Available online at www.sciencedirect.com
ScienceDirect
Gene 410 (2008) 89–96
GENE
www.elsevier.com/locate/gene

Introns outperform exons in analyses of basal avian phylogeny using clathrin heavy chain genes

Jena L. Chojnowski, Rebecca T. Kimball, Edward L. Braun *

Department of Zoology, 223 Bartram Hall, PO Box 118525, University of Florida, Gainesville, FL 32611, USA

Received 20 July 2007; received in revised form 28 November 2007; accepted 30 November 2007
Available online 11 January 2008
Received by T. Gojobori

Abstract

Neoaves is the most diverse major avian clade, containing ~95% of avian species, and it underwent an ancient but rapid diversification that has made resolution of relationships at the base of the clade difficult. In fact, Neoaves has been suggested to be a "hard" polytomy that cannot be resolved with any amount of data. However, this conclusion was based on slowly evolving coding sequences and ribosomal RNAs and some recent studies using more rapidly evolving intron sequences have suggested some resolution at the base of Neoaves. To further examine the utility of introns and exons for phylogenetics, we sequenced parts of two unlinked clathrin heavy chain genes (*CLTC* and *CLTCL1*). Comparisons of phylogenetic trees based upon individual partitions (i.e. introns and exons), the combined dataset, and published phylogenies using Robinson–Foulds distances (a metric of topological differences) revealed more similarity than expected by chance, suggesting there is structure at the base of Neoaves. We found that introns provided more informative sites, were subject to less homoplasy, and provided better support for well-accepted clades, suggesting that intron evolution is better suited to determining closely-spaced branching events like the base of Neoaves. Furthermore, phylogenetic power analyses indicated that existing molecular datasets for birds are unlikely to provide sufficient phylogenetic information to resolve relationships at the base of Neoaves, especially when comprised of exon or other slowly evolving regions. Although relationships among the orders in Neoaves cannot be definitively established using available data, the base of Neoaves does not appear to represent a hard polytomy. Our analyses suggest that large intron datasets have the best potential to resolve relationships among avian orders and indicate that the utility of intron data for other phylogenetic questions should be examined.

Keywords: Simulation; Power analysis; Congruence; Polytomy; Saturation

1. Introduction

The relationships among extant birds has been a subject of substantial debate since the earliest days of evolutionary biology, and the availability of molecular data has done little to resolve this debate (e.g., Cracraft et al., 2004; Poe and Chubb, 2004; Harshman, 2007). Although there is consensus that extant birds can be divided into three major clades (Paleognathae, Galloanserae, and Neoaves), relationships among orders within Neoaves (~95% of all avian species) remain unresolved. It has been suggested that the base of Neoaves represents a "hard" polytomy that will not be resolved with any amount of data (Poe and Chubb, 2004).

Attempts to use molecular phylogenetics to resolve relationships among orders in Neoaves have been complicated by their apparent rapid and ancient diversification (Poe and Chubb, 2004). Rapid radiations result in short internodes, with few changes that unite groups (Braun and Kimball, 2001). The majority of molecular studies have focused on exons (e.g., *RAG1* and *EGR1* [also called *Zenk*]) and mitochondrial sequences (coding and ribosomal RNAs). Studies using these sequences have had limited resolution at the base of Neoaves (e.g. Groth and Barrowclough, 1999; van Tuinen et al., 2000; Chubb, 2004; Watanabe et al., 2006; Gibb et al., 2007).

Abbreviations: CI, consistency index; *CLTC*, clathrin heavy chain; *CLTCL1*, clathrin heavy chain-like; *EGR1*, early growth response factor 1; *FGB*, β-fibrinogen; ML, maximum likelihood; MP, maximum parsimony.

* Corresponding author. Tel.: +1 352 846 1124; fax: +1 352 392 3704.
E-mail address: ebraun68@ufl.edu (E.L. Braun).

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.11.016

However, analyses of a single nuclear intron (β-fibrinogen [*FGB*] intron 7) appeared to support some deep branches in Neoaves (Prychitko and Moore, 2003; Fain and Houde, 2004). Fain and Houde (2004) had broader taxon sampling and concluded that *FGB* intron 7 supported splitting Neoaves into two clades they called Metaves and Coronaves. Ericson et al. (2006) corroborated this division using a combination of intron and exon regions (including *FGB* intron 7). This suggests that, in contrast to placental mammals where coding regions have successfully resolved relationships (Murphy et al., 2001), more rapidly evolving intronic regions may have the greatest potential to resolve relationships at the base of Neoaves.

To further examine the utility of introns, we obtained sequences from two paralogous clathrin heavy chain genes that arose in an ancient genome (or large-scale) duplication event. While both maintained the basic structural features of clathrin heavy chains, their interactions with regulatory proteins have diversified (Wakeham et al., 2005). Both are part of the polyhedral lattice surrounding coated pits and vesicles involved in intracellular trafficking of receptors and endocytosis of macromolecules. *CLTC* (clathrin heavy chain) is expressed ubiquitously in all vertebrates that have an ortholog, while *CLTCL1* (clathrin, heavy chain-like 1) is specialized in humans to have a distinct role in muscle tissues (Wakeham et al., 2005). The chicken (*Gallus gallus*) orthologs of *CLTC* and *CLTCL1* are on chromosomes 19 and 15, respectively. Although both genes are likely under selection to maintain their functional differences, our data primarily consists of introns (*CLTC* introns 6 and 7 and *CLTCL1* intron 7) and this non-coding data is expected to largely show neutral evolution.

The conflicting phylogenetic hypotheses of Poe and Chubb (2004), who proposed that Neoaves is a hard polytomy, and Fain and Houde (2004), who divided of Neoaves into Metaves and Coronaves, make fundamentally different predictions. If the base of Neoaves is a hard polytomy, then estimates of phylogeny based upon novel data will show no more similarity to phylogenetic trees in previous studies than expected by chance and power analyses will indicate that sufficient data are available to recover an accurate estimate of avian phylogeny. In contrast, if the base of Neoaves can be resolved, similar structure will be found in analyses of additional gene regions. We examine these questions by comparing tree distances between estimates of phylogeny obtained using our clathrin heavy chain data and previous publications. Finally, we estimate the rates of *CLTC* and *CLTCL1* sequence evolution, focusing on the implications of these rates to resolve avian relationships at the base of Neoaves.

2. Methods and materials

2.1. DNA amplification, sequencing, and alignment

Sequences (Genbank accession nos. EU302706–EU302791) from 43 taxa representing 21 orders (see Table S1 for tissue information) were obtained directly from PCR products using the ABI BigDye® Terminator v.3.1 chemistry and an ABI Prism™ 3100-Avant genetic analyzer (PE Applied Biosystems). Standard PCR conditions were used and the primer sequences are listed in Table S2. If length heterozygosities obscured parts of sequences, they were cloned into pGEM®-T Easy vector (Promega) and plasmids were isolated using the Eppendorf Perfectprep® Plasmid Mini kit before sequencing. Contigs were assembled using Sequencher™ 4.1 (Gene Codes Corp.) and intron-exon junctions were annotated based upon homology, checking for presence of GT-AG dinucleotides at the intron boundaries. Sequences were initially aligned using ClustalX (Thompson et al., 1997) and the alignment was refined by eye using MacClade 4.0 (Maddison and Maddison, 2000). A large insertion (226 bp) present only in the kagu and sunbittern *CLTCL1* intron sequences was excluded from phylogenetic analyses.

2.2. Phylogenetic analyses

Maximum likelihood (ML) analyses were performed on the combined (*CLTC* and *CLTCL1*) dataset and each individual partition; the combined dataset was also used for MP and Bayesian analyses. ML and MP analyses were conducted using PAUP* 4.0b10 (Swofford, 2003), ML bootstrap analyses and ML analyses of simulated datasets were conducted using RAxML-VI (Stamatakis, 2006), and Bayesian analyses were conducted using MrBayes 3.1.2 (Ronquist and Huelsenbeck, 2003). For the Bayesian analyses, we conducted two runs of four chains each that were run for 5 million generations (using default heating parameters), sampling every 100 generations and discarding the first 40,000 trees sampled as "burn-in". We used MODELTEST 3.06 (Posada and Crandall, 1998) and the AIC criterion to select the appropriate model for model-based (ML and Bayesian) analyses; RAxML analyses were conducted using the GTR+CAT model. ML bootstrap support was estimated using 100 replicates and MP bootstrap support was estimated using 1000 replicates with 10 random additions per replicate.

Insertions and deletions (indels) were coded using the simple indel coding method of Simmons and Ochoterena (2000) as implemented in the gap recoder program by Rick Ree (http://maen.huh.harvard.edu:8080/services/gap_recoder); indels from all three introns were combined to generate the intron partition. We used PAUP* to examine the consistency index (CI) of the indels on the ML tree estimated from the combined dataset. We then focused on those indel characters that had a CI excluding uninformative sites of 1 or 0.5 (those that exhibited little or no homoplasy relative to the ML tree) and counted the number of these indels supporting the well-established monophyletic groups in our taxon sample (lettered groups in Fig. 1).

2.3. Molecular clock analyses

We used non-parametric rate smoothing (Sanderson, 1997) as implemented in TreeEdit 1.0 (Rambaut and Charleston, 2002) and the Bayesian approach of Thorne and Kishino (2002) as implemented in Multidivtime.09.25.03. Analysis used branch lengths and parameter estimates from PAML 3.15 (Yang, 1997), with branch lengths for TreeEdit reflecting a four

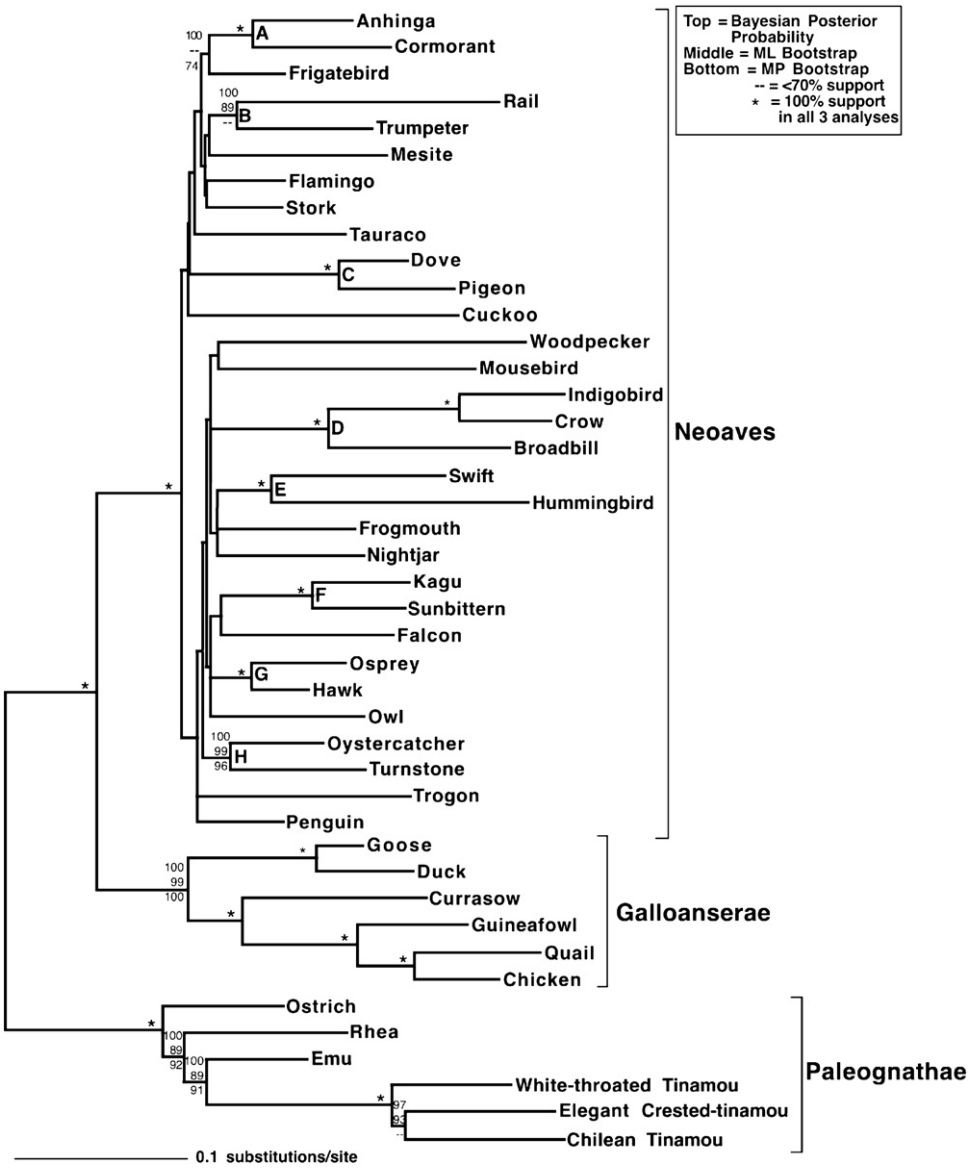

Fig. 1. Maximum likelihood estimate of avian phylogeny based upon the combined dataset. ML (RAxML) and MP (PAUP*) bootstrap support is presented along with Bayesian posterior probabilities (from MrBayes). Stars indicate branches with 100% support in all analyses. Absence of a star or support values indicates that support was <70% in all analyses. The letters identify well-accepted clades that are included in Table 2.

partition multigene analysis with linked branch length parameters (using baseml from PAML) and mutidivtime used parameter estimates obtained using baseml. A diverse set of avian fossils was used to calibrate the clock (Table S3, see Supplementary Information for additional details).

2.4. Tree comparisons

To compare the differences among trees generated using different data partitions and to compare our ML tree from the combined dataset with previous studies, we used Robinson and Foulds (1981) distances. Phylogenies used for comparison include the *FGB* intron 7 tree of Fain and Houde (2004; Fig. 2), the combined nuclear intron and exon tree of Ericson et al. (2006), the morphological trees of Mayr and Clarke (2003) and Livezey and Zusi (2007), the synthesis of recent studies from Cracraft et al. (2004), and the DNA–DNA hybridization "Tapestry" of Sibley and Ahlquist (1990). When taxa differed among trees, we used the best substitutions based upon current classification or kept groupings unresolved for monophyletic lineages where taxon sampling differed. In some cases, published phylogenies did not include all lineages included in our data set (e.g., Mayr and Clarke, 2003 did not include Piciformes). In these cases, we placed the absent taxa in an unresolved position at the base of the Neoaves.

To establish a null distribution for the Robinson–Foulds distances, we compared trees to a set of random trees. Completely random trees would be expected to contain nodes that would contradict well-established monophyletic groups. Therefore, we constrained random trees to maintain certain well-established relationships. The constraints included the nodes that unite Neoaves, Neognathae, Galloanserae (and all

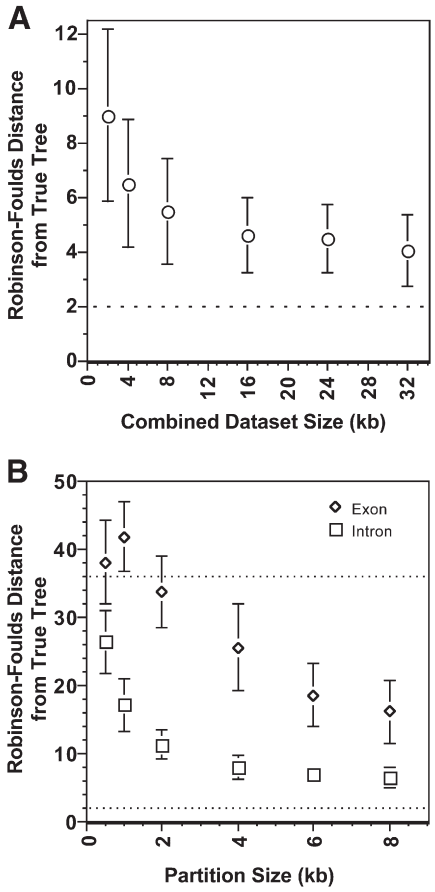

Fig. 2. Results of simulations used to establish the power of phylogenetic analyses. The distance from the true tree (the tree used to simulate the data) is indicated along with error bars indicating the standard deviation for 100 simulations. The minimum distance from the true tree for fully resolved trees is indicated by the dotted line (the lower dotted line in part B), which is greater than zero due because the true tree includes two polytomies. Simulated intron and exon data was generated using parameters estimated from the clathrin heavy chain data. A. Results of ML analyses using heterogeneous datasets with 75% simulated intron data and 25% simulated exon data (similar to the clathrin heavy chain dataset). B. Results of ML analyses of simulated datasets corresponding to the individual partitions. The values given for the intron partitions are the maximum distance for the three simulated introns. The upper dotted line is the distance expected for random trees given the constraints described in the text.

nodes within Galloanserae), Tinamidae (tinamous) and all lettered nodes in Fig. 1, since these represent well-accepted monophyletic groups in. PAUP* was used to generate the random trees (with constraints) and to calculate tree distances.

2.5. Power analyses

Absolute rates of sequence evolution were estimated by dividing the ML branch length (substitutions/site) by the time since a lineage diverged from a common ancestor (see supplementary material). These rates were only estimated for terminal branches, and the median rate was used to estimate the amount of data necessary to be 95% confident that at least one synapomorphy uniting a group will be observed (Braun and Kimball, 2001). The branch length used in the polytomy power analysis was 0.5 million years, reflecting a short internode near the base of Neoaves (shown in supplemental material on clock analyses).

Simulations used the evolver program from PAML package (Yang, 1997), using the GTR+Γ+inv model for intron partitions and the codon-based model of Yang et al. (1998) for exons (with parameter estimates obtained using PAML). The simulations of individual partitions were concatenated, and then RAxML was used to analyze the individual and combined datasets. All simulations used four equal-length partitions, three of which used parameter estimates from each of the three clathrin heavy chain introns and the fourth used the exon parameter estimates.

3. Results and discussion

3.1. Molecular evolution of CLTC and CLTCL1

The best fitting model for the combined dataset was GTR+Γ+inv, but the best fitting models for individual intron and exon partitions show a systematic difference reflecting differences in the base frequency parameters. When the partitions were examined individually, the best fitting exon models had equal base frequencies while the best fitting intron models had unequal base frequencies (Table 1). This is consistent with the observation that exons have a higher GC content (~50%) than introns (~41%). This was expected based upon the clear functional constraints in exons due to codon structure, and that exons exhibit greater variance in site to site rate heterogeneity than introns. In fact, the best fitting model for *CLTCL1* intron 7 does not include invariant sites, and those intron partitions that include invariant sites in the best fitting model have a small number of invariant sites (<3%). Likewise, the intron partitions have higher Γ-distribution shape parameters (α) than the exon partitions (which also have >50% invariant sites). These patterns are not surprising because most sites in introns are likely free to evolve (although some introns do contain regulatory elements that are presumably subject to purifying selection; see Le Hir et al., 2003) while nonsynonymous sites in coding exons are subject to constraint (the ratio of the nonsynonymous to synonymous rate [ω] for clathrin heavy chain exons is very low [ω=0.0293] suggesting strong

Table 1
The best fitting evolutionary model the clathrin heavy chain data partitions

Partition	Best fitting model[a]	Proportion of invariant sites	α	GC%	CI[b]
Combined	GTR+Γ+inv	0.163	4.02	42	0.461
CLTC	GTR+Γ+inv	0.144	3.84	42	0.454
CLTCL1	TVM+Γ+inv	0.221	5.37	44	0.487
CLTC int6	GTR+Γ+inv	0.0309	5.86	40	0.463
CLTC int7	TVM+Γ+inv	0.0321	3.15	42	0.455
CLTCL1 int7	GTR+Γ	0	8.39	41	0.493
CLTC exons	K80+Γ+inv	0.530	0.626	48	0.361
CLTCL1 exons	TVMef+Γ+inv	0.555	1.31	49	0.436

[a] The best fitting model identified by MODELTEST (Posada and Crandall, 1998).
[b] Consistency index (CI) excluding uninformative sites given the ML tree.

constraints). Overall, the intron and exon partitions exhibited more differences than two different paralogs.

Most partitions exhibited similar levels of homoplasy based upon the CI of parsimony informative sites calculated using the ML tree for the combined dataset (Table 1). The exon partitions, however, had a lower CI than the intron partitions (when *CLTC* and *CLTCL1* were combined, the CI excluding uninformative sites was 0.388 for the exon and 0.467 for the intron), indicating that the exons are more subject to homoplasy than the associated introns. Neither the introns nor the exons showed any evidence of saturation in a standard saturation plot (Fig. S1), despite the greater divergence of the introns. There were also a large number of intronic indels, which exhibited less homoplasy than either intronic or exonic nucleotide changes (CI excluding uninformative indels = 0.550).

3.2. Estimates of avian phylogeny using clathrin heavy chain genes

Phylogenetic analyses of both *CLTC* and *CLTCL1* largely result in identification of the same well-supported nodes (Fig. 1), although estimates of phylogeny obtained using *CLTC*, a larger region when compared to *CLTCL1* that comprises two introns, exhibits higher bootstrap support (data not shown). As expected, Paleognathae and Neognathae were well-supported clades (assuming the root of the avian tree falls between Paleognathes and Neognathes; see Braun and Kimball, 2002), as were Galloanserae and Neoaves. Other lineages that were expected to be monophyletic were found using the combined dataset as well as the larger and/or more variable partitions (*CLTC*, *CLTCL1*, and each intron) with relatively high support (Table 2). In contrast, analyses of the combined exon partition rarely found expected monophyletic lineages, or found them with low bootstrap support (Table 2). Overall, the introns clearly outperformed the exons not only in finding expected clades but also in having higher support for clades.

Since the indels exhibited little homoplasy, we expected to find indel support for many relationships. Most well-supported nodes had strong indel support (Table 2). Of interest is the single large (227 bp) insertion in *CLTCL1* that was found exclusively in the kagu and sunbittern. A search of the chicken genome using BLASTN indicated that this large insertion did not show homology to a previously identified transposable element, such as CR1.

Table 2
Maximum likelihood bootstrap support by data partition for well-accepted clades[a]

Partitions	Length[b]	A	B	C	D	E	F	G	H
Combined	3549	100	89	100	100	100	100	100	99
CLTC	2665	100	70	100	100	100	100	100	98
CLTCL1	884	69	68	100	82	43	98	69	26
CLTC int6	1498	94	32	100	100	78	100	100	76
CLTC int7	851	100	67	100	100	93	100	100	53
CLTCL1 int7	665	73	52	100	76	37	98	80	46
All exons	535	–	–	90	61	–	63	–	–
Indels (CI = 1/CI = 0.5)[c]		2/0	0/3	5/3	13/5	1/4	4/2	2/0	2/1

[a] The well-accepted clades correspond to those indicated with letters on Fig. 1. The clades are A = Anhinga/Cormorant, B = Trumpeter/Rail, C = Dove/Pigeon, D = Crow/Indigobird/Broadbill, E = Hummingbird/Swift, F = Sunbittern/Kagu, G = Hawk/Osprey, H = Turnstone/Oystercatcher.

[b] Aligned base pairs excluding the Kagu/Sunbittern insert.

[c] The number of indels supporting the indicated clades are listed, with indels that have CI = 1 to the left and those that have CI = 0.5 to the right.

3.3. Estimates of avian divergence times

The internodes at the base of Neoaves are very short (Fig. 1; see also van Tuinen et al., 2000; Poe and Chubb, 2004). Our molecular clock showed substantial rate heterogeneity among lineages and did not fit a molecular clock based upon the likelihood ratio test ($P < 0.0001$), so we used rate smoothing. After calibration (see Fig. S2), we found that the divergences occurred near the Cretaceous–Tertiary boundary (Fig. S3), in agreement with the fossil record (e.g. Bleiweiss, 1998; Dyke and van Tuinen, 2004) and some molecular studies (e.g., Ericson et al., 2006).

Reconstructing phylogenetic relationships when many cladogenic events occurred during a short time requires the use of markers that are evolving at sufficiently high rates; therefore, mutations are likely to accumulate along the short internodes that define clades but low enough homoplasy that some of those mutations will persist through time. The rate of the clathrin heavy chain introns is fairly similar and all accumulate substitutions ~3.5-fold more rapidly than the exons (data not shown), but they actually exhibit less homoplasy than the exon regions (see above). This greater homoplasy of exon regions likely reflects that most substitutions in coding regions are synonymous; because some synonymous sites can only be occupied by two different nucleotides they are expected to saturate more rapidly.

3.4. Tree comparisons

Although we did not find a high degree of support for relationships among orders in Neoaves, the phylogenies estimated using individual partitions appeared similar. To examine the congruence of phylogenies based upon each of the individual partitions quantitatively, we measured the distances between trees using the method of Robinson and Foulds (1981) (this measures the number of branches that appear in each tree but not in both trees, so larger numbers reflect greater topological differences). This approach also allowed us to compare the combined clathrin heavy chain phylogeny to phylogenies from previous studies that used similar taxa.

The estimates of phylogeny based upon the individual introns are closer to each other and to the combined phylogeny than either is to the exon phylogeny (Table 3), despite the similar length of the exons and *CLTCL1* intron 7. The larger Robinson–Foulds distances in comparisons that include exons (*CLTC* and *CLTCL1*) relative to those that just include introns suggests there may be conflict between exons and introns, even within the same paralog. In sharp contrast, the intron phylogenies show more similarity to each other and to the combined analyses than to random trees. The combined *CLTC*

Table 3
Robinson–Foulds distances for ML trees for each of our data partitions

Partitions[a]	Combined	*CLTC*	*CLTCL1*	*CLTC* int6	*CLTC* int7	*CLTCL1* int7	All exons[b]	Random trees[c]
Combined	–							36–44
CLTC	14	–						36–46
CLTCL1	33	35	–					31–39
CLTC int6	9	21	34	–				37–45
CLTC int7	0	14	33	9	–			36–44
CLTCL1 int7	11	21	34	2	11	–		37–45
All exons[b]	44.5	46.5	37.5	45.5	44.5	45.5	–	42–51

[a] The combined partition includes all *CLTC* and *CLTCL1* data except the Kagu/Sunbittern insert.
[b] Two maximum likelihood trees were equally likely therefore the data presented is an average of the two trees.
[c] Range of distances to a set of random trees constrained to include all well-accepted clades.

topology is closer to the combined and intron phylogenies than the *CLTCL1* topology, though this is expected given that *CLTC* is longer and thus makes a greater contribution to the combined dataset. The difference between the performance of introns and exons probably reflects the greater homoplasy in exons combined with a lower rate of exon evolution (~28% of the mean rate of intron evolution).

Among the six published studies we compared to our combined phylogeny (Table 4), the most divergent topology is that of the Tapestry (Sibley and Ahlquist, 1990). The difference between the Tapestry and our combined tree was larger than observed in any random tree, emphasizing strong incongruence. The Tapestry has received substantial criticism (summarized by Harshman, 1994) and the differences we observed probably reflect these issues. Excluding the Tapestry, the morphological analyses (Mayr and Clarke, 2003; Livezey and Zusi, 2007) were most divergent from our combined phylogeny and others based upon sequence data (Fain and Houde, 2004; Ericson et al., 2006). The differences between analyses based upon morphology and sequence data suggest they might reflect distinct phylogenetic signals.

The published topology with the smallest Robinson–Foulds distance from our combined tree was the consensus avian tree (TOL) from Cracraft et al. (2004). The TOL topology also showed similarity to the Fain and Houde (2004) and Ericson et al. (2006) topologies, probably reflecting the fact that the TOL is a synthesis of recently published studies that only included reliable groups (based upon expert opinion) and otherwise left groups unresolved. The limited resolution of TOL is expected to reduce the Robinson–Foulds distance to other trees, so it is important to consider distances in light of the distance to random trees (which is lower for TOL than for other trees; Table 4).

Table 4
Robinson–Foulds Test for the maximum likelihood tree obtained from our combined dataset compared to several published topologies

Partitions[a]	Combined	Cracraft	FH	Ericson	Tapestry	MC	LZ	Random trees[b]
Combined	–							36–44
Cracraft	26	–						28–32
FH	34	22	–					32–38
Ericson	32	20	22	–				34–38
Tapestry	47	29	41	39	–			41–49
MC	41	29	37	31	42	–		41–47
LZ	43	27	39	35	40	36	–	45–51

[a] The combined partition includes all *CLTC* and *CLTCL1* data except the Kagu/Sunbittern insert. Cracraft=Cracraft et al. (2004), FH=Fain and Houde (2004), Ericson=Ericson et al. (2006), Tapestry=Sibley and Ahlquist (1990), MC=Mayr and Clarke (2003), LZ=Livezey and Zusi (2007).
[b] Range of distances to a set of random trees constrained to include all well-accepted clades.

Since Robinson–Foulds distances are the number of branches that differ between trees they are related to the number of clades that differ between two trees. Still, many studies focus on the presence or absence of specific clades (e.g. Edwards et al., 2002). Our phylogeny is not congruent with the basal divergence of Fain and Houde (2004; see also Ericson et al., 2006). Nor is our phylogeny congruent with the basal divergence in Livezey and Zusi (2007) or that found in studies of whole mitochondrial genomes (Watanabe et al., 2006; Gibb et al., 2007), emphasizing the lack of consensus regarding the basal divergences among Neoaves.

3.5. Power of phylogenetic estimation using introns and exons

Given the large differences within the clathrin heavy chain partitions and between our phylogeny and published avian phylogenies, we wanted to determine whether our dataset (or other published datasets) were likely to have sufficient power to correctly resolve relationships along the short internodes at the base of Neoaves. A simple method of phylogenetic power analysis is to determine the amount of sequence data necessary to be confident that at least one synapomorphic change along a short internal branch occurred (Braun and Kimball, 2001). The internal branches at the base of Neoaves are short in absolute terms, with different branches reflecting 0.5 to 5 million years (see branch lengths in Fig. S3). The median rate of intron evolution is ~0.0014 substitutions per site per million years, so the minimum length of sequence necessary to be 95% certain of a single synapomorphy along the shortest internal branches is ~4 kb. In sharp contrast, the median rate of exon evolution is ~0.0004 substitutions per site per million years, suggesting ~15 kb of exon sequence are necessary. However, we note that the synapomorphy may not persist through time and that multiple synapomorphies will be necessary to support relationships, indicating that substantially more sequence data will likely be necessary to accurately reconstruct multiple divergences during a short period of time.

To extend the previous analysis, which provides an absolute minimum sequence length, we used simulation to explore the power of a heterogeneous dataset to resolve the base of Neoaves in the context of the complete tree. The simulated dataset had three regions of equal length that evolved at the intronic rate and a fourth that evolved under a codon model at the exonic partition rate, like the clathrin heavy chain dataset. These simulations (Fig. 2A) suggest that at least 16 kb of data are required to approach the correct tree (within 3 branches of the correct tree); even larger datasets (greater than 32 kb) of data evolving at this rate will be necessary to reliably obtain the correct tree.

Although many datasets contain a heterogeneous mixture of introns and exons, we also wanted to examine the performance of introns and exons independently. For similar amounts of sequence data, analyses of introns show much smaller Robinson–Foulds distances from the correct tree than analyses of exons (Fig. 2B). Given that some studies using nuclear sequences have focused on exon data (e.g. Groth and Barrowclough, 1999; Chubb, 2004), it is of interest that estimates of phylogeny using the simulated exon data are as different from the true tree as random trees when limited amounts of data are collected (e.g., under 2 kb). This suggests that substantial exon data are required to overcome stochastic variation in phylogenetic estimation.

It is well known that the evolutionary history of individual genes can differ from the species phylogeny. Thus, it is necessary to examine multiple gene trees to determine the correct species tree. In this context, the fact that *CLTC* and *CLTCL1* intron phylogenies exhibit more similarity than expected by chance (Table 3) is contrary to the predictions given a hard polytomy (Poe and Chubb, 2004). Although this study, like others (Poe and Chubb, 2004; Ericson et al., 2006) used multiple unlinked gene regions, our simulations mimicked the analysis of a single gene. Because many open reading frames are shorter than 2 kb, our power analyses suggest that many genes will have insufficient exon data to accurately estimate the gene tree when very short internodes must be resolved. Combined with our other results, these simulations provide strong evidence that introns have greater potential than exons to reconstruct basal avian phylogeny.

3.6. Conclusions

The phylogenetic relationships among orders within Neoaves (as well as the monophyly of some orders) have been the subject of substantial debate (Cracraft et al., 2004; Harshman, 2007), largely due to short internodes and the apparent rapid radiation of this group (e.g., van Tuinen et al., 2000). Poe and Chubb (2004) interpreted their failure to reject the null hypothesis of independent evolution for the gene trees they examined as evidence that Neoaves represents a hard polytomy. However, the observation that the unlinked *CLTC* and *CLTCL1* intron partitions exhibit more similarity than expected by chance (Table 3) provides evidence that there is structure at the base of Neoaves. Furthermore, the fact that some published trees exhibited more similarity for basal nodes than expected by chance (Table 4), even if that similarity is relatively limited, provides additional corroboration for the hypothesis that Neoaves can be resolved.

The short internodes and low bootstrap support for many key groups within the Neoaves suggest this study does not have enough data to draw definitive conclusions regarding specific phylogenetic relationships at the base of Neoaves. Consideration of evolutionary rates and the use of simulations corroborated this hypothesis by showing that the amounts of nuclear sequence data that have been used to examine relationships among avian orders is unlikely to be sufficient to fully resolve Neoaves. Our analyses of clathrin heavy chain genes, combined with our analysis of evolutionary rates and the use of simulations, suggest that a strategy placing greater emphasis on the collection of intron data is likely to have the greatest potential to resolve relationships at the base of Neoaves. The approaches we used to examine the utility of introns for resolving the base of Neoaves should be useful in other groups as well.

Acknowledgments

This paper benefited from helpful suggestions by members of the Kimball/Braun lab and John Harshman. We are grateful to Tamaki Yuri for help with indel analyses and to the museums (Table S1) that supplied tissues. This research was funded by a National Science Foundation grant (DEB-0228682) to R.T.K., E.L.B. and D.W. Steadman and facilitated by additional grants to the EarlyBird consortium (DEB-0228675, DEB-0228688, and DEB-0228617).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2007.11.016.

References

Bleiweiss, R., 1998. Fossil gap analysis supports early tertiary origin of trophically diverse avian orders. Geology 26, 323–326.

Braun, E.L., Kimball, R.T., 2001. Polytomies, the power of phylogenetic inference, and the stochastic nature of molecular evolution: a comment on Walsh et al. (1999). Evolution 55, 1261–1263.

Braun, E.L., Kimball, R.T., 2002. Examining Basal avian divergences with mitochondrial sequences: model complexity, taxon sampling, and sequence length. Syst. Biol. 51, 614–625.

Chubb, A.L., 2004. New nuclear evidence for the oldest divergence among neognath birds: the phylogenetic utility of ZENK (i). Mol. Phylogenet. Evol. 30, 140–151.

Cracraft, J., et al., 2004. Phylogenetic relationships among modern birds (Neornithes): toward an avian tree of life. In: Cracraft, J., Donoghue, M.J. (Eds.), Assembling the tree of life. Oxford University Press, Inc., New York, pp. 468–489.

Dyke, G.J., van Tuinen, M., 2004. The evolutionary radiation of modern birds (Neornithes): reconciling molecules, morphology and the fossil record. Zool. J. Linn. Soc. 141, 153–177.

Edwards, S.V., Fertil, B., Giron, A., Deschavanne, P.J., 2002. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. Syst. Biol. 51, 599–613.

Ericson, P.G., et al., 2006. Diversification of Neoaves: integration of molecular sequence data and fossils. Biol. Lett. 2, 543–547.

Fain, M.G., Houde, P., 2004. Parallel radiations in the primary clades of birds. Evolution Int. J. Org. Evolution 58, 2558–2573.

Gibb, G.C., Kardailsky, O., Kimball, R.T., Braun, E.L., Penny, D., 2007. Mitochondrial genomes and avian phylogeny: complex characters and resolvability without explosive radiations. Mol. Biol. Evol. 24, 269–280.

Groth, J.G., Barrowclough, G.F., 1999. Basal divergences in birds and the phylogenetic utility of the nuclear RAG-1 gene. Mol. Phylogenet. Evol. 12, 115–123.

Harshman, J., 1994. Reweaving the Tapestry — what can we learn from Sibley and Ahlquist (1990). Auk 111, 377–388.

Harshman, J., 2007. Classification and phylogeny of birds. In: Jamieson, B.G.M. (Ed.), Reproductive biology and phylogeny of birds. Science Publishers, Inc., Enfield, NH, pp. 1–35.

Le Hir, H., Nott, A., Moore, M.J., 2003. How introns influence and enhance eukaryotic gene expression. Trends Biochem. Sci. 28, 215–220.

Livezey, B.C., Zusi, R.L., 2007. Higher-order phylogeny of modern birds (Theropoda, Aves: Neornithes) based on comparative anatomy. II. Analysis and discussion. Zool. J. Linn. Soc. 149, 1–95.

Maddison, D., Maddison, W., 2000. MacClade 4: Analysis of Phylogeny and Character Evolution. Sinauer Associates, Inc., Sunderland, MA.

Mayr, G., Clarke, J., 2003. The deep divergences of neornithine birds: a phylogenetic analysis of morphological characteristics. Cladistics 19, 527–553.

Murphy, W.J., Eizirik, E., Johnson, W.E., Zhang, Y.P., Ryder, O.A., O'Brien, S.J., 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409, 614–618.

Poe, S., Chubb, A.L., 2004. Birds in a bush: five genes indicate explosive evolution of avian orders. Evolution 58, 404–415.

Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14, 817–818.

Prychitko, T.M., Moore, W.S., 2003. Alignment and phylogenetic analysis of beta-fibrinogen intron 7 sequences among avian orders reveal conserved regions within the intron. Mol. Biol. Evol. 20, 762–771.

Rambaut, A., Charleston, M., 2002. Phylogenetic Tree Editor v1.0. Oxford University, Oxford.

Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Sanderson, M.J., 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 19, 1218–1231.

Sibley, C.G., Ahlquist, J.E., 1990. Phylogeny and Classification of Birds: A Study in Molecular Evolution. Yale University Press, New Haven, CT.

Simmons, M.P., Ochoterena, H., 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol. 49, 369–381.

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

Swofford, D.L., 2003. PAUP*: Phylogenetic Analysis Using Parsimony (* and other Methods). Sinauer, Sunderland, MA.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.

Thorne, J.L., Kishino, H., 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol. 51, 689–702.

van Tuinen, M., Sibley, C.G., Hedges, S.B., 2000. The early history of modern birds inferred from DNA sequences of nuclear and mitochondrial ribosomal genes. Mol. Biol. Evol. 17, 451–457.

Wakeham, D.E., Abi-Rached, L., Towler, M.C., Wilbur, J.D., Parham, P., Brodsky, F.M., 2005. Clathrin heavy and light chain isoforms originated by independent mechanisms of gene duplication during chordate evolution. Proc. Natl. Acad. Sci. U. S. A. 102, 7209–7214.

Watanabe, M., et al., 2006. New candidate species most closely related to penguins. Gene 378, 65–73.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.

Yang, Z., Nielsen, R., Hasegawa, M., 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. 15, 1600–1611.