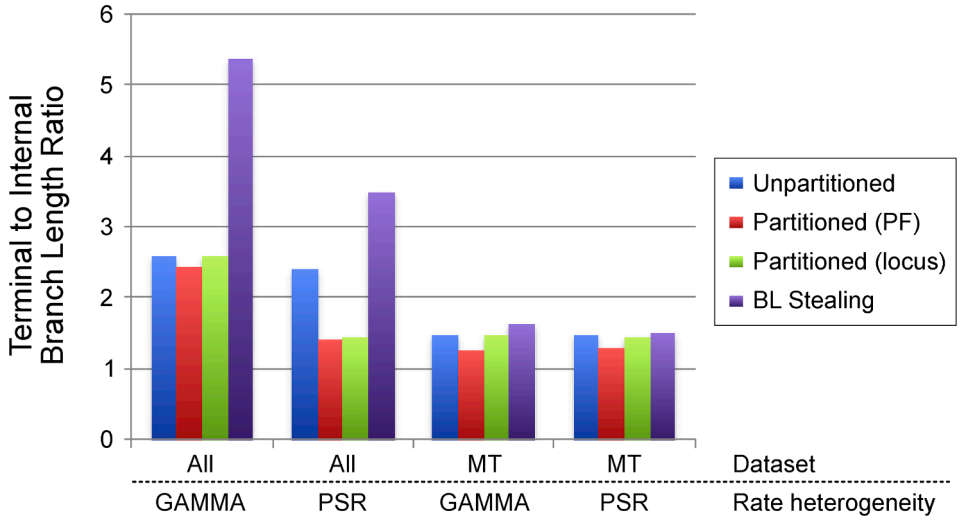

Fig. 5. Ratio of terminal to internal branch lengths after re-estimating branch lengths using the ExaML unpartitioned topology.

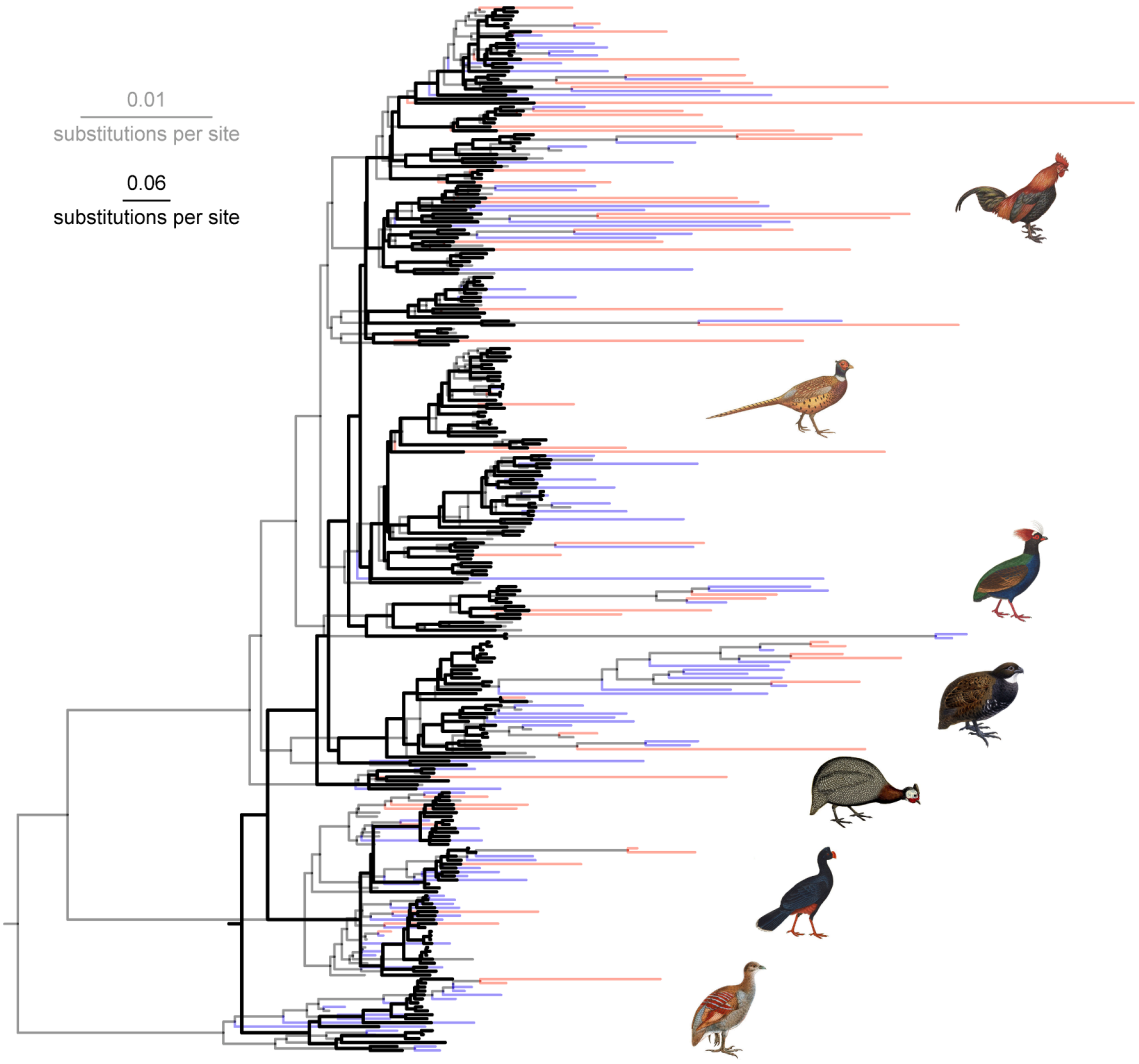

Fig. 6. Branch length estimates for the unpartitioned ExaML topology generated using the best-sampled mitochondrial gene regions (CYB and ND2) are presented using the dark black tree, which is superimposed on the phylogram from the analysis of the complete supermatrix (from Fig. 3). We scaled the terminal branch for *Gallus gallus* to be proportional for both phylograms, although we emphasize that CYB + ND2 treelength is longer than the treelength for the supermatrix (note scale bars).

A. Hosner: Formal analysis, Investigation, Data curation, Writing - review & editing. **Edward L. Braun:** Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Funding acquisition.

Acknowledgements

We thank Jorden Holland and Emily Griffith for assistance with our efforts to compile data, and Sarah Kurtis for making ultrametric trees. We would like to thank the many natural history museums that have supported our efforts to sequence galliforms over the years. We appreciate comments from two anonymous reviewers who helped us improve this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2021.107091.

References

Balasubramaniam, P., Rotenberry, J.T., 2016. Elevation and latitude interact to drive life-history variation in precocial birds: a comparative analysis using galliformes. J. Anim. Ecol. 85, 1528–1539.
Bininda-Emonds, O.R.P., 2004. The evolution of supertrees. Trends Ecol. Evol. 19, 315–322.
Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L., Purvis, A., 2007. The delayed rise of present-day mammals. Nature 446, 507–512.
Bininda-Emonds, O.R.P., Gittleman, J.L., Purvis, A., 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). Biol. Rev. 74, 143–175.
Braun, E.L., 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: Kraus, R. H.S. (Ed.), Avian Genomics in Ecology and Evolution - From the Lab into the Wild. Springer, Switzerland, pp. 151–210.
Brown, J.W., Wang, N., Smith, S.A., 2017. The development of scientific consensus: Analyzing conflict and concordance among avian phylogenies. Mol. Phylogenet. Evol. 116, 69–77.
Burleigh, J.G., Driskell, A.C., Sanderson, M.J., 2006. Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. Syst. Biol. 55, 426–440.
Burleigh, J.G., Kimball, R.T., Braun, E.L., 2015. Building the avian tree of life using a large-scale, sparse supermatrix. Mol. Phylogenet. Evol. 84, 53–63.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Cotton, J.A., Wilkinson, M., 2009. Supertrees join the mainstream of phylogenetics. Trends Ecol. Evol. 24, 1–3.
Crowe, T.M., Bowie, R.C.K., Bloomer, P., Mandiwana, T.G., Hedderson, T.A.J., Randi, E., Pereira, S.L., Wakeling, J., 2006. Phylogenetics, biogeography and classification of, and character evolution in gamebirds (Aves: Galliformes): Effects of character exclusion, data partitioning and missing data. Cladistics 22, 1–38.
Darriba, D., Weiss, M., Stamatakis, A., 2016. Prediction of missing sequences and branch lengths in phylogenomic data. Bioinformatics 32, 1331–1337.
Davison, G.W.H., 1985. Avian spurs. J. Zool. 206, 353–366.
de Queiroz, A., Gatesy, J., 2007. The supermatrix approach to systematics. Trends Ecol. Evol. 22, 34–41.
Driskell, A.C., Ane, C., Burleigh, J.G., McMahon, M.M., O'Meara, B.C., Sanderson, M.J., 2004. Prospects for building the tree of life from large sequence databases. Science 306, 1172–1174.
Dufort, M.J., 2016. An augmented supermatrix phylogeny of the avian family Picidae reveals uncertainty deep in the family tree. Mol. Phylogenet. Evol. 94, 313–326.
Eo, S.H., Binnida-Emonds, O.R.P., Carroll, J.P., 2009. A phylogenetic supertree of the fowls (Galloanserae, Aves). Acta Zoologica Scripta 38, 465–481.
Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61, 717–726.
Fulton, T.L., Strobeck, C., 2006. Molecular phylogeny of the Arctoidea (Carnivora): Effect of missing data on supertree and supermatrix analyses of multiple gene data sets. Mol. Phylogenet. Evol. 41, 165–181.
Gatesy, J., Baker, R.H., 2005. Hidden likelihood support in genomic data: Can forty-five wrongs make a right? Syst. Biol. 54, 483–492.
Gatesy, J., Matthee, C., DeSalle, R., Hayashi, C., 2002. Resolution of a supertree/supermatrix paradox. Syst. Biol. 51, 652–664.
Goloboff, P.A., Catalano, S.A., Miranda, J.M., Szumik, C.A., Arias, J.S., Kallersjo, M., Farris, J.S., 2009. Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups. Cladistics 25, 211–230.
Hackett, S.J., Kimball, R.T., Reddy, S., Bowie, R.C.K., Braun, E.L., Braun, M.J., Chojnowski, J.L., Cox, W.A., Han, K.-L., Harshman, J., Huddleston, C.J., Marks, B., Miglia, K.J., Moore, W.S., Sheldon, F.H., Steadman, D.W., Witt, C.C., Yuri, T., 2008. A phylogenomic study of birds reveals their evolutionary history. Science 320, 1763–1768.
Hosner, P.A., Braun, E.L., Kimball, R.T., 2015. Land connectivity changes and global cooling shaped the colonization history and diversification of New World quail (Aves: Galliformes: Odontophoridae). J. Biogeogr. 42, 1883–1895.
Hosner, P.A., Braun, E.L., Kimball, R.T., 2016a. Rapid and recent diversification of curassows, guans, and chachalacas (Galliformes: Cracidae) out of Mesoamerica: Phylogeny inferred from mitochondrial, intron, and ultraconserved element sequences. Mol. Phylogenet. Evol. 102, 320–330.
Hosner, P.A., Faircloth, B.C., Glenn, T.C., Braun, E.L., Kimball, R.T., 2016b. Avoiding missing data biases in phylogenomic inference: An empirical study in the landfowl (Aves: Galliformes). Mol. Biol. Evol. 33, 1110–1125.
Hosner, P.A., Owens, H.L., Braun, E.L., Kimball, R.T., 2020. Phylogeny and diversification of the gallopheasants (Aves: Galliformes): Testing roles of sexual selection and environmental niche divergence. Zool. Scr. 49, 549–562.
Hosner, P.A., Tobias, J.A., Braun, E.L., Kimball, R.T., 2017. How do seemingly non-vagile clades accomplish trans-marine dispersal? Trait and dispersal evolution in the landfowl (Ayes: Galliformes). Proc. R. Soc. B 284, 20170210.
Jetz, W., Thomas, G.H., Joy, J.B., Hartmann, K., Mooers, A.Ø., 2012. The global diversity of birds in space and time. Nature 491, 444–448.
Katoh, K., Misawa, K., Kuma, K.-I., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066.
Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30, 772–780.
Kim, S., 2015. ppcor: An R Package for a fast calculation to semi-partial correlation coefficients. Commun Stat Appl Met 22, 665–674.
Kimball, R.T., Braun, E.L., 2008. A multigene phylogeny of Galliformes supports a single origin of erectile ability in non-feathered facial traits. J. Avian Biol. 39, 438–445.
Kimball, R.T., Braun, E.L., 2014. Does more sequence data improve estimates of galliform phylogeny? Analyses of a rapid radiation using a complete data matrix. Peerj 2, e361.
Kimball, R.T., Braun, E.L., Ligon, J.D., Lucchini, V., Randi, E., 2001. A molecular phylogeny of the peacock-pheasants (Galliformes : Polyplectron spp.) indicates loss and reduction of ornamental traits and display behaviours. Biol. J. Linn. Soc. 73, 187–198.
Kimball, R.T., St. Mary, C.M., Braun, E.L., 2011. A macroevolutionary perspective on multiple sexual traits in the Phasianidae (Galliformes). Int. J. Evol. Biol. 2011, 423938.
Kimball, R.T., Oliveros, C.H., Wang, N., White, N.D., Barker, F.K., Field, D.J., Ksepka, D. T., Chesser, R.T., Moyle, R.G., Braun, M.J., Brumfield, R.T., Faircloth, B.C., Smith, B. T., Braun, E.L., 2019. A phylogenomic supertree of birds. Diversity-Basel 11, 109.
Kohany, O., Gentles, A.J., Hankus, L., Jurka, J., 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7, 474.
Kolm, N., Stein, R.W., Mooers, A.Ø., Verspoor, J.J., Cunningham, E.J.A., 2007. Can sexual selection drive female life histories? A comparative study on Galliform birds. J. Evol. Biol. 20, 627–638.
Kozlov, A.M., Aberer, A.J., Stamatakis, A., 2015. ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics 31, 2577–2579.
Krakauer, A.H., Kimball, R.T., 2009. Interspecific brood parasitism in galliform birds. Ibis 151, 373–381.
Ksepka, D.T., Balanoff, A.M., Smith, N.A., Bever, G.S., Bhullar, B.S., Bourdon, E., Braun, E.L., Burleigh, J.G., Clarke, J.A., Colbert, M.W., Corfield, J.R., Degrange, F.J., De Pietri, V.L., Early, C.M., Field, D.J., Gignac, P.M., Gold, M.E.L., Kimball, R.T., Kawabe, S., Lefebvre, L., Marugan-Lobon, J., Mongle, C.S., Morhardt, A., Norell, M. A., Ridgely, R.C., Rothman, R.S., Scofield, R.P., Tambussi, C.P., Torres, C.R., van Tuinen, M., Walsh, S.A., Watanabe, A., Witmer, L.M., Wright, A.K., Zanno, L.E., Jarvis, E.D., Smaers, J.B., 2020. Tempo and Pattern of Avian Brain Size Evolution. Curr Biol. 30, 2026–2036.e3.
Ksepka, D.T., Phillips, M.J., 2015. Avian Diversification Patterns across the K-Pg Boundary: Influence of Calibrations, Datasets, and Model Misspecification. Ann. Missouri Bot. Gard. 100, 300–328.
Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T., Calcott, B., 2017. PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34, 772–773.
Lartillot, N., Philippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21, 1095–1109.
Lislevand, T., Figuerola, J., Szekely, T., 2009. Evolution of sexual size dimorphism in grouse and allies (Aves: Phasianidae) in relation to mating competition, fecundity demands and resource division. J. Evol. Biol. 22, 1895–1905.
Litsios, G., Salamin, N., 2012. Effects of phylogenetic signal on ancestral state reconstruction. Syst. Biol. 61, 533–538.
Marcondes, R.S., 2019. Realistic scenarios of missing taxa in phylogenetic comparative methods and their effects on model selection and parameter estimation. PeerJ 7, e7917.
McEntee, J.P., Tobias, J.A., Sheard, C., Burleigh, J.G., 2018. Tempo and timing of ecological trait divergence in bird speciation. Nat. Ecol. Evol. 2, 1120–1127.
McGowen, M.R., Spaulding, M., Gatesy, J., 2009. Divergence date estimation and a comprehensive molecular tree of extant cetaceans. Mol. Phylogenet. Evol. 53, 891–906.
Meiklejohn, K.A., Danielson, M.J., Faircloth, B.C., Glenn, T.C., Braun, E.L., Kimball, R.T., 2014. Incongruence among different mitochondrial regions: A case study using complete mitogenomes. Mol. Phylogenet. Evol. 78, 314–323.
Meiklejohn, K.A., Faircloth, B.C., Glenn, T.C., Kimball, R.T., Braun, E.L., 2016. Analysis of a rapid evolutionary radiation using ultraconserved elements (UCEs): Evidence for a bias in some multi-species coalescent methods. Syst. Biol. 65, 612–627.
Nadeau, N.J., Burke, T., Mundy, N.I., 2007. Evolution of an avian pigmentation gene correlates with a measure of sexual selection. Proc. R. Soc. B 274, 1807–1813.
Paradis, E., Schliep, K., 2019. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35, 526–528.
Persons, N.W., Hosner, P.A., Meiklejohn, K.A., Braun, E.L., Kimball, R.T., 2016. Sorting out relationships among the grouse and ptarmigan using intron, mitochondrial, and ultra-conserved element sequences. Mol. Phylogenet. Evol. 98, 123–132.
Pulido-Santacruz, P., Weir, J.T., 2016. Extinction as a driver of avian latitudinal diversity gradients. Evolution 70, 860–872.
Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.
Salisbury, B.A., Kim, J.H., 2001. Ancestral state estimation and taxon sampling density. Syst. Biol. 50, 557–564.
Sanderson, M.J., Purvis, A., Henze, C., 1998. Phylogenetic supertrees: assembling the trees of life. Trends Ecol. Evol. 13, 105–109.
Shakya, S.B., Sheldon, F.H., 2017. The phylogeny of the world's bulbuls (Pycnonotidae) inferred using a supermatrix approach. Ibis 159, 498–509.
Stamatakis, A., 2006. Phylogenetic models of rate heterogeneity: A high performance computing perspective. Proceedings of 20th IEEE/ACM International Parallel and Distributed Processing Symposium (IPDPS2006), High Performance Computational Biology Workshop, Proceedings on CD, Rhodos, Greece, April 2006.
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.
Stamatakis, A., Aberer, A.J., 2013. Novel parallelization schemes for large-scale likelihood-based phylogenetic inference. 2013 IEEE 27th International Symposium on Parallel and Distributed Processing. IEEE, pp. 1195-1204.
Stein, R.W., Brown, J.W., Mooers, A.Ø., 2015. A molecular genetic time scale demonstrates Cretaceous origins and multiple diversification rate shifts within the order Galliformes (Aves). Mol. Phylogenet. Evol. 92, 155–164.
Swofford, D.L. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
Sun, K.P., Meiklejohn, K.A., Faircloth, B.C., Glenn, T.C., Braun, E.L., Kimball, R.T., 2014. The evolution of peafowl and other taxa with ocelli (eyespots): A phylogenomic approach. Proc. R. Soc. B 281, 20140823.
Team, R.C., 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Torices, R., 2010. Adding time-calibrated branch lengths to the Asteraceae supertree. J Syst Evol 48, 271–278.
Wang, N., Kimball, R.T., Braun, E.L., Liang, B., Zhang, Z., 2013. Assessing phylogenetic relationships among Galliformes: A multigene phylogeny with expanded taxon sampling in Phasianidae. PLoS ONE 8, e64312.
Wang, N., Kimball, R.T., Braun, E.L., Liang, B., Zhang, Z.W., 2017. Ancestral range reconstruction of Galliformes: the effects of topology and taxon sampling. J. Biogeogr. 44, 122–135.
Warnow, T., 2018. Supertree construction: Opportunities and challenges. arXiv, 1805.03530.
Webb, C.O., Ackerly, D.D., Kembel, S.W., 2008. Phylocom: software for the analysis of phylogenetic community structure and trait evolution. Bioinformatics 24, 2098–2100.