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