A peer-reviewed version of this preprint was published in PeerJ on 22 April 2014.

View the peer-reviewed version (peerj.com/articles/361), which is the preferred citable publication unless you specifically need to cite this preprint.

Kimball RT, Braun EL. 2014. Does more sequence data improve estimates of galliform phylogeny? Analyses of a rapid radiation using a complete data matrix. PeerJ 2:e361 https://doi.org/10.7717/peerj.361

Does more sequence data improve estimates of galliform phylogeny? Analyses of a rapid radiation using a complete data matrix.

Rebecca T. Kimball and Edward L. Braun

Department of Biology, University of Florida, P.O. Box 118525, Gainesville, FL 32611 USA

Corresponding Author: Rebecca T. Kimball

Department of Biology

P.O. Box 118525

University of Florida

Gainesville, FL 32611

rkimball@ufl.edu

Phone: 1-352-846-3737

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

Abstract

The resolution of rapid evolutionary radiations or "bushes" in the tree of life has been one of the most difficult and interesting problems in phylogenetics. The avian order Galliformes appears to have undergone several rapid radiations that have limited the resolution of prior studies and obscured the position of taxa important both agriculturally and as model systems (chicken, turkey, Japanese quail). Here we present analyses of a multi-locus data matrix comprising over 15,000 sites, primarily from nuclear introns but also including three mitochondrial regions, from 46 galliform taxa with all gene regions sampled for all taxa. The increased sampling of unlinked nuclear genes provided strong bootstrap support for all but a small number of relationships. Coalescent-based methods to combine individual gene trees and analyses of datasets independent of published data indicated that this well-supported topology is likely to reflect the galliform species tree. Some of the key findings include support for a second major clade within the core phasianids that includes the chicken and Japanese quail and clarification of the phylogenetic relationships of turkey. Jackknifed datasets suggested that there is an advantage to sampling many independent regions across the genome rather than obtaining long sequences for a small number of loci, possibly reflecting the differences among gene trees that differ due to incomplete lineage sorting. Despite the novel insights we obtained using this increased sampling of gene regions, some nodes remain unresolved, likely due to periods of rapid diversification. Resolving these remaining groups will likely require sequencing a very large number of gene regions, but our analyses now appear to support a robust backbone for this order.

Keywords – Galliformes; rapid radiation; sampling strategies; data matrix size

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

Introduction

Continuing improvements in data collection and methods of phylogenetic analyses have revolutionized our understanding of evolutionary relationships (Delsuc, Brinkmann & Philippe, 2005). However, many relationships in the tree of life remain in question despite intensive study. These difficult to resolve relationships, sometimes called "bushes" in the tree of life (Rokas and Carroll, 2006), are thought to reflect rapid evolutionary radiations. Some of these cases may reflect true hard polytomies (simultaneous speciation events), though others may represent soft polytomies that could be resolved if datasets of sufficient size are collected and analyzed appropriately. Questions remain about the best methods to resolve these difficult problems (e.g., Patel, Kimball & Braun, 2013). While taxon sampling may improve phylogenetic resolution in some cases, in others more data per species has to potential to yield greater improvements to resolution than adding in data from additional species (reviewed by Nabhan and Sarkar, 2012).

The avian order Galliformes, informally called gamebirds or landfowl, includes some of the best-studied avian species. This includes the economically important chicken and turkey, the first two avian genomes sequenced (Dalloul et al., 2010; Hillier et al., 2004), as well as other important model systems such as the Japanese quail (e.g., Poynter, Huss & Lansford, 2009). As such, there exists a large body of literature about galliform physiology, reproduction, genetics, development and behavior. This order has also been extensively studied phylogenetically, and recent studies have typically separated the order into five families (reviewed in Wang et al., 2013). However, relationships within the largest family, the Phasianidae (which includes the key galliform model systems), have remained problematic. Indeed, the placement of some key phasianid taxa has varied among recent studies. This is likely to reflect the existence of several very short internodes within Phasianidae (e.g., Kimball, St. Mary & Braun, 2011), suggesting

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

that rapid radiations may have occurred and that these radiations are likely to have led to some of the difficulties in resolving relationships within this group. Analysis of supermatrices, constructed either by combining sequence data from multiple gene regions (Kimball, St. Mary & Braun, 2011) or by combining sequence data with morphological and behavioral traits (Crowe et al., 2006), also exhibited limited support for a number of relationships. Those studies included more taxa than previous studies but they used data matrices with substantial amounts of missing data. To explore the utility of taxon sampling with a more consistent sampling of genes, Wang et al. (2013) sequenced six nuclear loci and two mitochondrial regions for 88 galliforms. This improved the support for some relationships and clarified the position of several previously poorly sampled lineages, but it provided limited improvement to the backbone phylogeny of the Phasianidae. These remaining unresolved and poorly supported nodes make it difficult to place the large body of literature about galliforms into an evolutionary framework.

Simulations based on parameters estimated from avian data demonstrate that increasing partition size leads to more accurate resolution of short internodes and that analyses of introns yields better results (with smaller amounts of sequence data) than that of exons (Chojnowski, Kimball & Braun, 2008). Those simulations were focused on the power of a single locus to resolve gene trees rather than the power of a multi-locus dataset to resolve a species tree. However, it seems reasonable to postulate that sequencing additional unlinked loci will improve phylogenetic resolution. It is also likely to prove beneficial to include mitochondrial sequence data in such a data matrix. Although phylogenetic analyses using avian mitochondrial sequences have proven to be sensitive to model selection and taxon sampling when applied to deep branches in the avian tree (Braun and Kimball, 2002; Pratt et al., 2009), the smaller population size for the mitochondrial genome increases the probability that the mitochondrial gene tree will

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

match the species tree (Moore, 1995). Indeed, including both multiple unlinked nuclear loci along with mitochondrial sequences has been shown to improve phylogenetic estimation under some circumstances (Corl and Ellegren, 2013; Sanchez-Garcia and Castresana, 2012). On the other hand, Kimball et al. (2013) found that increasing the number of sites in a supermatrix nearly two-fold to examine higher-level avian relationships did not improve bootstrap support or resolve additional nodes, suggesting there may be limits to the improvement that can be obtained with larger datasets (although truly genome-scale datasets have been analyzed for only a few clades in the tree of life). This raises the question of whether additional data could substantially improve resolution for the rapid radiations in the galliforms.

Here we collected data from 15 nuclear loci and three mitochondrial regions to generate a data matrix comprising more than 15,000 sites and examined whether sampling these additional loci provided improved resolution and support within the galliforms. Our taxon sampling was focused on the Phaisanidae and specifically targeted to include several difficult to resolve parts of the galliform tree (i.e., relationships defined by short internodes). However, we also included representatives of other galliform families to provide outgroups. The data matrix was fully sampled, such that each species was represented by sequence data for each nuclear intron and mitochondrial region. Thus, missing data could not have an impact upon our conclusions. First, we used this dataset to determine whether we could improve our understanding of relationships among difficult nodes within the Phasianidae both by conducting analyses of concatenated sequences and using methods that incorporate the multispecies coalescent. Second, we also analyzed a subset of these taxa to allow direct comparisons to an earlier study (Kimball and Braun, 2008) that used fewer loci to directly assess the impact of adding more sequence data per species on bootstrap support. Finally, we generated jackknifed datasets of varying size from the

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

concatenated data to further explore the impact of increasing dataset size on phylogenetic reconstruction in galliforms.

Methods

Molecular methods

The species included in this analysis represent a range of galliform lineages, with an emphasis on the main phasianid lineages where there has been conflict among recent studies. We combined published sequences from our earlier studies (Armstrong, Braun & Kimball, 2001; Cox et al., 2007; Kimball and Braun, 2008; Kimball et al., 1999, 2001; Randi et al., 2000, 2001; Wang et al., 2013) with novel sequences collected as part of this study. This study focuses on the same samples included in Kimball and Braun (2008), since that study included broad sampling of most of the Phasianidae, including clades that appear likely to have undergone a rapid radiation, as well as a set of taxa from the other galliform families. However, Kimball and Braun (2008) lacked a representative of one key phasianid clade, the Arborophilinae. It also did not include the argus pheasant, which was placed in an unexpected position in the Kimball, St. Mary & Braun (2011) supermatrix analysis. To address these issues, this study included the crested partridge (*Rollulus rouloul*), a representative of the Arborophilinae, and the argus pheasant (*Argusianus argus*) as well as the samples from Kimball and Braun (2008).

Primers for PCR amplification and sequencing conditions are described in Cox et al. (2007), Kimball et al. (2009), Wang et al. (2013), and in Supplemental Material (Table S1). PCR products were amplified using standard protocols, and amplified products were cleaned for sequencing by precipitation using an equal volume of PEG:NaCl (20 %:2.5M). Sequencing of PCR products was done using either ABI BigDye® Terminator v.1.0, BigDye® Terminator v.3.1,

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

or Beckman DTCS Quickstart® chemistries. Manufacturers' recommendations were followed, except reaction volumes were cut to 1/2 - 1/6 of the recommended volume. Sequences were analyzed on an ABI Prism™ 3100-Avant genetic analyzer (PE Applied Biosystems) or a CEQ™ 8000 (Beckman-Coulter™) genetic analysis system. Double-stranded contigs were assembled using Sequencher™ 4.1 (Gene Codes Corp.).

A number of the nuclear loci were heterozygous, and in some cases the alleles differed in size within an individual. These length polymorphisms made it impossible to sequence the PCR products cleanly in both directions. For these samples, PCR products were cloned using the pGEM®-T Easy vector (Promega Corp.). Plasmid preparations were done using Eppendorf Perfectprep® Plasmid Mini kit, and the resulting plasmids were sequenced in each direction. In these cases, the clean portions of the original sequences from the PCR products as well as those from the plasmids were used to assemble the final contig.

Alignment and phylogenetic analysis

Sequences of the mitochondrial coding regions were equal in length and did not have any insertions or deletions, so alignment was straightforward. Preliminary alignments for the nuclear introns and the 12S rRNA region were generated using ClustalX (Thompson et al., 1997) and then optimized by eye. Regions that were difficult to align with confidence were identified and excluded from analyses. Six microinversions (cf. Braun et al., 2011) in the nuclear introns were identified by eye during the alignment process. These were treated as described in Hackett et al. (2008) and excluded from analysis. A large autapomorphic insertion (an ERV transposable element insertion) in FGB intron 7 was also excluded from analyses; two synapomorphic transposable element insertions were included in analyses.

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

We used PAUP* 4.0b10 (Swofford, 2003) to estimate the maximum likelihood (ML) tree on the concatenated dataset, a concatenated nuclear dataset, a concatenated mitochondrial dataset, and each individual locus or mitochondrial region. The appropriate models for ML analyses of the various concatenated matrices and for each independent nuclear locus or mitochondrial region were determined using the Akiake information criterion (AIC) using Modeltest 3.6 (Posada and Crandall, 1998). ML trees were identified using a heuristic search with 10 random sequence additions and the parameters recommended by Modeltest.

Bootstrap support using maximum likelihood was performed for the three concatenated datasets described above, as well as the individual loci or mitochondrial regions, using the GTRGAMMA model in RAxML 7.2.8 (Stamatakis, 2006) with 500 standard bootstrap replicates. Partitioned ML analyses (using each locus or mitochondrial region as a separate partition) of the multi-locus datasets were also conducted using RAxML using 500 bootstrap replicates.

To allow a direct comparison between analyses of this dataset the results of Kimball and Braun (2008), where only four nuclear loci and two mitochondrial regions were examined, we excluded two species (*Argusianus argus* and *Rollulus rouloul*) to obtain a dataset with the same species as included in that study. We also conducted additional analyses using only the data that were not used in Kimball and Braun (2008), which corresponded to 11 additional loci and a single mitochondrial region. This allowed us to obtain an independent assessment of galliform relationships. In both cases, we then obtained an ML tree using PAUP* (Swofford 2003) and a bootstrap consensus tree, using unpartitioned and partitioned analyses, in RAxML (Stamatakis 2006) as described above for the concatenated data for all loci and mitochondrial regions.

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

Bayesian analyses of the concatenated dataset were conducted using MrBayes 3.1.2 (Ronquist and Huelsenbeck, 2003). The appropriate model for each partition (locus or mitochondrial region) was determined using the AIC (limiting the set of models under consideration to those implemented in MrBayes). We used two simultaneous searches with four chains each (three heated chains and one cold chain) that were run for, 20 million generations, sampling every 1000 generations and discarding the first 2,000,000 generations as "burn-in". The runs appeared to converge, as the harmonic means of the different runs were similar, the posterior scale reduction factors were essentially 1, and the average deviations of the split frequencies were substantially smaller than 0.01.

We estimated a species tree from individual gene trees using two approaches. First, we used BUCKy 1.4.0 (Larget et al., 2010) to estimate both the primary concordance tree and the population (species) tree. To obtain the input trees, we used MrBayes 3.1.2 (Ronquist and Huelsenbeck, 2003) on the CIPRES Science Gateway (Miller, Pfeiffer & Schwartz, 2010) for each nuclear locus or for the concatenated mitochondrial data. The appropriate model for analyses was determined by using the Akiake Information Criterion in Modeltest 3.6 (Posada and Crandall, 1998), limiting our consideration to those models implemented in MrBayes. Since the mitochondrion is a single locus, we concatenated the sequences but partitioned by mitochondrial region in the analyses. Searches were run as described above for the concatenated dataset. Our second approach was to use NJ_{st} (Liu and Yu, 2011) as implemented on the Species Tree Webserver (Shaw et al., 2013). To accommodate uncertainty in the gene tree estimates, we used the 500 bootstrap replicates from RAxML for each of the individual nuclear loci and combined mitochondrial data (see above for details).

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

We generated datasets of varying lengths by jackknifing to explore the relationship between sequence length and the power to resolve relationships. Since the mitochondrial genome is limited in size and is thought to form a single non-recombining locus (Berlin and Ellegren, 2001; Berlin, Smith & Ellegren, 2004), obtaining larger and larger datasets will by necessity involve sampling from the nuclear genome. Thus, we only used the nuclear sequence data to generate the jackknifed datasets. This also avoided the much more rapidly evolving sites in the mitochondrial genome. We generated 100 jackknifed datasets that were 450, 550, 650, 750, 1000, 1100, 1650, 2000, 5000, 8000, and 11,000 bp (the alignment of the nuclear loci, after excluding the sites indicated above, was 12731 bp in length). This led to jackknifed datasets similar in length to the individual loci or mitochondrial regions (to allow comparison with those) as well as longer datasets to explore the affect of increasing the amount of data. For each dataset, we estimated the ML tree in PAUP*, using the same approach as described above. We then examined differences between the ML tree estimated from the complete nuclear data matrix and the trees estimated from the jackknifed datasets as well as the individual partitions by calculating RF distances (Robinson and Foulds, 1981).

Finally, we explored the differences among individual loci by comparing the RAxML bootstrap results of individual loci (see above) among several clades that received high bootstrap support (97-100%) but very different concordance factors (see below). Several of these nodes have been well-supported in a range of recent galliform phylogenies (e.g., Bonilla, Braun, & Kimball, 2010; Cohen et al., 2012; Cox et al., 2007; Crowe et al., 2006; Kaiser, van Tuinen & Ellegren, 2007; Kriegs et al., 2007; Meng et al., 2008; Nadeau, Burke & Mundy, 2007; Shen et al., 2010), including those defining the “erectile” clade (A), the “core” Phasianidae (all phasianids except the Arborophilinae, which is represented by *Rollulus* in this study) (B),

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

Phasianidae (C), and the clade comprising Phasianidae and Odontophoridae (excluding other galliform families) (D). We included two other clades that received strong bootstrap support that have not been recovered consistently in recent studies that we also examined. One of these is the clade comprising the Argus pheasants (represented here by *Argusianus*) and the peafowl (*Pavo* and *Afropavo*) (E), and the other is a clade comprising the phasianids that are neither included in the erectile clade nor Arborophilinae (this "non-erectile clade" forms a grade in many studies) (F).

Results and Discussion

The nuclear loci sequenced were distributed on 12 chromosomes in the chicken genome, including macro- and microchromosomes, providing both a broad sampling of the genome and sampling of loci that likely differ in patterns of molecular evolution (Axelsson et al., 2005). When the 13970 sites from the nuclear genome were combined with the 3265 sites obtained from the mitochondrial genome a total evidence data matrix of 17235 sites was obtained. The majority of sites from the nuclear genome were non-coding, with only, 208 sites from coding exons (two amplicons spanned two introns and the intervening exons were sequenced for these loci). After we excluded the hard to align regions, the microinversions, and a long autapomorphic TE insertion in FGB (a total of 1564 sites) there were 15,866 sites used in analyses (12,731 nuclear and 3135 mitochondrial). Synapomorphic TE insertions are most likely to be found on long, and uncontroversial branches (e.g., Han et al., 2011), as was true with the synapomorphic TE insertions that were included in the analyses. The mitochondrial data had a greater proportion of parsimony informative sites (41%) than the nuclear data (36%), though there were slightly fewer variables sites in the mitochondrial data (48%) relative to the nuclear data (52%).

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

Relationships within galliforms based upon mitochondrial data and 15 nuclear loci

The total evidence partitioned ML tree (Fig. 1) had relatively high support at most nodes, particularly many of the nodes that define relationships among genera. There were some topological differences among the analyses. For example, the unpartitioned and partitioned ML analyses had three conflicting branches (RF distance = 6). However, all of these differences involved rearrangements within genera (*Gallus*, *Lophura* and *Polyplectron*). Topologically, the consensus tree from the Bayesian MCMC analysis was almost identical to the partitioned ML tree; the only difference being a rearrangement within *Polyplectron* (Fig. 1) that was not highly supported (a posterior probability of 0.94). Only one other node received a posterior probability less than 1.0, and it was also within *Polyplectron* (both of these poorly supported nodes received less than 70% support in the bootstrap analyses). Thus, outside of some conflicts among recently diverged taxa, the total evidence phylogeny was well supported and consistent among analyses.

The partitioned and unpartitioned ML trees estimated using the nuclear data were topologically identical (Supplementary Information). However, partitioning the mitochondrial data (by gene region) resulted in several differences when compared with the unpartitioned mitochondrial topology (Supplementary Information). This included two differences that were within genera (*Gallus* and *Polyplectron*) while a third difference involved relationships between *Phasianus* and *Chrysolophus*. The partitioned total evidence tree was much more similar to the tree estimated from the nuclear-only data than the mitochondrial-only data (RF distance = 2 versus 18), probably reflecting, at least in part, the fact that there was nearly 4-fold more sites in the nuclear dataset.

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

Both estimates of the species tree from individual gene trees (Fig. 2) – the BUCKy population tree (which should provide an unbiased estimate of the species tree if discordance among gene trees reflects the multispecies coalescent) and the NJ_{st} tree – were very similar. In fact, they differed at just a single node (within *Polyplectron*). The species trees were similar to the total evidence trees obtained by analysis of the concatenated data, being more similar to the partitioned (RF distance to the BUCKy tree = 8) than the unpartitioned (RF distance to the BUCKy tree = 14) tree. All differences to the total evidence partitioned ML topology were within the Phasianidae, although they did involve rearrangements within genera, among genera, and among higher-level clades (see circled nodes in Fig. 2). Most of the differences were at nodes with low bootstrap support in the NJ_{st} analyses, had low concordance factors (Fig. 2) and/or were absent in the primary concordance tree. These problematic relationships also corresponded to relationships that have varied among recent studies (reviewed in Wang et al., 2013) and thus represent some of the more challenging nodes within the Phasianidae that may require substantially more data from unlinked genes to resolve with confidence.

We identified a set of relationships within the Phasianidae that appeared to be robustly supported by this data (Fig. 3) by excluding those nodes that were poorly supported (those with less than 70% bootstrap support or 0.95 posterior probability) and those that were incongruent between the total evidence (Fig. 1) and species trees (Fig. 2). This set of relationships divided the "core phasianids" (the members of Phasianidae excluding Arborophilinae) into two clades: the erectile clade (Braun and Kimball 2008) and a non-erectile clade. The erectile clade has been recovered in a number of studies and is often recovered with high support in analyses of individual loci (Kimball and Braun 2008). Moreover, the use of better-fitting models increased support for this clade in analyses of a region (CYB) that does not support the clade in many

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

analyses (Kimball et al. 2006). In contrast, the non-erectile clade was not recovered in many prior studies (reviewed by Wang et al. 2013). When it has been recovered it was either poorly supported or its recovery varied among analyses [in Wang et al. (2013), analysis of concatenated data recovered the clade whereas the estimate of the species tree obtained using NJ_{st} did not]. In this study, this major clade received >80% bootstrap support in both the species tree analysis and analyses of concatenated data. This provides strong support for the existence of a clade (the non-erectile clade) including both of the important galliform model systems [the chicken (*Gallus gallus*) and Japanese quail (*Coturnix japonica*)] and separating them from the agriculturally important turkey (*Meleagris gallopavo*) in the erectile clade.

Support for relationships within these two major clades varied. Within the erectile clade, relationships among genera were largely resolved (with the exception the positions of *Phasianus* and *Chrysolophus*). Our analyses found strong support for uniting *Meleagris* with the grouse, and those taxa with the koklass pheasant (*Pucrasia macrolopha*); the sister group of the turkey has been variable and the support for whatever relationship was found has been limited in previous studies (reviewed in Wang et al. 2013). In the non-erectile clade, however, there were four well-defined clades but relationships among those four clades remained problematic. The other poorly resolved nodes (represented by polytomies in Fig. 3) were all within genera (*Lophura*, *Gallus*, and two in *Polyplectron*). In every genus where three or more species were sampled, at least some of the relationships were poorly supported. Rearrangements within these genera represented many of the differences among analyses of the different data partitions (total evidence, nuclear and mitochondrial) as well as the analysis of concatenated data and estimates of the species tree using methods that incorporated the multispecies coalescent. The difficulty in resolving these problematic relationships likely reflect a combination of relatively rapid

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

radiations leading to short internodes that did not allow for sufficient substitutional variation to accumulate, and gene tree discordance due to lineage sorting. However, other processes (i.e., problems estimating the gene tree due to patterns of molecular evolution) may also contribute. The amount of additional data that will be needed to establish these relationships with confidence, if resolution is possible, is unclear.

The impact of larger datasets

This dataset included more loci than other previously published galliform phylogenies (e.g., Bonilla, Braun, & Kimball, 2010; Kimball and Braun, 2008; Nadeau, Burke, & Mundy, 2007; Wang et al., 2013), but, like those other studies, there are still some nodes that lacked support. Simulations generally show that adding data initially results in a rapid approach to the true tree followed by a more gradual improvement as data continues to be added (e.g., Chojnowski, Kimball & Braun, 2008). This raises the issue of whether increasing the size of the dataset that we report here has resulted in appreciable improvements to the phylogenetic resolution within Galliformes. Specifically, are we in the relatively rapid phase of improvement or the more gradual phase?

To address this question with a direct comparison between this study and a study using less data, we conducted a partitioned ML bootstrap using the dataset assembled in this study restricting the taxa analyzed to the 44 taxa included in Kimball and Braun (2008), which included four nuclear loci and two mitochondrial regions (Supplementary Information). In general, increasing the dataset approximately 3-fold (5533 sites in Kimball and Braun [2008] compared to 15866 in this study) resulted in higher bootstrap support for a number of nodes. However, the differences were modest. There were seven nodes that increased by 5% or more in

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

bootstrap support in the larger dataset (including two that were present but had less than 50% in Kimball and Braun [2008]), while only three nodes decreased by 5% or more. Several nodes differed, all involving relationships that remained problematic (see shaded nodes in Fig. 3).

Increasing the number of loci has also been suggested to improve analyses of species trees (e.g., Corl and Ellegren, 2013; McCormack, Huang & Knowles, 2009). Kimball and Braun (2008) had only analyzed concatenated data, and so the results of that study could not be compared to the species tree estimated assuming the multispecies coalescent (Fig. 2). Instead, we estimated a species tree in NJ_{st} using the loci included in Kimball and Braun (2008), but with the same bootstrap input trees used in Figure 2. This corresponded to the comparison of a tree based upon 46 taxa and 16 loci (the NJ_{st} shown in Fig. 2) to one with the same taxa but only five loci (Supplementary Information). Increasing the number of loci resulted in an increase of at least 5% bootstrap for six nodes and a decrease of at least 5% for only one node. Two nodes that received moderate to high (73 and 98%) bootstrap support in the 16-locus analysis were unresolved (e.g., received less than 50% bootstrap support) in the 5-locus species tree. As with the comparison using concatenated data, there were five nodes that were topologically different (all of these were within genera). Inclusion of a mitochondrial partition has been suggested to lead to a greater improvement relative to the addition of a nuclear locus (Corl and Ellegren, 2013), so we also compared the 16-locus species tree with a 15-locus tree where the mitochondrial partition was excluded. While exclusion of the mitochondrial data reduced 4 nodes by at least 5% bootstrap support, 3 other nodes increased by at least 5%. Thus, the inclusion of mitochondrial data appeared to have a modest impact upon our coalescent-based analyses. Overall, the level of improvement with the inclusion of additional data for the species tree analyses was modest and similar to that obtained in analyses of the concatenated data.

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

We also explored the impact of increasing dataset size by creating jackknifed datasets of various different sizes, and comparing the ML tree from those datasets with our partitioned ML tree. As expected, increasing the numbers of base pairs in the dataset led to ML trees that were closer to the total evidence tree (Fig. 4A). However, the degree of improvement decreased as data were added and the benefits to increasing the number of loci quickly become quite limited. Thus, the relatively modest improvements that we observed are consistent with the increase in dataset size form that analyzed in Kimball and Braun (2008) to that examined here. These results suggest it may require substantially greater amounts of data (relative to the almost three-fold increase reported here) to resolve these additional nodes with confidence.

Localized biases – insights from independent evidence?

Analyzing multi-locus data matrices has been shown to result in the recovery of many strongly supported clades, even for difficult problems (e.g., Dunn et al., 2008; Hackett et al., 2008). However, there are instances where one or a few genes exhibit strong localized biases that do not reflect evolutionary history. In some cases they may reflect well-understood phenomena such as convergence in base composition (e.g., Katsu et al., 2009) while in others the biological basis for the localized bias is more obscure (e.g., Kimball et al., 2013). Regardless of their basis, these biases can affect phylogenetic signal in idiosyncratic ways. In fact, previous studies have even established that conflicting phylogenetic signals are associated with different mitochondrial regions (Cox, Kimball & Braun, 2007), even though the avian mitochondrial genome is non-recombining (Berlin and Ellegren, 2001; Berlin, Smith & Ellegren, 2004) so all regions are expected to have the same gene tree. Thus, localized biases exist in the mitochondrial genome, possibly reflecting the complex patterns of sequence evolution that make it difficult to extract

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

366 phylogenetic signal from the mitochondrial genome (cf. Braun and Kimball, 2002; Powell,
367 Barker & Lanyon, 2013). Biases can drive non-historical relationships both in analyses of
368 concatenated data and in coalescent-based estimates of species trees (e.g., Kimball et al., 2013),
369 making it desirable to identify localized biases, if they exist.

370 We have suggested that analyzing independent, multi-locus datasets (i.e., those with no
371 overlapping loci between datasets) can help identify nodes that may be present due to localized
372 biases (Kimball et al., 2013). If no conflict is identified, analysis of independent datasets can lead
373 to increased confidence in relationships and greater justification for combining datasets across
374 different studies (e.g., Smith, Braun & Kimball, 2013; Wang, Braun & Kimball, 2012). To
375 conduct such an independent evidence analysis for Galliformes we compared the results from
376 Kimball and Braun (2008) with an ML analysis of the 11 nuclear loci and 1 mitochondrial region
377 in this study that were not included in the Kimball and Braun (2008) dataset. Although the
378 mitochondrial regions form a single locus, since analyses of distinct mitochondrial regions can
379 result in different topologies (e.g., Cox, Kimball & Braun, 2007), presumably due to distinct
380 patterns of molecular evolution in each region, we included the mitochondrial 12S rRNA in the
381 independent evidence data matrix along with the 11 nuclear loci. This data matrix was limited to
382 the 44 taxa that were included in Kimball and Braun (2008), and a comparison of the ML
383 analysis of this independent data matrix revealed several differences from the Kimball and Braun
384 (2008) tree (Supplementary Information). However, all differences involved relationships
385 supported at less than 70% (often less than 50%) bootstrap support, suggesting there is little or
386 no conflicting signal between the two datasets. The topology and levels of support were also very
387 similar when the 44-taxon tree (which allowed a direct comparison with Kimball and Braun,

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

388 2008) and the 46-taxon tree were compared. Thus, this analysis did not reveal any evidence for
389 localized biases that might affect our conclusions and indicate that including all loci is justified.
390
391 *Sampling loci throughout the genome*
392 Expanded sampling of loci has the potential to reduce the impact of discordance among
393 gene trees by providing a better sample of the topologies that resulted from the multispecies
394 coalescent during evolutionary radiations. To determine whether sampling more loci (versus
395 sampling the same number of total sites but from fewer loci) is advantageous we compared the
396 performance of individual loci (and mitochondrial regions) against jackknifed datasets of
397 comparable size (Fig. 4B). The jackknifed datasets represent sites sampled from all loci, and thus
398 include a mixture of sites that come from loci that are evolving at different rates, exhibit different
399 patterns of evolution, and different evolutionary histories (i.e., distinct gene trees). Although
400 analyses of six loci yielded trees that were at least as similar to the total evidence tree as the
401 average jackknifed dataset (GAPDH, OVM, HSP90B1, CHRNG, CLTC, and HMGN2), the
402 other loci (and all three of the mitochondrial regions) yielded gene trees that were more
403 divergent from the total evidence tree than the jackknifed datasets of similar sizes. Not only did
404 the majority of loci perform worse with respect to recovering the total evidence tree than average
405 samples of sites from the data matrix, they typically performed substantially worse. Thus, on
406 average, for a given number of base pairs, sampling sites from many loci around the genome
407 appears to perform better than sampling from one or a few loci.
408 There are several reasons why sampling multiple loci is likely to be advantageous. It has
409 long been recognized that patterns of molecular evolution may make different loci ideal for the
410 resolution of nodes at different depths in the tree (Graybeal, 1994). Unlinked genes are also

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

411 expected to have somewhat distinct evolutionary histories (Oliver, 2013), so inclusion of many
412 different loci provides more information about these differences among gene trees. Finally,
413 broader sampling of loci should reduce the impact of localized biases, if they are present. The
414 relative contributions of these phenomena to the observed incongruence among estimates of gene
415 trees can be difficult to establish, though an examination of the performance of individual loci
416 can illustrate the differences among loci. For example, when several key clades that received
417 high bootstrap support (97-100%) were examined we found substantial variation in the
418 concordance factors (nodes labeled A-F, Fig. 2). Different nuclear loci and mitochondrial gene
419 regions show substantial variation in whether these key clades are found, and if found, the levels
420 of bootstrap support (Table 1). Thus, it appears that the specific clades supported by a locus are
421 quite variable – some loci supported one clade whereas other loci provided signal supporting
422 other clades. Consistent with our previous results (Fig. 4A), the two longest loci (FGB and
423 SERPINB14) supported the greatest number of clades with at least 50% bootstrap support, while
424 a relatively short locus (OVM) supports none of these clades with at least 50% bootstrap support.
425 However, we note that the RF distance from the total evidence tree (Fig. 1) was not substantially
426 lower for the longest loci than it was for the shorter loci (Fig. 4A). This suggests that at least
427 some of the difference between gene trees and the species tree may reflect discordance due to the
428 coalescent rather than the lack of power due to their short length. Combining these loci resulted
429 in a well-supported phylogeny for many clades. Not surprisingly, the number of loci or regions
430 that support a specific clade does appear related to the concordance factor, which is an estimate
431 of the proportion of the genome that supports a clade. Finally, we note that although FGB
432 exhibits localized biases that appear to provide non-historical signal for some avian relationships
433 (e.g., Kimball et al., 2013; Mayr, 2011) it appears to perform well in galliforms.

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

Conclusions

A consensus regarding many relationships within the Phasianidae now appears to be emerging, despite the fact that some challenges remain before we obtain a phylogenetic tree for this group that is both well resolved and strongly supported. While taxon sampling may help in some cases (e.g., Wang et al., 2013), we have shown here that adding loci is also extremely important. Regarding nodes that have been problematic in previous studies, we found strong evidence for a second major clade within the core phasianids (the non-erectile clade). We also found strong support for uniting the grouse and turkey (*Meleagris*) within the erectile clade and for placing the koklass (*Pucrasia*) sister to that grouse-turkey clade. As expected, however, the degree of improvement with increasing dataset size was modest, reflecting the difficulty of the relationships that remain unresolved. Indeed, some relationships remain problematic even with the larger dataset used in this study. Many of these problematic relationships were within genera, though a few poorly supported relationships among the higher-level clades still remain (Fig. 3). Whether the remaining unresolved nodes represent hard polytomies, that cannot be resolved, or soft polytomies that might be resolved with even larger datasets remains to be determined. Given the modest improvement evident in this study relative to Kimball and Braun (2008), it will likely require a much larger dataset, perhaps one with an order of magnitude increase in the number of variable sites, before it becomes clear whether the remaining unresolved nodes represent hard polytomies or soft polytomies that can eventually be resolved.

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

Acknowledgments

Ben Burkley, Andrew Cox, Amanda Hudson, Doug Storch, and Padi Tester assisted with data collection. We are grateful to the University of Florida Genetics Institute for providing computational resources. We thank Louisiana State University Natural History Museum for providing tissues for *Argusianus argus* (Accession B13314) and *Rollulus rouloul* (Accession B24971). This research was funded by the US National Science Foundation (Grant DEB-1118823 to RTK and ELB).

References

Armstrong, MH, Braun, EL, Kimball, RT. 2001. Phylogenetic utility of avian ovomucoid intron G: A comparison of nuclear and mitochondrial phylogenies in Galliformes. *Auk* 118:799-804.

Axelsson, E, Webster, MT, Smith, NGC, Burth, DW, Ellegren, H. 2005. Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. *Genome Research* 15:120-125.

Berlin, S, Ellegren, H. 2001. Evolutionary genetics: Clonal inheritance of avian mitochondrial DNA. *Nature* 413:37-38.

Berlin, S, Smith, NGC, Ellegren, H. 2004. Do avian mitochondria recombine? *Journal of Molecular Evolution* 58:163-167.

Bonilla, AJ, Braun, EL, Kimball, RT. 2010. Comparative molecular evolution and phylogenetic utility of 3'-UTRs and introns in Galliformes. *Molecular Phylogenetics and Evolution* 56:536-542.

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

480 Braun, EL, Kimball, RT. 2002. Examining basal avian divergences with mitochondrial
481 sequences: model complexity, taxon sampling, and sequence length. *Systematic Biology*
482 51:614-625.
483 Braun, EL, Kimball, RT, Han, KL, Iuhasz-Velez, NR, Bonilla, AJ, Chojnowski, JL, Smith, JV,
484 Bowie, RCK, Braun, MJ, Hackett, SJ, Harshman, J, Huddleston, CJ, Marks, B, Miglia,
485 KJ, Moore, WS, Reddy, S, Sheldon, FH, Witt, CC, Yuri, T. 2011. Homoplastic
486 microinversions and the avian tree of life. *BMC Evolutionary Biology* 11:141.
487 Chojnowski, JL, Kimball, RT, Braun, EL. 2008. Introns outperform exons in analyses of basal
488 avian phylogeny using clathrin heavy chain genes. *Gene* 410:89-96.
489 Cohen, C, Wakeling, JL, Mandiwana-Neudani, TG, Sande, E, Dranzoa, C, Crowe, TM, Bowie,
490 RCK. 2012. Phylogenetic affinities of evolutionarily enigmatic African galliforms: the
491 Stone Partridge *Ptilopachus petrosus* and Nahan's Francolin *Francolinus nahani*, and
492 support for their sister relationship with New World quails. *Ibis* 154:768-780.
493 Corl, A, Ellegren, H. 2013. Sampling strategies for species trees: The effects on phylogenetic
494 inference of the number of genes, number of individuals, and whether loci are
495 mitochondrial, sex-linked, or autosomal. *Molecular Phylogenetics and Evolution* 67:358-
496 366.
497 Cox, WA, Kimball, RT, Braun, EL. 2007. Phylogenetic position of the New World quail
498 (Odontophoridae): Eight nuclear loci and three mitochondrial regions contradict
499 morphology and the Sibley-Ahlquist tapestry. *Auk* 124:71-84.
500 Crowe, TM, Bowie, RCK, Bloomer, P, Mandiwana, TG, Hedderson, TAJ, Randi, E, Pereira, SL,
501 Wakeling, J. 2006. Phylogenetics, biogeography and classification of, and character
502 evolution in gamebirds (Aves: Galliformes): Effects of character exclusion, data

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

503 partitioning and missing data. *Cladistics* 22:1-38.
504 Dalloul, RA, Long, JA, Zimin, AV, Aslam, L, Beal, K, Blomberg, LA, Bouffard, P, Burt, DW,
505 Crasta, O, Crooijmans, RPMA, Cooper, K, Coulombe, RA, De, S, Delany, ME, Dodgson,
506 JB, Dong, JJ, Evans, C, Frederickson, KM, Flicek, P, Florea, L, Folkerts, O, Groenen,
507 MAM, Harkins, TT, Herrero, J, Hoffmann, S, Megens, H-J, Jiang, A, Jong, Pd, Kaiser, P,
508 Kim, H, Kim, K-W, Kim, S, Langenberger, D, Lee, M-K, Lee, T, Mane, S, Marcais, G,
509 Marz, M, McElroy, AP, Modise, T, Nefedov, M, Notredame, C, Paton, IR, Payne, WS,
510 Pertea, G, Prickett, D, Puiu, D, Qioa, D, Raineri, E, Ruffier, M, Salzberg, SL, Schatz,
511 MC, Scheuring, C, Schmidt, CJ, Schroeder, S, Searle, SMJ, Smith, EJ, Smith, J,
512 Sonstegard, TS, Stadler, PF, Tafer, H, Tu, ZJ, Tassell, CPV, Vilella, AJ, Williams, KP,
513 Yorke, JA, Zhang, L, Zhang, H-B, Zhang, X, Zhang, Y, Reed, KM. 2010. Multi-platform
514 next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome
515 assembly and analysis. *PLoS Biology* 8:e1000475.
516 Delsuc, F, Brinkmann, H, Philippe, H. 2005. Phylogenomics and the reconstruction of the tree of
517 life. *Nature Review Genetics* 6:361-375.
518 Graybeal, A. 1994. Evaluating the phylogenetic utility of genes: A search for genes informative
519 about deep divergences among vertebrates. *Systematic Biology* 43:174-193.
520 Hackett, SJ, Kimball, RT, Reddy, S, Bowie, RCK, Braun, EL, Braun, MJ, Chojnowski, JL, Cox,
521 WA, Han, KL, Harshman, J, Huddleston, CJ, Marks, BD, Miglia, KJ, Moore, WS,
522 Sheldon, FH, Steadman, DW, Witt, CC, Yuri, T. 2008. A phylogenomic study of birds
523 reveals their evolutionary history. *Science* 320: 1763-1768.
524 Han, KL, Braun, EL, Kimball, RT, Reddy, S, Bowie, RCK, Braun, MJ, Chojnowski, JL, Hackett,
525 SJ, Harshman, J, Huddleston, CJ, Marks, BD, Miglia, KJ, Moore, WS, Sheldon, FH,

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

526 Steadman, DW, Witt, CC, Yuri, T. 2011. Are transposable element insertions homoplasy
527 free?: An examination using the avian tree of life. *Systematic Biology* 60:375-386.
528 Hillier, LW, Miller, W, Birney, E, Warren, W, Hardison, RC, Ponting, CP, Bork, P, Burt, DW,
529 Groenen, MAM, Delany, ME, Dodgson, JB, Chinwalla, AT, Cliften, PF, Clifton, SW,
530 Delehaunty, KD, Fronick, C, Fulton, RS, Graves, TA, Kremitzki, C, Layman, D,
531 Magrini, V, McPherson, JD, Miner, TL, Minx, P, Nash, WE, Nhan, MN, Nelson, JO,
532 Oddy, LG, Pohl, CS, Randall-Maher, J, Smith, SM, Wallis, JW, Yang, SP, Romanov,
533 MN, Rondelli, CM, Paton, B, Smith, J, Morrice, D, Daniels, L, Tempest, HG, Robertson,
534 L, Masabanda, JS, Griffin, DK, Vignal, A, Fillon, V, Jacobbson, L, Kerje, S, Andersson,
535 L, Crooijmans, RPM, Aerts, J, van der Poel, JJ, Ellegren, H, Caldwell, RB, Hubbard, SJ,
536 Grafham, DV, Kierzek, AM, McLaren, SR, Overton, IM, Arakawa, H, Beattie, KJ,
537 Bezzubov, Y, Boardman, PE, Bonfield, JK, Croning, MDR, Davies, RM, Francis, MD,
538 Humphray, SJ, Scott, CE, Taylor, RG, Tickle, C, Brown, WRA, Rogers, J, Buerstedde,
539 JM, Wilson, SA, Stubbs, L, Ovcharenko, I, Gordon, L, Lucas, S, Miller, MM, Inoko, H,
540 Shiina, T, Kaufman, J, Salomonsen, J, Skjoedt, K, Wong, GKS, Wang, J, Liu, B, Wang,
541 J, Yu, J, Yang, HM, Nefedov, M, Koriabine, M, deJong, PJ, Goodstadt, L, Webber, C,
542 Dickens, NJ, Letunic, I, Suyama, M, Torrents, D, von Mering, C, Zdobnov, EM, Makova,
543 K, Nekrutenko, A, Elnitski, L, Eswara, P, King, DC, Yang, S, Tyekucheva, S,
544 Radakrishnan, A, Harris, RS, Chiaromonte, F, Taylor, J, He, JB, Rijnkels, M, Griffiths-
545 Jones, S, Ureta-Vidal, A, Hoffman, MM, Severin, J, Searle, SMJ, Law, AS, Speed, D,
546 Waddington, D, Cheng, Z, Tuzun, E, Eichler, E, Bao, ZR, Flicek, P, Shteynberg, DD,
547 Brent, MR, Bye, JM, Huckle, EJ, Chatterji, S, Dewey, C, Pachter, L, Kouranov, A,
548 Mourelatos, Z, Hatzigeorgiou, AG, Paterson, AH, Ivarie, R, Brandstrom, M, Axelsson, E,

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

549 Backstrom, N, Berlin, S, Webster, MT, Pourquie, O, Reymond, A, Ucla, C, Antonarakis,
550 SE, Long, MY, Emerson, JJ, Betran, E, Dupanloup, I, Kaessmann, H, Hinrichs, AS,
551 Bejerano, G, Furey, TS, Harte, RA, Raney, B, Siepel, A, Kent, WJ, Haussler, D, Eyras,
552 E, Castelo, R, Abril, JF, Castellano, S, Camara, F, Parra, G, Guigo, R, Bourque, G,
553 Tesler, G, Pevzner, PA, Smit, A, Fulton, LA, Mardis, ER, Wilson, RK. 2004. Sequence
554 and comparative analysis of the chicken genome provide unique perspectives on
555 vertebrate evolution. *Nature* 432:695-716.
556 Kaiser, VB, van Tuinen, M, Ellegren, H. 2007. Insertion events of *CR1* retrotransposable
557 elements elucidate the phylogenetic branching order in galliform birds. *Molecular*
558 *Biology and Evolution* 24:338-347.
559 Katsu, Y, Braun, EL, Guilette, LJ Jr., Iguchi, T. 2009. From reptilian phylogenomics to reptilian
560 genomes: analyses of c-Jun and DJ-1 proto-oncogenes. *Cytogenetics and Genome*
561 *Research* 127:79-93.
562 Kimball, RT, Braun, EL. 2008. A multigene phylogeny of Galliformes supports a single origin of
563 erectile ability in non-feathered facial traits. *Journal of Avian Biology* 39:438-445.
564 Kimball, RT, Braun, EL, Barker, FK, Bowie, RCK, Braun, MJ, Chojnowski, JL, Hackett, SJ,
565 Han, K-L, Harshman, J, Heimer-Torres, V, Holznagel, W, Huddleston, CJ, Marks, BD,
566 Miglia, KJ, Moore, WS, Reddy, S, Sheldon, FH, Smith, JV, Witt, CC, Yuri, T. 2009. A
567 well-tested set of primers to amplify regions spread across the avian genome. *Molecular*
568 *Phylogenetics and Evolution* 50:654-660.
569 Kimball, RT, Braun, EL, Ligon, JD, Randi, E, Lucchini, V. 2006. Using molecular phylogenetics
570 to interpret evolutionary changes in morphology and behavior in the Phasianidae. *Acta*
571 *Zoologica Sinica* 52(Supplement):362-365.

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

572 Kimball, RT, Braun, EL, Ligon, JD, Lucchini, V, Randi, E. 2001. A molecular phylogeny of the
573 peacock-pheasants (Galliformes: *Polyplectron* spp.) indicates loss and reduction of
574 ornamental traits and display behaviours. *Biological Journal of the Linnean Society*
575 73:187-198.
576 Kimball, RT, Braun, EL, Zwartjes, PW, Crowe, TM, Ligon, JD. 1999. A molecular phylogeny of
577 the pheasants and partridges suggests that these lineages are not monophyletic. *Molecular*
578 *Phylogenetics and Evolution* 11:38-54.
579 Kimball, RT, St Mary, CM, Braun, EL. 2011. A macroevolutionary perspective on multiple
580 sexual traits in the Phasianidae (Galliformes). *International Journal of Evolutionary*
581 *Biology* 2011:423938.
582 Kimball, RT, Wang, N, Heimer-McGinn, V, Ferguson, CN, Braun, EL. 2013. Identifying
583 localized biases in large datasets: A case study using the Avian Tree of Life. *Molecular*
584 *Phylogenetics and Evolution* 69:1021-1032.
585 Kriegs, JO, Matzke, AJM, Churakov, G, Kuritzin, A, Mayr, G, Brosius, J, Schmitz, J. 2007.
586 Waves of genomic hitchhikers shed light on the evolution of gamebirds (Aves:
587 Galliformes). *BMC Evolutionary Biology* 7:190.
588 Larget, BR, Kotha, SK, Dewey, CN, Ane, C. 2010. BUCKy: Gene tree/species tree
589 reconciliation with Bayesian concordance analysis. *Bioinformatics* 26:2910-2911.
590 Liu, L, Yu, L. 2011. Estimating species trees from unrooted gene trees. *Systematic Biology*
591 60:661-667.
592 Liu, L, Yu, L, Pearl, DK, Edwards, SV. 2009. Estimating species phylogenies using coalescence
593 times among sequences. *Systematic Biology* 58:468–477.
594 Mayr, G. 2011. Metaves, Mirandornithes, Strisores and other novelties – a critical review of the

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

higher-level phylogeny of neornithine birds. *Journal of Zoological Systematics and Evolutionary Research* 49:58–76.

McCormack, JE, Huang, HT, Knowles, LL. 2009. Maximum likelihood estimates of species trees: How accuracy of phylogenetic inference depends upon the divergence history and sampling design. *Systematic Biology* 58:501-508.

Meng, Y, Da, i.B, Ran, JH, Li, J, Yue, BS. 2008. Phylogenetic position of the genus *Tetraophasis* (Aves, Galliformes, Phasianidae) as inferred from mitochondrial and nuclear sequences. *Biochemical Systematics and Ecology* 36:626-637.

Miller, MA, Pfeiffer, W, Schwartz, T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE), New Orleans, LA, pp. 1 - 8.

Moore, WS. 1995. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. *Evolution* 49:718-726.

Nabhan, AR, Sarkar, IN. 2012. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in Bioinformatics* 13:122-134.

Nadeau, NJ, Burke, T, Mundy, NI. 2007. Evolution of an avian pigmentation gene correlates with a measure of sexual selection. *Proceedings of the Royal Society B-Biological Sciences* 274:1807-1813.

Oliver, JC. 2013. Microevolutionary processes generate phylogenomic discordance at ancient divergences. *Evolution* 67:1823-1830

Patel, S, Kimball, RT, Braun, EL, 2013. Error in phylogenetic estimation for bushes in the Tree of Life. *Journal of Phylogenetics and Evolutionary Biology* 1:110.

Posada, D, Crandall, KA. 1998. MODELTEST: testing the model of DNA substitution.

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

Bioinformatics 14:817-818.

Powell, AF, Barker, FK, Lanyon, SM. 2013. Empirical evaluation of partitioning schemes for phylogenetic analyses of mitogenomic data: an avian case study. *Molecular Phylogenetics and Evolution* 66:69-79.

Poynter, G, Huss, D, Lansford, R. 2009 Japanese quail: an efficient animal model for the production of transgenic avians. Cold Spring Harb Protoc doi: 10.1101/pdb.emo112.

Pratt, RC, Gibb, GC, Morgan-Richards, M, Phillips, MJ, Hendy, MD, Penny, D. 2009. Toward resolving deep Neoaves phylogeny: Data, signal enhancement, and priors. *Molecular Biology and Evolution* 26:313-326.

Randi, E, Lucchini, V, Armijo-Prewitt, T, Kimball, RT, Braun, EL, Ligon, JD. 2000. Mitochondrial DNA phylogeny and speciation in the tragopans. *Auk* 117:1003-1015.

Randi, E, Lucchini, V, Hennache, A, Kimball, RT, Braun, EL, Ligon, JD. 2001. Evolution of the mitochondrial DNA control region and cytochrome *b* genes and the inference of phylogenetic relationships in the avian genus *Lophura* (Galliformes). *Molecular Phylogenetics and Evolution* 19:187-201.

Robinson, DF, Foulds, LR. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53:131-147.

Rokas, A, Carroll, SB. 2006. Bushes in the tree of life. *PLoS Biology* 4:e352.

Ronquist, F, Huelsenbeck, JP. 2003. MrBayes 3: Bayesian phylogeneic inference under mixed models. *Bioinformatics* 19:1572-1574.

Sanchez-Gracia, A, Castresana, J. 2012. Impact of deep coalescence on the reliability of species tree inference from different types of DNA markers in mammals. *PLoS ONE* 7: e30239.

Shaw, TI, Ruan, Z, Glenn, TC, Liu, L. 2013. STRAW: Species TRee Analysis Web server.

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

Nucleic Acids Research, doi: 10.1093/nar/gkt1037.

Shen, YY, Liang, L, Sun, YB, Yue, BS, Yang, XJ, Murphy, RW, Zhang, YP. 2010. A mitogenomic perspective on the ancient, rapid radiation in the Galliformes with an emphasis on the Phasianidae. *BMC Evolutionary Biology* 10:132.

Smith, JV, Braun, EL, Kimball, RT. 2013. Ratite nonmonophyly: Independent evidence from 40 novel loci. *Systematic Biology* 62:35-49.

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

Swofford, DL. 2003. PAUP*: Phylogenetic analysis using parsimony (* and other methods), version 4.0. Sinauer, Sunderland, MA.

Thompson, JD, Gibson, TJ, Plewniak, F, Jeanmougin, F, Higgins, DG. 1997. The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* 25:4876-4882.

Wang, N, Braun, EL, Kimball, RT. 2012. Testing hypotheses about the sister group of the Passeriformes using an independent 30-locus data set. *Molecular Biology and Evolution* 29:737-750.

Wang, N, Kimball, RT, Braun, EL, Liang, B, Zhang, Z. 2013. Assessing phylogenetic relationships among Galliformes: a multigene phylogeny with expanded taxon sampling in Phasianidae. *PLoS ONE* 8:e64312.

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

661 Table 1. Performance of individual loci and mitochondrial gene regions in recovering six
662 different nodes (identified on Figs. 2 and 3). Values are bootstrap percents from RAxML. A
663 dashed line (-------) indicates the node was not in the gene tree; bold denotes >50% support.

	Length (bp)	A	B	C	D	E	F
Bootstrap (Fig. 1)	15866	100	100	100	100	98	97
Concordance Factor (Fig. 2)	15866	0.99	0.54	0.82	0.69	0.36	0.28
ALDOB	555	**92**	25	**62**	**75**	-------	-------
CALB1	623	**93**	-------	35	-------	-------	33
CHRNG	673	**97**	-------	36	-------	49	-------
CLTC	735	**97**	**93**	**96**	**78**	-------	-------
CLTCL1	427	**76**	-------	**92**	**93**	-------	-------
CRYAA	1061	**100**	**90**	**99**	**95**	34	-------
EEF2	975	**95**	-------	-------	**100**	-------	-------
FGB	1645	**100**	**97**	**100**	**100**	**72**	38
GAPDH	417	**96**	-------	**77**	**83**	38	-------
HMGN2	758	**100**	36	**86**	-------	49	45
HSP90B1	656	**100**	**91**	**76**	**94**	-------	-------
OVM	519	28	-------	-------	47	-------	-------
PCBD1	575	**89**	**73**	**74**	**57**	-------	17
RHO	1105	**100**	**50**	**98**	-------	-------	**56**
SERPIN	2007	**100**	48	**100**	**63**	**76**	**66**
ND2	1041	**97**	**50**	**84**	39	21	17
CYB	1143	**60**	**64**	**73**	42	-------	-------
12S	951	**90**	-------	36	**51**	**60**	-------

664

665

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

Figure Legends

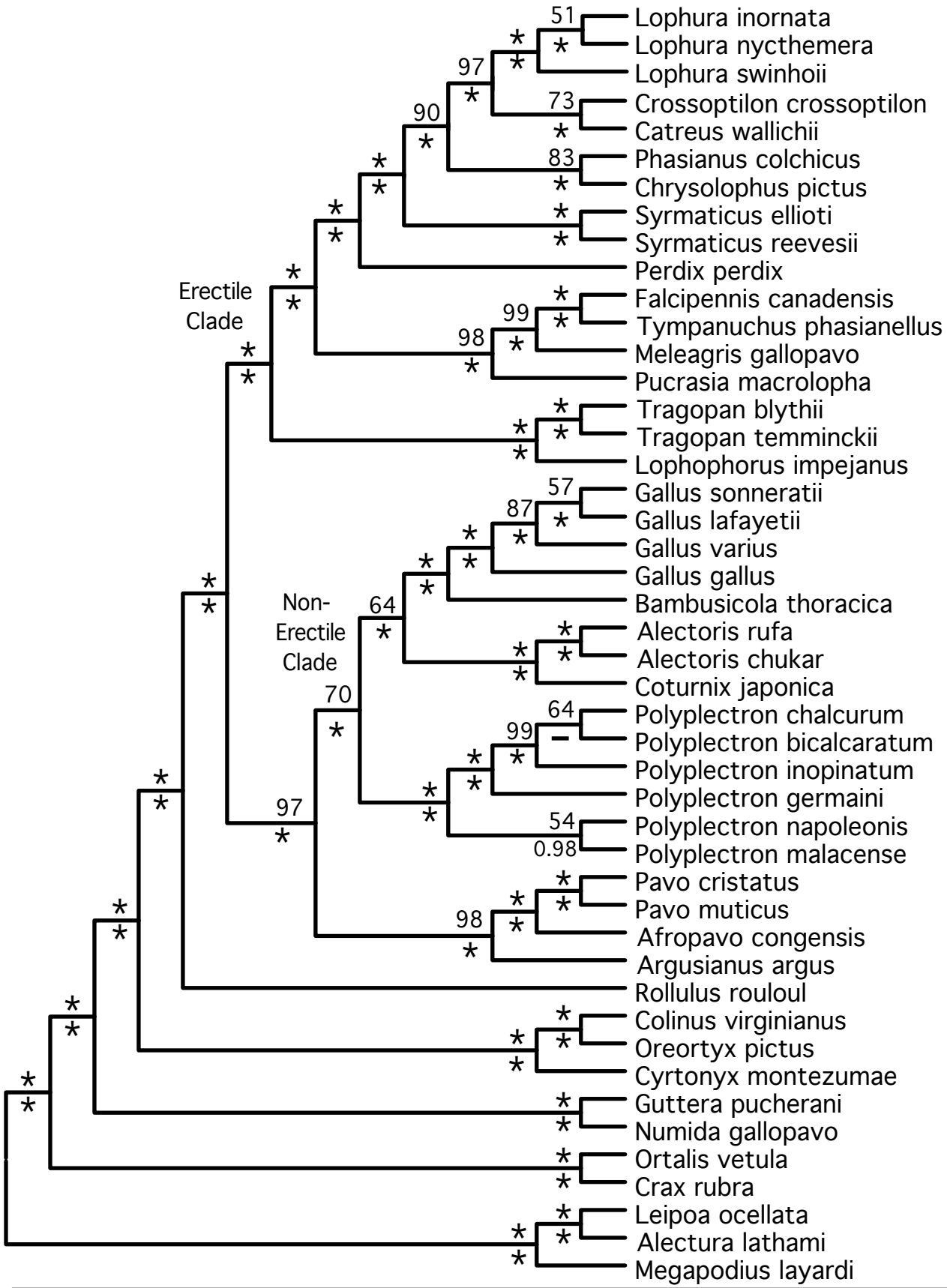
Figure 1. Phylogenetic tree estimated from the total evidence (nuclear + mitochondrial) dataset. Values above nodes are the % bootstrap support from a partitioned ML analysis. Values below nodes are the posterior probabilities from a Bayesian analysis. * represents either 100% bootstrap support or a posterior probability of 1.0.

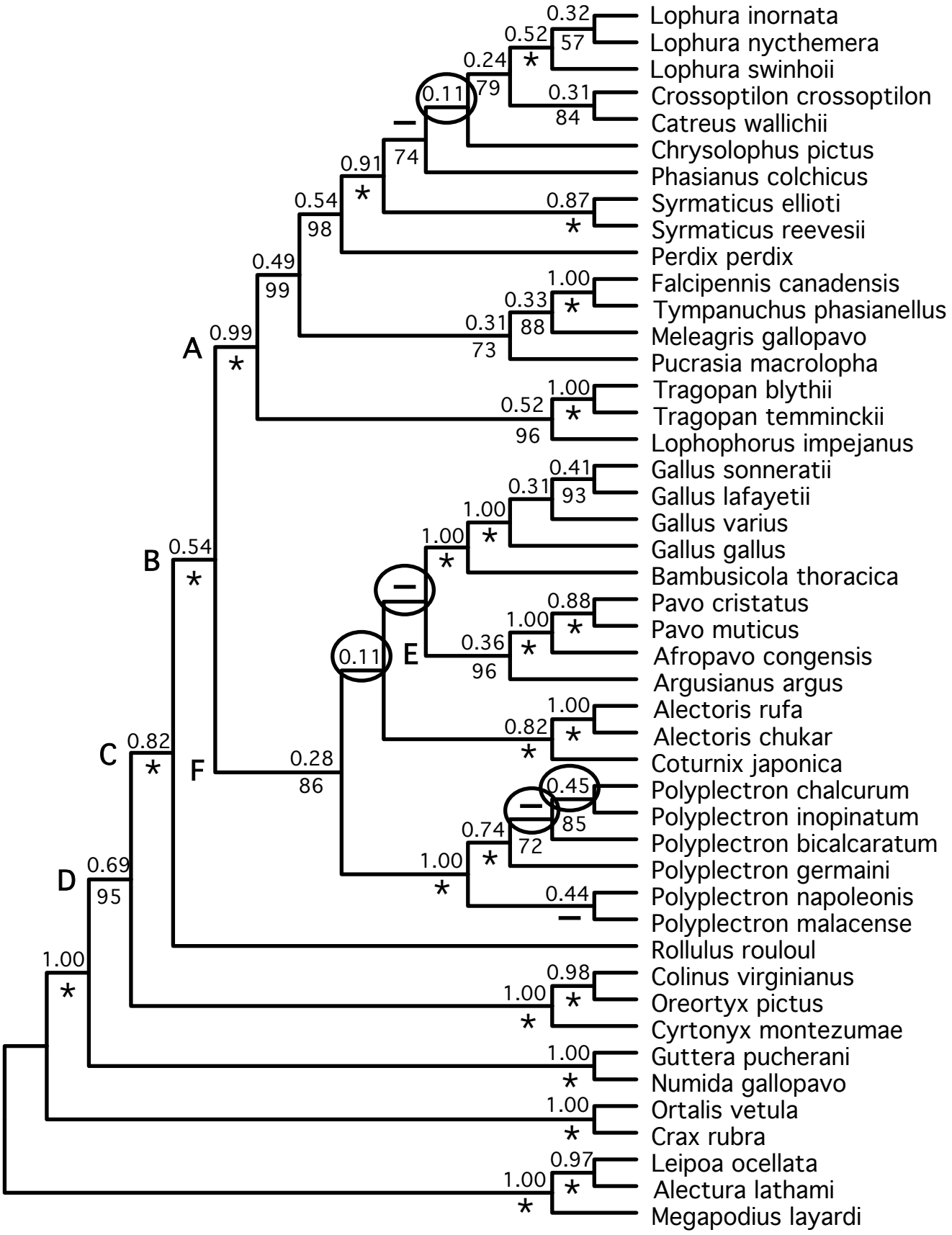
Figure 2. Species tree estimated from individual gene trees. Shown is the BUCKy population tree. Values above nodes are the concordance factors and values below the nodes are the % bootstrap support from the NJ_{st} analysis; only values ≥ 50% are shown. Letters (A-F) correspond to the key nodes nodes included in Table 1. Nodes that are circled represent differences between this topology and that using concatenation (Figure 1). Dashed lines above the node indicate that the primary concordance tree exhibited a different topology from the BUCKy population tree and dashed lines below the node indicate that the NJ_{st} tree exhibited a different topology.

Figure 3. Summary tree that represents our best estimate of galliform phylogeny, with poorly supported nodes or those that are in conflict between Figures 1 and 2 collapsed and shaded.

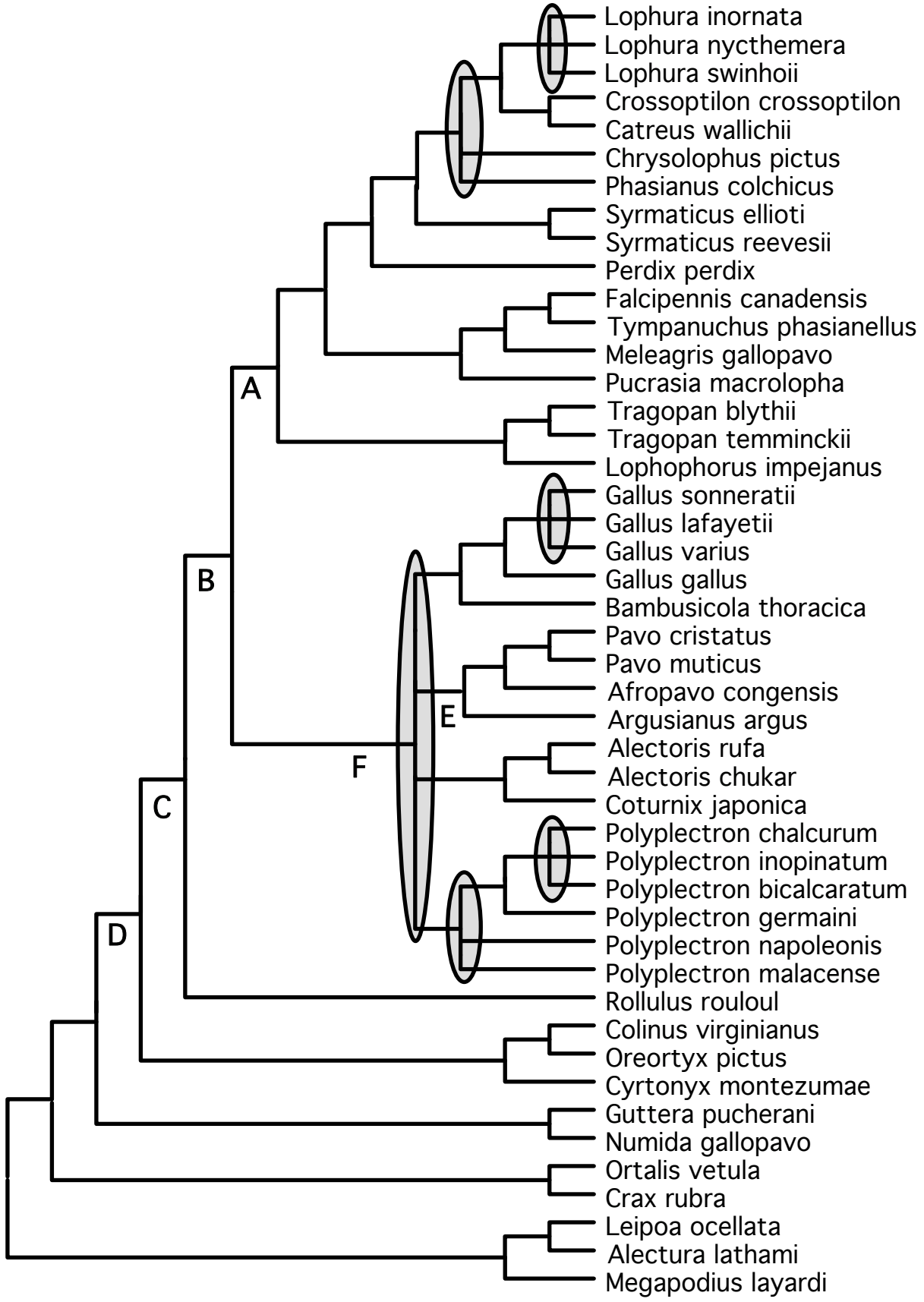
Figure 4. RF distances between the unpartitioned ML topology and estimates of phylogeny based upon different data matrices. Diamonds represent the average RF distance of the 100 jackknifed datasets from the ML tree and the error bars represent the standard deviation of the RF distances. A. RF distances between jackknifed datasets and the ML tree estimated from the nuclear dataset for datasets of varying sizes. B. RF distances between the ML trees estimated from each locus or mitochondrial region and the nuclear tree are shown.

PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1

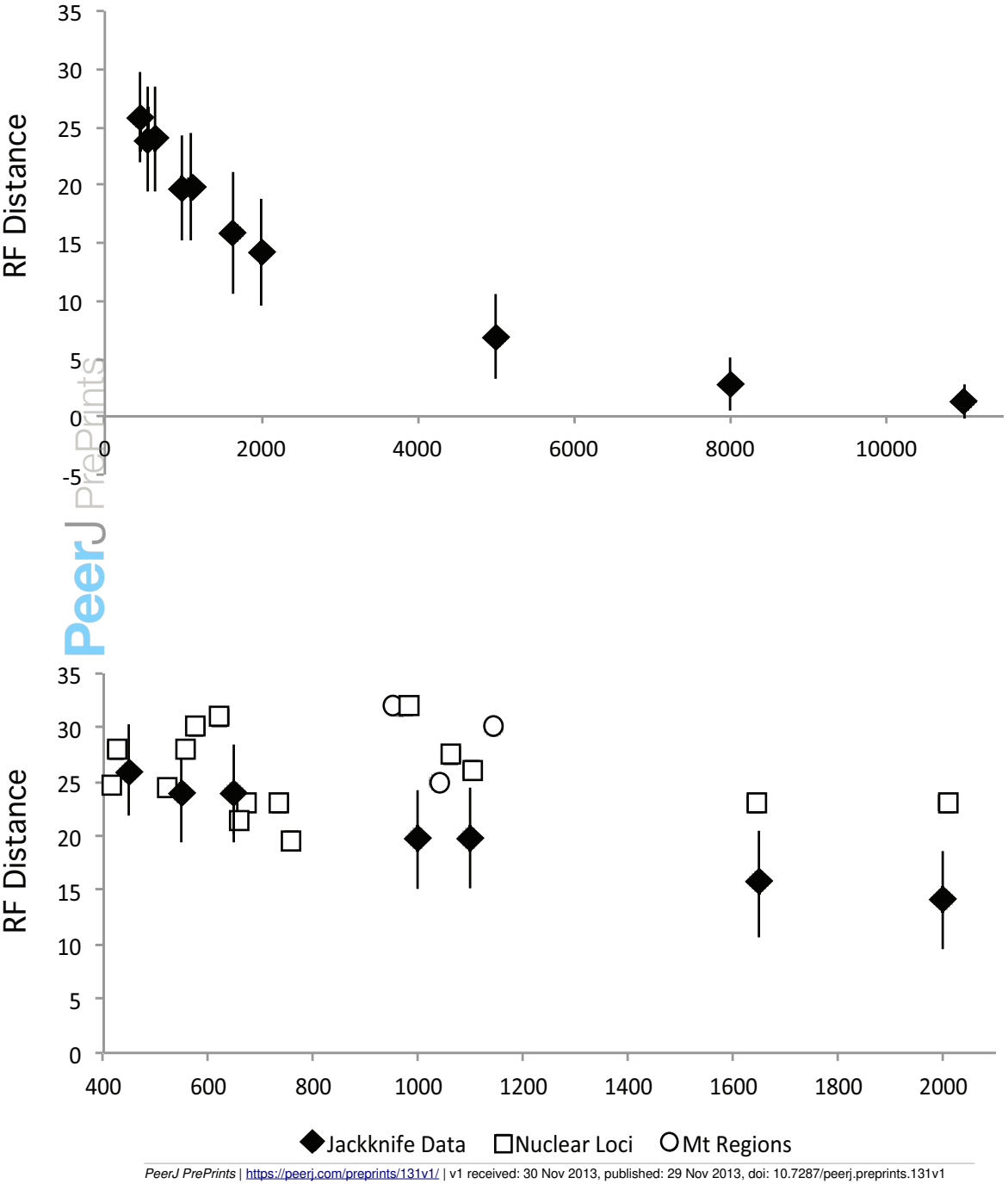




PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1



PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1



PeerJ PrePrints | https://peerj.com/preprints/131v1/ | v1 received: 30 Nov 2013, published: 29 Nov 2013, doi: 10.7287/peerj.preprints.131v1