Deletion Bias in Avian Introns over Evolutionary Timescales

Kevin P. Johnson
Illinois Natural History Survey, Champaign, Illinois

The role that introns play in the function and evolution of nuclear genomes is not fully understood. Recent models of intron evolution suggest that selection and drift may interact to maintain introns in multicellular organisms. In addition, deletion mutations are more likely to become fixed than insertion mutations. Examination of indel substitutions over macroevolutionary timescales in pigeons and doves (Aves: Columbiformes) reveals that deletion substitutions outnumber insertion substitutions by over six times. The length of indel events is variable.

Introduction

Introns are a major component of eukaryotic nuclear genomes. However, the function of introns and the forces maintaining introns over evolutionary timescales are only beginning to be explored (Duret 2001). The evolution of intron size is only poorly understood. The role of selection on intron size is a matter of debate (Clark, Leicht, and Muse 1996; Carvalho and Clark 1999; Comeron and Kreitman 2000; Duret 2001; Lynch 2002; Waltari and Edwards 2002). Intron size can be quite stable over long evolutionary timescales as shown in some vertebrates (Waltari and Edwards 2002). In Drosophila, intron size increases in regions of low recombination. Comeron and Kreitman (2000) suggest that there is indirect selection for large introns in regions of low recombination because they can reduce the load caused by deleterious mutations by increasing the recombination rate. However, both very small and very large introns tend to be more prevalent in regions of low recombination. Given that selection is generally weaker in regions of low recombination, Carvalho and Clark (1999) suggested that selection may act directly on introns to stabilize their size. However, Lynch (2002) pointed out that the reduced efficiency of selection in regions of low recombination may lead to an increase in intron size if small introns provide a slightly improved transcription efficiency or splicing accuracy. Resolving this debate will require an understanding of the underlying mutation rates and substitution properties of insertion and deletion events (indels) in introns. Studies of nuclear pseudogenes and their functional counterparts in humans and mice have revealed that deletions outnumber insertions by about 2.7 to 1 (Ophir and Graur 1997). Similarly, deletions were estimated to outnumber insertions by 2 to 1 in nuclear pseudogenes and spacer regions in primates (Saitou and Ueda 1994). In Drosophila, polymorphic sites indicate that deletions outnumber insertions by 1.35 to 1 (Comeron and Kreitman 2000). Polymorphisms are presumably a closer reflection of the underlying mutation bias, because selection is relatively stronger over evolutionary timescales such that deleterious mutations may be eliminated (Lynch 2002). Both mutation and selection interact to determine the substitution rates of insertions and deletions over evolutionary timescales. Recent theoretical work has suggested that there are reasons why insertion events in introns might be deleterious relative to deletion events, including reduced transcription efficiency and decreased splicing accuracy (Lynch 2002). However, it is unclear how these selection biases might interplay with underlying mutation biases.

A detailed examination of the substitution pattern of indels in introns over evolutionary timescales has not been conducted. Prychitko and Moore (2003) examined molecular evolution in the β-fibrinogen intron 7 across 28 species of birds. Although they did report a deletion bias, it was not quantified. Mapping indels over a phylogeny provides polarity information, which allows identification of indels as either insertions or deletions. In the present study, I reconstruct a phylogeny for pigeons and doves (Aves: Columbiformes) using mitochondrial protein coding and nuclear intron DNA sequences. I then reconstruct the history of indel substitutions for the β-fibrinogen intron 7 over the phylogeny, using parsimony. This reconstruction provides an assessment of the relative frequency and size distribution of indel events over evolutionary timescales. In addition, this study provides an assessment of the degree of convergence (or homoplasy) of intron indels.

Materials and Methods

DNA was extracted using a Qiagen DNAeasy Tissue kit with manufacturer’s protocols from frozen muscle tissue of 78 species and subspecies of pigeons and doves (Aves: Columbiformes) and an outgroup taxon (Aerodramus salanganus). Polymerase chain reaction (PCR) was used to amplify portions of the mitochondrial cytochrome b (cyt b) and ATPase8 genes, as well as the entire nuclear β-fibrinogen intron 7 (FIB7) gene. Primers and PCR protocols followed Johnson and Clayton (2000a) for cyt b and FIB7. The primers BRUs and t-lys (Fleischer et al. 2000) were used to amplify ATPase8. The PCR amplifications were purified using a Qiagen PCR Purification kit with manufacturer’s protocols. These cleaned PCR products were sequenced using cycle sequencing and were visualized as described by Johnson and Clayton (2000a). Complementary chromatograms were resolved using Sequencher 3.1 and aligned across species by eye using the same program. Because the two mitochondrial genes were protein coding, alignment was very straightforward. In addition, indels in the FIB7 gene were relatively infrequent (see Results) so this region could also be reliably aligned by eye (alignment is shown in the online Supplementary Material). All sequences were deposited in
GenBank, and many of the sequences were published previously (Johnson and Clayton 2000a, 2000b; Johnson et al. 2001; accession numbers AY443577–AY443699).

Phylogenies were reconstructed from the combined sequence data using both parsimony and Bayesian maximum likelihood searches. FIB7 indels were initially treated as missing data in the phylogenetic analysis, such that they could be independently reconstructed over the tree in subsequent analyses. In parsimony analyses, 100 random addition replicate searches with all characters unweighted and unordered were conducted using PAUP* (Swofford 2001). To assess the sensitivity of this tree to character resampling (Felsenstein 1985), 1,000 bootstrap replicate searches were conducted. Bayesian maximum likelihood analyses were conducted using MrBayes (Huelsenbeck 2001) with a GTR + I + G model. Posterior probabilities for various nodes in this tree were calculated sampling trees every 1,000 generations from a chain of 2 million generations and ignoring the first 500,000 burn-in trees. Two runs were performed and the average posterior probability is reported.

Using the alignment of β-fibrinogen intron 7, the evolutionary timing of insertion and deletion events within the ingroup (Columbiformes) was evaluated with respect to the base position of the aligned sequences. Indels involving the same position and number of base pairs were considered homologous. With MacClade (Maddison and Maddison 1992), parsimony was used to reconstruct the number of indel events, and to determine whether these events were insertions or deletions. For each indel, all most parsimonious optimizations were examined. A string of T repeats at aligned positions 559–565 was not included in this analysis because of difficulty in evaluating the homology of various T indels.

Results

Parsimony analysis of the combined sequence data recovered two most parsimonious trees (fig. 1). The difference between these trees involved a single rearrangement that did not affect the outcome of any of the indel character mapping. Most of the nodes in this tree were supported by greater than 50% of bootstrap replicates. Bayesian maximum likelihood analysis produced a very similar tree (not shown). Differences between this tree and the parsimony tree did not affect the results of the indel character mapping. Most of the nodes in this tree received higher than 90% posterior probability support (fig. 1).

Parsimony mapping of insertion/deletion events in FIB7 over the tree indicated a very high consistency for these characters (mean CI = 0.912). In fact, only 2 of 48 indels exhibited any degree of homoplasy; the remaining 46 constituted single events (see fig. 1). One of the homoplasious deletions (position 533) involved a weakly supported node (that uniting Macropygia with Patagioenas), and an alternate arrangement of this node would eliminate the homoplasy of this deletion event. The 8-bp deletion at position 1046 of the aligned sequence occurred repeatedly in a variety of taxa, and this deletion involved the loss of a single repeat of an 8-bp segment (AGAATCAT). Pyritchko and Moore (2003) reported a 77-bp insertion in Columbiformes relative to other avian taxa. Although this region was found in the present study in comparison to the outgroup Aerodramus, with the current outgroup sampling this indel cannot be polarized, and this region is not consistently 77 bp across all Columbiformes.

Deletions outnumbered insertions 44 to 7 (or 43 to 8 in an alternate equally parsimonious reconstruction). The proportion of indels that were deletions (86.3%) was significantly greater than chance (Sign test: \( P < 0.0001 \)) and also significantly greater than 73% (Wilcoxon signed rank test, \( P < 0.0005 \)), which is the highest bias previously reported for nuclear deletion events (Ophir and Graur 1997). The size of intron indels ranged from 1 to 167 base pairs (fig. 2). The mean insertion length (3.9 bp) was smaller than the mean deletion length (12.2 bp), but this difference was not significant (\( r \)-test, \( P = 0.48 \)).

Although some deletions involved deletion of repeat units, insertions always involved repeat elements. All four single-bp insertions were a repeat of the previous base. The insertion at position 386 was a 3-bp repeat of the previous sequence GAT. The 14-bp insertion at position 891 was a repeat of the previous 14 bp. Finally, a 6-bp insertion at position 460 involved a repeat of the previous 5 bp AAGTA, with an additional insertion of a G becoming AAGTGA.

Discussion

Mapping of insertion/deletion events in β-fibrinogen intron 7 over a phylogeny for pigeons and doves (Aves: Columbiformes) revealed that deletions outnumbered insertions by a ratio of more than 6 to 1. This is more than double the mutation ratio estimated from nuclear pseudogenes in previous studies (Saitou and Ueda 1994; Ophir and Graur 1997; Comeron and Kreitman 2000). Mapping of indels over phylogenetic timescales reflects both mutation and selection biases. Given that the ratio of deletions to insertions for polymorphisms in Drosophila is only 1.35 to 1, the high deletion bias detected in the present study indicates that selection may have an important effect on indels in nuclear introns. Lynch (2002) predicted that intron insertions should be more deleterious than deletions because of reduced transcription and splicing efficiency. The deletion bias found in FIB7 in Columbiformes supports the predictions of Lynch (2002).

The size of indels was variable, but it did not differ significantly between insertions and deletions. Large insertions and deletions occur in a single mutational step, and there is no evidence of intermediates for these large indels. The prevalence of repeat elements in indel regions, especially for insertions, supports the slipped-strand mispairing mechanism for the origin of insertion events (Levinson and Gutman 1987). This mechanism may be an important process in intron regions, and not just in regions where repeats are already prevalent (Kelchner 2000).

Given the strong bias toward deletion substitutions, what keeps the introns from disappearing over evolutionary time and intron size relatively stable (Waltari and Edwards 2002)? It may be that once introns get below
Fig. 1.—One of two most parsimonious trees (length = 5,863, consistency index [CI] = 0.296) from unweighted analysis of combined cyt b, ATPase8, and FIB7 gene sequences. Branch lengths proportional to the number of changes. An * indicates branches supported in >50% of parsimony bootstrap replicates and which also received >90% Bayesian posterior probability. A † indicates branches receiving >50% parsimony bootstrap support only. A # indicates branches receiving >90% Bayesian posterior probability only. Branches with no symbol did not receive high support from either of these measures. Bold hash marks indicate branches on the tree on which indels of the FIB7 gene occurred. The first number indicates the aligned position for the start of the indel, and the second number indicates the size and direction of the indel (positive = insertion, negative = deletion). Italics indicate indels that exhibited homoplasy.
a certain size, mutations that knock out the splice sites become more likely (Lynch 2002). Over evolutionary time, such indirect selection may counteract the bias toward deletion substitutions in introns. Much more work is needed on the mutation properties of nuclear introns and the fate of these mutations over evolutionary time.

Acknowledgments

I thank the following individuals and institutions who provided tissue specimens for sequencing: D. Clayton, S. de Kort, M. Kennedy, Field Museum of Natural History, University of Kansas Museum of Natural History, Louisiana State University Museum of Natural Science, University of Washington Burke Museum, U.S. National Museum of Natural History, and the Tracy Aviary. This work was supported in part by National Science Foundation grant DEB-0107891.

Literature Cited


Huelsenbeck, J. P. 2001. MrBayes 1.10 (Bayesian Analysis of Phylogeny). University of Rochester, Rochester, N.Y.


Scott Edwards, Associate Editor

Accepted November 3, 2003