

Structural Conservation and Variation in the Mitochondrial Control Region of Fringilline Finches (*Fringilla* spp.) and the Greenfinch (*Carduelis chloris*)

H. Dawn Marshall and Allan J. Baker

Centre for Biodiversity and Conservation Biology, Royal Ontario Museum, and Department of Zoology, University of Toronto

We sequenced the entire control region and portions of flanking genes (tRNA^{Phe}, tRNA^{Glu}, and ND6) in the common chaffinch (*Fringilla coelebs*), blue chaffinch (*F. teydea*), brambling (*F. montifringilla*), and greenfinch (*Carduelis chloris*). In these finches the control region is similar in length (1,223–1,237 bp) and has the same flanking gene order as in other birds, and contains a putative TAS element and the highly conserved CSB-1 and F, D, and C boxes recognizable in most vertebrates. Cloverleaf-like structures associated with the TAS element at the 5' end and CSB-1 at the 3' end of the control region may be involved with the stop and start of D-loop synthesis, respectively. The pattern of nucleotide and substitution bias is similar to that in other vertebrates, and consequently the finch control region can be subdivided into a central, conserved G-rich domain (domain II) flanked by hyper-variable 5'-C-rich (domain I) and 3'-AT-rich (domain III) segments. In pairwise comparisons among finch species, the central domain has unusually low transition/transversion ratios, which suggests that increased G+T content is a functional constraint, possibly for DNA primase efficiency. In finches the relative rates of evolution vary among domains according to a ratio of 4.2 (domain III) to 2.2 (domain I) to 1 (domain II), and extensively among sites within domains I and II. Domain I and III sequences are extremely useful in recovering intraspecific phylogeographic splits between populations in Africa and Europe, Madeira, and a basal lineage in Nefza, Tunisia. Domain II sequences are highly conserved, and are therefore only useful in conjunction with sequences from domains I and III in phylogenetic studies of closely related species.

Introduction

The control region is usually considered to be the most variable portion of the mitochondrial DNA (mtDNA) molecule (reviewed by Simon 1991). However, between two species of rat, Saccone, Attimonelli, and Sbisà (1987) found levels of divergence similar to those among protein-coding genes, although more distant comparisons (e.g., rat to mouse) showed the expected elevated differences. Furthermore, interspecific divergences between six species of *Jalmenus* butterflies were too low to be useful for phylogeny construction, despite little conservation in primary sequence among Lepidopteran genera (Taylor et al. 1993). In such cases, putatively functional features of the control region can be difficult to identify, underscoring the need for comparative data from a range of levels of phylogenetic separation (Taylor et al. 1993).

Among vertebrates, the control region varies in length from 0.73 kb (white sturgeon; Buroker et al. 1990) to 2.1 kb (*Xenopus*; Saccone, Attimonelli, and Sbisà 1987), and is flanked by the genes for tRNA^{Pro} (tRNA^{Glu} in birds) and tRNA^{Phe}. Most primary sequence variability is apportioned between two hypervariable domains flanking a conserved central domain (Brown et al. 1986), and consists of substitutions, small indels, large deletions or duplications, and variation in copy number of tandem repeats. The domains are further characterized by differences in base composition and by the presence of particular conserved motifs and putative

secondary structures. The central domain, low in L-strand adenine, is responsible for the formation of the three-strand displacement (D-) loop structure (Clayton 1991). The domain closest to the tRNA^{Pro} (tRNA^{Glu} in birds) gene is characterized by high adenine and low guanine content, and short termination-associated sequences (TASs; Doda, Wright, and Clayton 1981) near a potential cloverleaf structure and the 3' terminus of the D-loop. The tRNA^{Phe}-adjacent domain contains a conserved sequence block (CSB-1) associated with another cloverleaf structure and the 5' end of the D-loop, as well as the origin for heavy strand replication (O_H), and the light (LSP) and heavy (HSP) strand transcription promoters (Walberg and Clayton 1981; Clayton 1982, 1984). High in adenine, it also tends to be the most variable domain and the one in which length variation due to repeat structure is found (Brown et al. 1986; Saccone, Attimonelli, and Sbisà 1987).

Avian control region studies (reviewed by Baker and Marshall 1997) to date have largely used these sequences to elucidate population structure, although enough comparative data exist to reveal a number of structural or functional features. Birds differ from other vertebrates in that the 5'-flanking gene is tRNA^{Glu} instead of tRNA^{Pro}; this results from a mitochondrial rearrangement involving the ND6 gene (Desjardins and Morais 1990). Additional differences include the presence of one bidirectional transcription promoter rather than a separate one for each strand (L'Abbé et al. 1991), and the occurrence of CSB-1-like repeats in Galliformes (Desjardins and Morais 1990, 1991) and Anseriformes (Ramirez, Savoie, and Morais 1993). Rates and patterns of variability are similar to those in other vertebrates, and the control region has proven useful in revealing previously undetectable genetic structure within and among such closely related taxa as subspecies of dunlins

Key words: mitochondrial control region, conserved sequence blocks, secondary structure, variability, *Fringilla*, *Carduelis*.

Address for correspondence and reprints: H. Dawn Marshall, Centre for Biodiversity and Conservation Biology, Royal Ontario Museum, 100 Queen's Park, Toronto, Ontario M5S 2C6, Canada. E-mail: hdm@zoo.utoronto.ca.

(Wenink, Baker, and Tilanus 1993; Wenink et al. 1996), populations and subspecies of grey-crowned babblers (Edwards 1993a), and populations of lesser snow geese (Quinn 1992).

Although partial control region sequences of babblers have been studied, and the complete sequence has been obtained from several nonpasserine birds (Quinn and Wilson 1993), no complete passerine control regions have been published. Finches in the genus *Fringilla* are an ideal species group in which to examine the structure, variability, and evolution of the control region of these passerines. *Fringilla* comprises three closely related species, one of which (the common chaffinch, *F. coelebs*) is widely distributed throughout Europe, northern Africa, and the Atlantic Islands. Populations vary in age and degree of isolation from recently established populations experiencing extensive gene flow to morphologically and genetically distinguishable subspecies (Baker et al. 1990). The blue chaffinch (*F. teydea*) of the Canary Islands is the sister species to the common chaffinch, and these two form a sister group with the brambling of Eurasia (Baker and Marshall 1997). Along with the related greenfinch (*Carduelis chloris*) these species provide an excellent opportunity to compare sequences from a range of taxonomic levels. We describe complete sequences and conserved structural features and rates and patterns of variability among common chaffinch, blue chaffinch, brambling, and greenfinch control regions, and compare them to those of other birds for a more comprehensive understanding of avian control region evolution. Additionally, we include samples of sequences representing the two variable flanking domains for four disparate populations of common chaffinches and one population of bramblings. We then assess the utility of the control region for population and phylogenetic studies of these finches.

Materials and Methods

Collection of Samples and DNA

Two specimens (collection localities in brackets) each of the greenfinch (Uppsala, Sweden), the brambling (near Holt, Norway), and the blue chaffinch (Tenerife, Canary Islands) were sequenced for the entire control region. An additional eight bramblings from Holt were examined for portions of the 5'- and 3'-flanking domains. The complete control region sequence was obtained from four common chaffinches, one from Madeira Island (*F. c. maderensis*), one from Rabat, Morocco (*F. c. africana*) and one each from Segovia, Spain and Asiago, Italy (*F. c. coelebs*). Flanking domains were sequenced for 9 more birds from Madeira, 8 from Asiago, 9 (5') or 10 (3') from Nefza, Tunisia (nominally *F. c. africana*), and 14 (5') or 9 (3') from Oslo, Norway (*F. c. coelebs*).

Genomic DNA was extracted from liver, heart, or spleen using standard proteinase K-phenol-chloroform methods (Sambrook, Fritsch, and Maniatis 1989). Briefly, tissues were homogenized in 100 mM Tris-HCl, pH 8.0; 10 mM EDTA; 100 mM NaCl; 0.1% SDS; and 10 μ g/ml proteinase K, and incubated overnight at 55°C.

The homogenate was extracted twice with Tris-saturated phenol and once with chloroform: isoamyl alcohol (24:1). Finally, nucleic acids were precipitated with sodium chloride or sodium acetate and ethanol, and resuspended in distilled water.

Polymerase Chain Reaction (PCR) and Sequencing

PCR primers were developed specifically for the chaffinches, the brambling, and the greenfinch as follows. First, amplifications were performed using three primers matching portions of the flanking genes for tRNA^{Pro} (CRTPRO), ND6 (FND6; P. Boag, personal communication), and tRNA^{Phe} (H1261; Wenink, Baker, and Tilanus 1994). Subsequently, specific internal primers were made for the 5'- or tRNA^{Glu}-adjacent (FCRI5' and FCRI3') and 3'- or tRNA^{Phe}-adjacent (F304) regions. Another primer, GSLGLU (P. Boag, personal communication), was used for some amplifications of the latter region. Primer sequences, positions, and direction are given in Baker and Marshall (1997). For the 5' region, FCRI5' was used as a sequencing primer; for the 3' region, either GSLGLU or F304 and H1261 were used. Some of these primers were used to produce sequence from flanking genes.

Double-stranded amplification reactions contained 10 mM Tris-HCl, pH 8.3; 1.5 mM MgCl₂; 50 mM KCl; 50 μ M each dNTP; 0.4 μ M each primer; and 1 U *Taq* DNA polymerase (Boehringer Mannheim) in a 25- μ l volume. Amplification was achieved through a thermal cycle of 93°C for 30 s, 48–50°C for 30 s, and 72°C for 60 s, repeated 35 times. Products were purified using agarose separation followed by binding to glass beads (Gene Clean; BIO 101), and were sequenced using either the Sequenase 2.0 (United States Biochemical) or AmpliCycle (Perkin Elmer) sequencing kit, according to the manufacturer's instructions. The complete control region sequences were obtained from at least two individuals for each species. Additionally, for at least one of these individuals, most of the sequence was attained in both directions or from an additional amplification and sequencing reaction in the same direction. Generally, the population sequences were obtained in one direction only.

Sequence Analysis

Multiple sequence alignment was achieved using the default options of CLUSTAL V (Higgins, Bleasby, and Fuchs 1992). Visualization of secondary structure was facilitated with the programs PCFOLD and MOLE-CULE (Zuker 1989). Sequence length and nucleotide composition, percent sequence similarity, and frequency and distribution of substitutions and indels were obtained from ESEE (Cabot and Beckenbach 1989) or MEGA (version 1.01; Kumar, Tamura, and Nei 1993). To investigate rate variation among sites in the control region, we used the approximate method of Yang and Kumar (1996) as implemented in the PAML package (Yang 1995) to estimate the average number of substitutions per site ($\bar{\mu}$) along the tree of the four species. The α parameter of the gamma distribution was estimated using the maximum likelihood method of the

same package. The input tree was (((chaffinch, blue chaffinch) brambling) greenfinch).

To assess the population genetic and phylogenetic utility of the control region in the finch sequences, we constructed neighbor-joining trees (Saitou and Nei 1987) and calculated bootstrap confidence levels and branch length confidence probabilities using MEGA. Sequence divergence values were calculated using either the Jukes-Cantor (Jukes and Cantor 1969), Kimura two-parameter (Kimura 1980), or Tajima-Nei (Tajima and Nei 1984) algorithms, ignoring alignment gap sites in pairwise comparisons, according to the criteria discussed by Kumar, Tamura, and Nei (1993). That is, when the Jukes-Cantor estimate of d was 0.05 or less or between 0.05 and 0.3 and accompanied by a low (<2) transition-to-transversion (ts/tv) ratio, the Jukes-Cantor estimate was used; when d was between 0.05 and 0.3 and the ts/tv ratio was high (>2) the Kimura two-parameter estimate was used; and when d was between 0.3 and 1 the Tajima-Nei distance was used. The combination of $d > 0.3$ and extensive rate variation among sites ($\alpha < 1$), which would necessitate the use of gamma distances, did not occur. For intraspecific comparisons, haplotypic diversity (h) and nucleotide diversity (π ; Nei and Li 1979) were calculated. Tajima's D (Tajima 1989) was calculated to test whether the chaffinch sequences conform to the expectations of neutral theory.

Results

Evidence Consistent with a Mitochondrial Origin of the Control Region Sequences

Contiguous sequence spanning the 3' terminus of the ND6 gene, the tRNA^{Glu} gene, the entire control region, and a small portion at the 5' end of the tRNA^{Phe} gene was obtained from 10 finches representing four species in two genera. For ND6, between 8 and 45 bp were sequenced; they show 60.0% to 65.8% similarity to the corresponding region in chicken mtDNA, and translate appropriately with the mitochondrial genetic code, having the same start codon as the chicken (ATG) and revealing no frameshift or nonsense mutations. The tRNA^{Glu} gene is 71 bp long in all individuals, and is between 55.6% and 61.1% similar to the chicken gene (with alignment gaps counted as mismatches) with no mismatches occurring in the anticodon loop (GAA). Additionally, when analyzed with PCFOLD (Zuker 1989), the proper cloverleaf structure for this gene was retrieved ($\Delta G = -18.4$ kcal/mol). Sequence similarity among finches ranges from 74.7% between the blue chaffinch and the greenfinch to 97.2% between the brambling and the chaffinch. When all sequences are compared, transition substitutions occur at 21 positions and there are no transversions. Although very little of the tRNA^{Phe} gene was sequenced (9–15 bp), it again appears to be homologous to the chicken gene in that nucleotide type (purine or pyrimidine) is maintained at each position, and sequence similarity is 60%–61.5% where 13–15 bp are available for comparison.

Primary Structure and Conserved Sequence Motifs

The alignment of the control region sequences of the four finch species is presented in figure 1 (GENBANK accession numbers are U76250 for the common chaffinch, U76249 for the blue chaffinch, U76251 for the brambling, and U56075 for the greenfinch). Uncorrected pairwise sequence similarity is 95.5%–99.6% among common chaffinches, 99.7% between the two bramblings and 100% between the two blue chaffinches and the two greenfinches. The length of the control region ranges from 1,223 bp in the blue chaffinch to 1,237 bp in the greenfinch. Both the brambling and the common chaffinch control regions are 1,233 or 1,234 bp. No obvious repeat structure, such as that identified in some species of Charadriiform birds (Berg, Moum, and Johansen 1995), was identified, although a long tract of C nucleotides near the 5' end of the control region is apparent. The among-species alignment is characterized by substitutions at 269 sites and alignment gaps at 37. Almost half (49.1%) of the substitution sites contain a transversion, and the alignment gaps are generally 1–2-bp indels.

A plot of number of variable sites in nonoverlapping 50-bp segments (fig. 2) was used to examine the distribution of variation across the control region. For descriptive purposes, we delimited domains to encompass regions of differing variability as follows: 5' domain I—bases 1–400; central domain II—bases 401–873; and 3' domain III—bases 874–1252. The frequency of both substitutions and gaps is highest in the third domain and lowest in the central region; base composition differences are also apparent among domains (table 1). Estimates of the α parameter of the gamma distribution and the average number of substitutions per site for the whole control region and for each domain are given in table 2. Consistent with figure 2, the estimate of number of substitutions per site is highest in the third domain and lowest in the central domain. The extent of rate variation among sites (inversely related to α) is highest in domain II, still extensive but lower in domain I, and lowest in the third domain.

Finches have conserved primary sequence motifs characteristic of the three domains in other vertebrates (Southern, Southern, and Dizon 1988). In the central domain, the F, D, and C boxes are readily recognizable (fig. 1), although quite divergent from the homologous features of nonpasserines. They share 56%–60% sequence similarity with the same regions in the chicken, values which exceed the chicken/human comparison only for the C box (58.6% vs. 44.4%). These boxes are, however, invariant among the species of the genus *Fringilla*, and are highly conserved (92%–96.6%) in comparisons to the greenfinch (*Carduelis*). The 28-bp CSB-1, spanning positions 876–903 of the among-species alignment, is also highly conserved among chaffinches and the greenfinch (four variable sites), and is 75.0%–78.6% similar to the chicken CSB-1. Neither CSB-2 nor CSB-3 could be unambiguously identified. The consensus TAS (TACATtAAAaYYYAAT where Y = C or T and lower case indicates an invariant base at

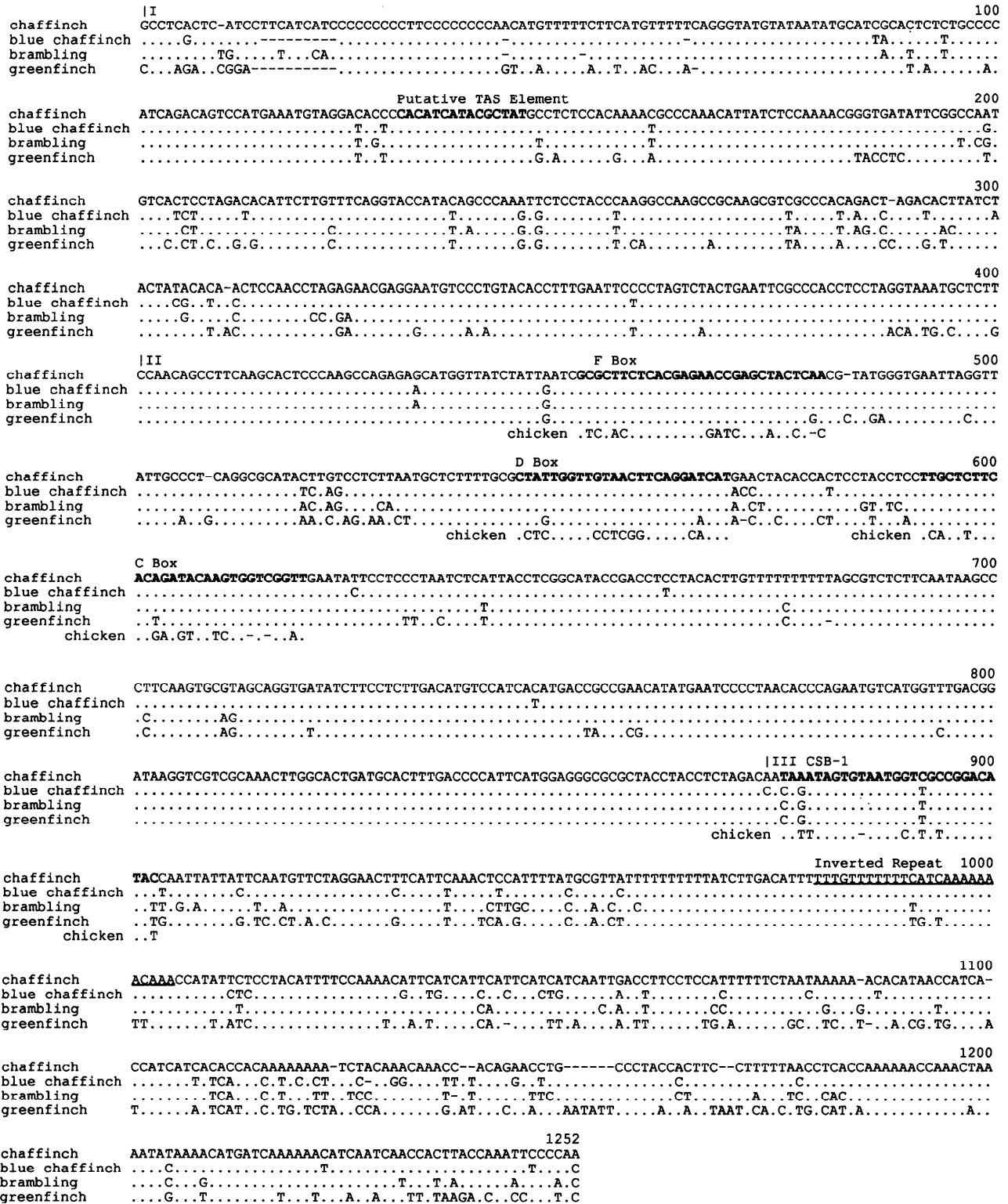


FIG. 1.—The alignment of the control region for the four finch species. Domains are indicated by a vertical dash above the first base of each domain followed by the number of the domain. Primary sequence features (TAS, F, D, and C boxes, and CSB-1) are shown in bold. The inverted repeat associated with putative secondary structure is underlined. For CSB-1 and the F, D, and C boxes the corresponding sequence from the chicken (Desjardins and Morais 1991) is shown.

an indel site) described by Foran, Hixson, and Brown (1988) could not be located in the finch domain I sequences. However, the sequence 5'-CACATCA-TACGCTAT-3' located at position 131 in the among-

species alignment might function as a TAS in finches, given that it directly follows the secondary structure formation in this domain (see below and fig. 3), and is conserved among all four species.

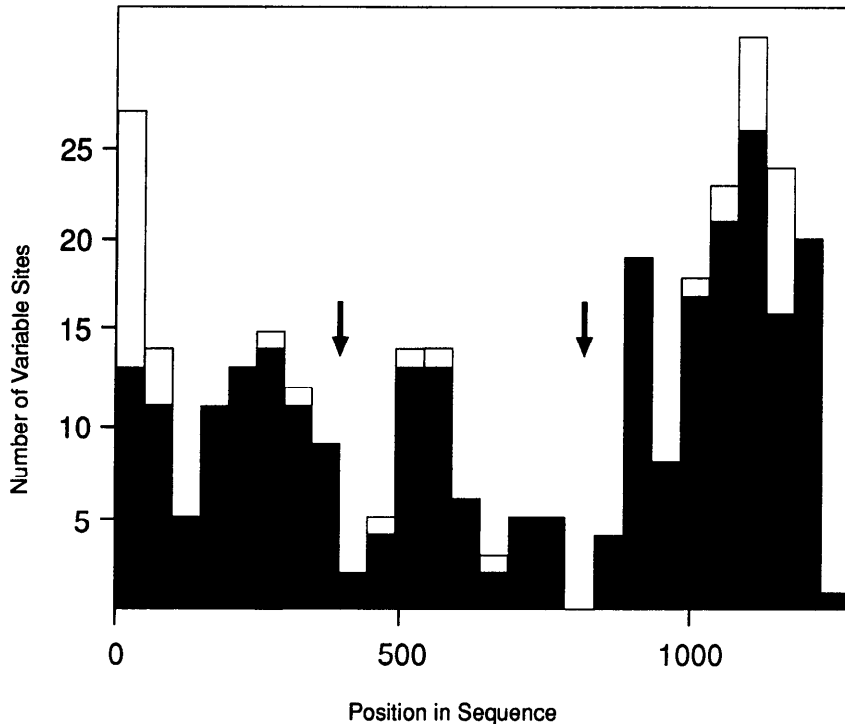


FIG. 2.—Plot of finch control region variability in 50-base windows. Substitutions are indicated with solid bars, and indels with clear ones. Arrows denote domain boundaries.

Potential Secondary Structure

In the finches an AT-rich sequence occurs about 150 bp downstream of CSB-1, and we searched for possible cruciform-forming sequences (inverted repeats) indicative of the bidirectional transcription promoter described in the chicken (L'Abbé et al. 1991). One was found ($\Delta G = -5.9$ kcal/mol) in the *Fringilla* species at position 980 of the multiple alignment (fig. 1), although it shares no primary sequence homology with the promoter of other birds. Other putative secondary structures were identified as follows. In domain I of the chaffinch, a stem-and-loop and an adjacent cloverleaf (fig. 3A; $\Delta G = -11.8$ kcal/mol) span positions 42–130 of the multiple alignment, directly upstream of the putative TAS. The brambling and greenfinch sequences in this region are capable of forming two stem-and-loops followed by a partial (two-loop) cloverleaf-like structure ($\Delta G = -12.3$ and -14.5 kcal/mol for the brambling and greenfinch structures, respectively) which incorporates the

TAS. In the blue chaffinch, the partial cloverleaf is situated between two stem-and-loop formations ($\Delta G = -9.3$ kcal/mol overall), the latter of which encompasses most of the TAS. Domain II and the first third of domain III of all four species of finches are characterized by extensive potential secondary structure formation, including a partial cloverleaf (fig. 3B; $\Delta G = -13.4$ to -13.9 kcal/mol) occupying bases 818–897 of the multiple alignment and incorporating CSB-1 at its 3' end.

Interspecific Variation

For the entire control region, pairwise sequence divergence values range from 7.8% (common chaffinch and blue chaffinch) to 19.6% (common chaffinch and greenfinch). The proportion of transversions increases with sequence divergence; transitions are four times more numerous than transversions between the common and blue chaffinches, but in any comparison with the greenfinch, transversions outnumber transitions. Consistent with the reduced G-content of the mitochondrial genome, C \leftrightarrow T transitions greatly outnumber A \leftrightarrow G transitions, and A \leftrightarrow C or A \leftrightarrow T transversions are more

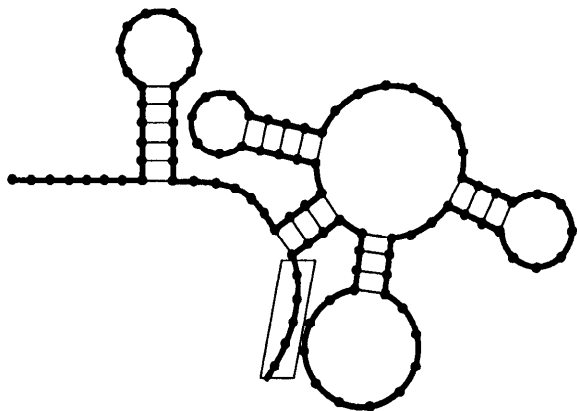
Table 1
Average Base Composition of the Entire Control Region and the Three Domains in Fringilline Finches and the Greenfinch

SEGMENT	NUCLEOTIDE FREQUENCIES (%)			
	A	C	G	T
Whole control region (1,252 bp) . . .	28.4	28.9	13.6	29.2
Domain I (400 bp)	26.5	32.9	14.4	26.2
Domain II (473 bp)	23.7	29.2	18.6	28.6
Domain III (379 bp)	36.4	24.3	6.3	33.1

Table 2
Estimation of the α Parameter of the Gamma Distribution and the Average Number of Substitutions per Site Along the Tree (μ) in Finches

Region	$\bar{\mu}$	α
Whole control region	0.2525	0.4196
Domain I	0.2467	0.4757
Domain II	0.1194	0.1563
Domain III	0.4309	1.0915

A



B

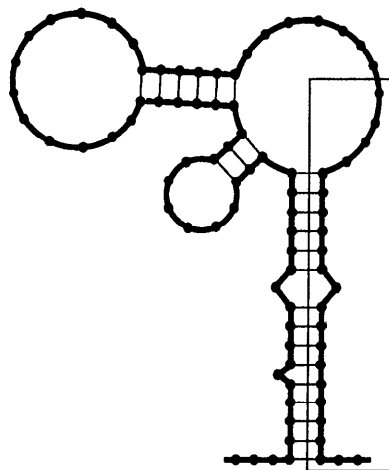


FIG. 3.—Putative secondary-structure formation in the finch control regions. *A*, The cloverleaf associated with domain I in the common chaffinch. The box indicates overlap with the putative TAS element. *B*, The partial cloverleaf formation in domain III in all four finch species. The box indicates bases that are part of CSB-1.

common than $G \leftrightarrow C$ or $G \leftrightarrow T$ transversions. A neighbor-joining tree constructed from the sequence data recovered the among-species relationships also obtained by Baker and Marshall (1997) for a portion of the control region (that is, [[chaffinch, blue chaffinch] brambling] greenfinch]).

Differences in variability among domains are evident from pairwise comparisons. Domain I sequences share similar divergence estimates with the whole control region, but have an elevated proportion of transitions (especially between purines) relative to transversions (particularly those involving A). Sequence divergence in domain II is greatly reduced (2.8%–9.1%), as are ts/tv ratios (0.6–3.3). In particular, $A \leftrightarrow C$ and $G \leftrightarrow T$ transversions are more numerous than in the control region as a whole. Divergences in domain III are almost double those for the whole control region, but ts/tv ratios remain similar to the overall values, reflecting a diminished proportion of transversions involving G in this G-deficient domain (table 1). Relative rate estimates (calculated by averaging among-species pairwise divergence estimates and dividing by the lowest value) for the whole control region and domains I, II, and III are 2.2, 2.2, 1, and 4.2, respectively. Compared with rates for the protein-coding genes cytochrome *b*, NADH dehydrogenase subunit 5, and ATPase 6 in these finches (1.5–1.9; unpublished data), the first domain and the whole control region are slightly faster than the coding genes, but the second domain evolves at a slower rate, and the third domain at a much higher rate. Neighbor-joining trees constructed from each of the three regions (not shown) have the same topology as the entire control region tree, but statistical support varies. For example, using the Kimura two-parameter method of distance estimation, the bootstrap confidence level (BCL, 500 replicates) and branch length confidence probability (CP) of the chaffinch/blue chaffinch node are both 99% for the entire control region and the third domain. Domain I values are still fairly high at 93 (BCL) and 92 (CP),

but for the less informative central domain these values are only 53 (BCL) and 49 (CP).

Intraspecific Variability

The sequence of a 300-bp segment of domain I, spanning positions 110–411 (fig. 1), was obtained for 42 chaffinches, and 22 variable positions define 17 haplotypes (fig. 4). Only the island (Madeira) and one continental (Nefza) populations are characterized by distinct haplotypes. Pairwise divergences are low and haplotypic diversities are high within populations (table 3), except in Madeira and Nefza. In the brambling population, seven haplotypes are denoted by eight variable sites; haplotypic diversity and sequence divergence are similar to those in the Oslo and Asiago populations. Nucleotide diversity is highest in Oslo and Nefza, and lowest in Madeira.

Domain III sequences comprising 285 bp between positions 929–1225 (fig. 1) obtained from 34 chaffinches yield a similar pattern of common and divergent haplotypes among regions (fig. 5). However, despite the presence of nearly twice as many variable sites as in domain I, only 12 haplotypes were identified. Within-population pairwise divergences and nucleotide diversities are lower in domain III than in domain I except in Madeira and Nefza. Other than in Oslo (where only five individuals were sampled) and Madeira, haplotypic diversities are also lower in domain III comparisons. In bramblings, one variable site defines two haplotypes, and average sequence divergence, haplotypic diversity, and nucleotide diversity are all lower in domain III than in domain I (table 3).

Neighbor-joining trees, bootstrap confidence levels (500 replicates), and branch lengths were calculated separately for the 17 domain I haplotypes (fig. 6A) and the 12 domain III haplotypes of the common chaffinch (fig. 6B), using the brambling sequence as an outgroup. In both cases a similar pattern was obtained. All haplotypes other than the divergent Nefza haplotype and the Mad-

$P > 0.10$) sequences. Neither data set yielded a D significantly different from zero, so conformity of the data to neutral expectations could not be rejected.

Discussion

Control Region Flanking Genes and Consistency with a Mitochondrial Origin

The gene order around the control region in Fringilline finches is the same as that of the babbler (Edwards 1993b), chicken (Desjardins and Morais 1990), quail (Desjardins and Morais 1991), duck (Ramirez, Savoie, and Morais 1993), and snow goose (Quinn and Wilson 1993) mtDNA genomes, but different from that of mammals (Saccone, Pesole, and Sbisà 1991) and *Xenopus* (Roe et al. 1985). Furthermore, the flanking genes do not appear to be evolving at a reduced rate relative to other mitochondrial genes (the common chaffinch and chicken mitochondrial cytochrome *b*, ATPase 6, and NADH dehydrogenase subunit 5 genes are 80.0%, 77.8%, and 73.1% similar, respectively; unpublished data). In fact, we found similar levels of divergence between the greenfinch and blue chaffinch tRNA^{Glu} genes as between the chicken and goose tRNA^{Glu} genes (27%; Quinn and Wilson 1993), a level of divergence which may be explained by the hypothesis that the tRNA^{Glu} gene is under reduced functional constraint in birds because it is not adjacent to the sense transcript of any gene (Quinn and Wilson 1993). Additionally, the cloverleaf formation the finch tRNA^{Glu} is capable of assuming suggests that it is a functional gene. These results, in combination with the identification of recognizable features in the control region, are consistent with a mitochondrial rather than a nuclear origin for the control region sequences presented here.

Structural Evolution of the Control Region in Passerines

In most respects, the finch control region is very similar structurally to other avian control regions (Desjardins and Morais 1990, 1991; Ramirez, Savoie, and Morais 1993; Quinn and Wilson 1993; Wenink, Baker, and Tilanus 1993), although several minor exceptions were revealed. In particular, the putative TAS-element sequence of finches is quite divergent from the consensus sequence for vertebrates (Foran, Hixson, and Brown 1988), which was unexpected given that the primary sequence of the TAS is thought to be important to its function. However, the position and association of the putative TAS elements with secondary structure at the 5' end of the finch control region indicate their involvement in termination of D-loop synthesis. Another significant feature is the absence of CSB-2 and CSB-3 in finches. The CSBs are thought to act as processing signals for the generation of the RNA primers needed for replication (Walberg and Clayton 1981). However, only CSB-1 is clearly conserved among diverse vertebrates, suggesting that primary sequence alone does not prescribe function in the other CSBs. These CSBs are also absent in some mammals (Saccone, Pesole, and Sbisà 1991), although they occur in the duck and possibly in

the chicken (Ramirez, Savoie, and Morais 1993). Lastly, the bidirectional promoter sequence motif (TPu-TATATA) located downstream from CSB-1 in the chicken (Desjardins and Morais 1990), quail (Desjardins and Morais 1991), and duck (Ramirez, Savoie, and Morais 1993) is not present in the finches, although a cruciform-forming structure occurring in this region may act as a substitute. An alternate explanation is that finches have two promoters, in which case neither inverted repeats nor primary sequence homology with other organisms would likely exist.

Although variation in form of the 5'-TAS-associated secondary structure was evident among finch species, such variability has been identified in other vertebrates; in mammals and in *Xenopus* the structure is a cloverleaf (Brown et al. 1986), but in the chicken and goose it is a stem-and-loop (Quinn and Wilson 1993). Interestingly, this secondary structure is adjacent to the tract of C nucleotides in the control region of finches, while in the snow goose a similar tract of C's is found within the stem-and-loop (Quinn and Wilson 1993). This suggests a function for this primary sequence motif, possibly in termination of D-loop synthesis. Unlike the putative secondary structures of domain I, the CSB-1-associated secondary-structure sequence of domain III in finches is highly conserved. In other vertebrates this region forms a full cloverleaf, thought to be involved in initiation of D-loop synthesis (Brown et al. 1986; Quinn and Wilson 1993), to which the conserved partial cloverleaf found here may be analogous.

Thus, flanking gene order, base composition, primary sequence features, and potential secondary-structure formation all appear to be conserved among birds, and often among vertebrates. While the specific functions of the conserved sequence blocks are uncertain, their conservation among vertebrates suggests that they are crucial for the regulatory functions encoded by the D-loop, which spans the central domain in which they are located (Southern, Southern, and Dizon 1988). Conversely, high levels of variation in parts of the control region suggest that secondary structure is sometimes more important than primary sequence in the maintenance of control region functionality. Among *Fringilla* species, for example, primary sequence in domain I has diverged enough that secondary structure maintenance is independent of sequence similarity. Moreover, the exact nature of the secondary structure does not seem to be important to the function of the molecule.

Patterns and Rates of Sequence Evolution

As in other birds and vertebrates, the finch control region can be divided into three domains of variability, with most of the substitutions and indels occurring in the 5'- and 3'-flanking domains. The average number of substitutions per site is highest in the third domain, and the extent of rate variation among sites varies with domain. According to Yang and Kumar (1996), the α parameter indicates that the shape of the gamma distribution is highly skewed over the entire control region and in domains I and II ($\alpha < 1$); most sites therefore vary at low rates while a few change at high rates in

these regions. The larger estimate for domain III ($\alpha = 1.09$) indicates that rates in this region are approximately normally distributed among sites.

Brown et al. (1986) suggested that the frequency of indels relative to base substitutions in the control region decreases with increasing divergence. This trend is not apparent with the finch sequences, nor does indel frequency approach substitution frequency even within populations. Population studies of control region sequences from white sturgeons (Brown, Beckenbach, and Smith 1993), dunlins (Wenink, Baker, and Tilanus 1993), grey-crowned babblers (Edwards 1993a), and humans (Aquadro and Greenberg 1983) also reveal low proportions of indels relative to substitutions. While the mode of control region evolution may vary among lineages, it does not appear that insertions and deletions occur at extremely high frequency in the short term in many vertebrates.

Unusually low ts/tv ratios in the conserved central domain might reflect a mechanism to maintain its elevated G+T content (47%). This is consistent with the substitution biases observed in this domain, as A \leftrightarrow C and G \leftrightarrow T transversions would be unconstrained but transitions disfavored. A possible functional advantage of increased G+T content is that TnGm sequences may be efficient templates for DNA primase (Murray 1990). The substitution bias observed in domain I whereby purine transitions are favored at the expense of transversions involving A is also not due to compositional bias, as A is reduced and C is elevated in this region; again, it may reflect a functional constraint.

For comparative purposes, the chaffinch control region appears to be evolving only slightly faster than several coding mtDNA genes assessed for this group of birds. The rate varies markedly among domains; in particular, it is reduced in domain II and elevated in domain III. The domain I rate in finches is similar to the overall rate for the whole control region. By way of contrast, a hypervariable portion of domain I evolves at a very high rate (20.6%/Myr/lineage) in geese (Quinn 1992), but this rate only applies to a small hypervariable segment and clearly cannot be extrapolated to the whole domain. Similarly, sequence corresponding to domain I was estimated to evolve at four times the whole mtDNA genome rate in sturgeon (Brown, Beckenbach, and Smith 1993), although rates are thought to vary widely in fish, and the observed rate may be affected by A+T bias (Zhu et al. 1994).

In birds, relatively few studies have addressed differences in rate among domains, either because the species being compared are too divergent to allow it, or because only one region was examined. In finches, the third domain is more divergent in comparisons among species, yet the first domain exhibits higher nucleotide and haplotypic diversity and greater pairwise divergences among individuals. In a comparison of dunlins and turnstones Wenink, Baker, and Tilanus (1994) also found more variability to be associated with domain III, but within dunlins domain I was the most variable (unpublished data). Confusion regarding domain variability may result from the level of comparison being evalu-

ated; it may be that change accumulates earlier in domain I, but that more overall divergence is tolerated in domain III. This is consistent with the pattern of rate heterogeneity observed among domains in finches; domain I is thought to contain a few highly variable sites (mutational "hotspots") among many invariant sites, while domain III exhibits a normally distributed array of rates. Variability among domains is similarly unpredictable in mammals. For instance, domain III contains the most variation in mice and rats (Brown et al. 1986), whereas domain I is more variable in cetaceans (Hoelzel, Hancock, and Dover 1991). Because relative amounts of variation in different control region segments likely result from many factors, these results emphasize the need for further research on the dynamics of variability in diverse animals.

Utility for Population and Phylogenetic Studies

For phylogenetic studies among closely related species, as were looked at here, the control region appears to exhibit an appropriate level of variability. The tree recovered for the finches is consistent with the biogeography of the group, as the blue chaffinch and chaffinch occur sympatrically in the Canaries, and the blue chaffinch is thought to be the product of an earlier invasion from the same ancestral stock that later gave rise to common chaffinches (Stresemann 1927–1934). The third domain or the entire control region would be most suitable, as the first two domains alone do not contain sufficient information to give significant statistical support to nodes or branches of the tree.

Both domain I and domain III show promise for within-population studies. Sequences from both domains recovered similar relationships among chaffinch haplotypes, revealing a phylogeographic split between island and continental haplotypes and the existence of a divergent, ancestral haplotype from Nefza. Higher bootstrap support for the domain III tree results from greater divergence among more distant comparisons in domain III relative to domain I. However, bootstrap values within the major clade of continental haplotypes are not notably higher in the domain III tree versus the domain I tree, and both haplotypic and nucleotide diversity are generally higher in domain I sequences among these haplotypes. Finally, sequences from both domains showed patterns of divergence consistent with neutrality, as in other studies of avian control region sequences (e.g., Edwards 1993b). Thus, the choice of domain for a population study depends on the level of divergence among haplotypes, and in many studies it will be advisable to sequence both domains to increase the resolution in gene trees.

Acknowledgments

We are grateful to O. Haddrath and C. Ritland for laboratory assistance and discussion, O. Haddrath and C. Ayley for assistance with manuscript preparation, and M. Peck for field assistance. Additionally, the comments and suggestions of C. Moritz and two anonymous reviewers were very helpful. Funding for this research was

provided by the Natural Sciences and Engineering Research of Canada (grant A0200 to A.J.B. and a Postgraduate Scholarship to H.D.M.). This is contribution no. 40 from the Royal Ontario Museum Centre for Biodiversity and Conservation Biology.

LITERATURE CITED

- AQUADRO, C. F., and B. D. GREENBERG. 1983. Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. *Genetics* **103**:287–312.
- BAKER, A. J., M. D. DENNISON, A. LYNCH, and G. LE GRAND. 1990. Genetic divergence in peripherally isolated populations of chaffinches in the Atlantic islands. *Evolution* **44**: 981–999.
- BAKER, A. J., and H. D. MARSHALL. 1997. Mitochondrial control-region sequences as tools for understanding the evolution of avian taxa. Pp. 49–80 in D. P. MINDELL, ed. *Avian molecular systematics and evolution*. Academic Press, New York.
- BERG, T., T. MOUM, and S. JOHANSEN. 1995. Variable numbers of simple tandem repeats make birds of the order ciconiiformes heteroplasmic in their mitochondrial genomes. *Curr. Genet.* **27**:257–262.
- BROWN, G. G., G. GADALETA, G. PEPE, C. SACCONI, and E. SBISA. 1986. Structural conservation and variation in the D-loop-containing region of vertebrate mitochondrial DNA. *J. Mol. Biol.* **192**:503–511.
- BROWN, J. R., A. T. BECKENBACH, and M. J. SMITH. 1993. Intraspecific DNA sequence variation of the mitochondrial control region of white sturgeon. *Mol. Biol. Evol.* **10**:326–341.
- BUROKER, N. E., J. R. BROWN, T. A. GILBERT, P. J. O'HARA, A. T. BECKENBACH, W. K. THOMAS, and M. J. SMITH. 1990. Length heteroplasmy of sturgeon mitochondrial DNA: an illegitimate elongation model. *Genetics* **124**:157–163.
- CABOT, E. L., and A. T. BECKENBACH. 1989. Simultaneous editing of multiple nucleic acid and protein sequences with ESEE. *Comput. Appl. Biosci.* **5**:233–234.
- CLAYTON, D. A. 1982. Replication of animal mitochondrial DNA. *Cell* **28**:693–705.
- . 1984. Transcription of the mammalian mitochondrial genome. *Annu. Rev. Biochem.* **53**:573–594.
- . 1991. Replication and transcription of vertebrate mitochondrial DNA. *Annu. Rev. Cell Biol.* **7**:453–478.
- DESJARDINS, P., and R. MORAIS. 1990. Sequence and gene organization of the chicken mitochondrial genome. A novel gene order in higher vertebrates. *J. Mol. Biol.* **212**:599–634.
- . 1991. Nucleotide sequence and evolution of coding and noncoding regions of a quail mitochondrial genome. *J. Mol. Evol.* **32**:153–161.
- DODA, J. N., C. T. WRIGHT, and D. A. CLAYTON. 1981. Elongation of displacement-loop strands in human and mouse mitochondrial DNA is arrested near specific template sequences. *Proc. Natl. Acad. Sci. USA* **78**:6116–6120.
- EDWARDS, S. V. 1993a. Long-distance gene flow in a cooperative breeder suggested by genealogies of mitochondrial DNA sequences. *Proc. R. Soc. Lond. B Biol. Sci.* **252**:177–185.
- . 1993b. Mitochondrial gene genealogy and gene flow among island and mainland populations of a sedentary songbird, the grey-crowned babbler (*Pomatostomus temporalis*). *Evolution* **47**:1118–1137.
- FORAN, D. R., J. E. HIXSON, and W. E. BROWN. 1988. Comparison of ape and human sequences that regulate mitochondrial DNA transcription and D-loop synthesis. *Nucleic Acids Res.* **16**:5841–5861.
- HIGGINS, D. G., A. J. BLEASBY, and R. FUCHS. 1992. CLUSTAL V: improved software for multiple sequence alignment. *CABIOS* **8**:189–191.
- HOELZEL, A. R., J. M. HANCOCK, and G. A. DOVER. 1991. Evolution of the cetacean D-loop region. *Mol. Biol. Evol.* **8**:475–493.
- JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 in H. N. MUNRO, ed. *Mammalian protein evolution*. Academic Press, New York.
- KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**:111–120.
- KUMAR, S., K. TAMURA, and M. NEI. 1993. MEGA: molecular evolutionary genetic analysis. Version 1.0. The Pennsylvania State University, University Park, Pa.
- L'ABBÉ, D., J. F. DUBOIS, B. F. LANG, and R. MORAIS. 1991. The transcription of DNA in chicken mitochondria initiates from one major bidirectional promoter. *J. Biol. Chem.* **266**: 10844–10850.
- MURRAY, A. 1990. All's well that ends well. *Nature* **346**:797–798.
- NEI, M., and W.-H. LI. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **76**:5269–5273.
- QUINN, T. W. 1992. The genetic legacy of mother goose—phylogeographic patterns of lesser snow goose *Chen caerulescens caerulescens* maternal lineages. *Mol. Ecol.* **1**: 105–117.
- QUINN, T. W., and A. C. WILSON. 1993. Sequence evolution in and around the mitochondrial control region in birds. *J. Mol. Evol.* **37**:417–425.
- RAMIREZ, V., P. SAVOIE, and R. MORAIS. 1993. Molecular characterization and evolution of a duck mitochondrial genome. *J. Mol. Evol.* **37**:296–310.
- ROE, B. A., D.-P. MA, R. K. WILSON, and J. F.-H. WONG. 1985. The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome. *J. Biol. Chem.* **260**:9759–9774.
- SACCONI, C., M. ATTIMONELLI, and E. SBISA. 1987. Structural elements highly preserved during the evolution of the D-loop-containing region in vertebrate mitochondrial DNA. *J. Mol. Evol.* **26**:205–211.
- SACCONI, C., G. PESOLE, and E. SBISA. 1991. The main regulatory region of mammalian mitochondrial DNA: structure-function model and evolutionary pattern. *J. Mol. Evol.* **33**:83–91.
- SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**:406–425.
- SAMBROOK, J., E. F. FRITSCH, and T. MANIATIS. 1989. *Molecular cloning*. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
- SIMON, C. 1991. Molecular systematics at the species boundary: exploiting conserved and variable regions of the mitochondrial genome of animals via direct sequencing from amplified DNA. Pp. 33–71 in G. M. HEWITT, A. W. B. JOHNSTON, and J. P. W. YOUNG, eds. *Molecular techniques in taxonomy*. NATO ASI series. Vol. 57. Springer-Verlag, Berlin.
- SOUTHERN, Š., P. J. SOUTHERN, and A. E. DIZON. 1988. Molecular characterization of a cloned dolphin mitochondrial genome. *J. Mol. Evol.* **28**:32–42.
- STRESEMANN, E. 1927–1934. *Aves*. Pp. 729–853 in G. KÜNTHAL, ed. *Handbuch der Zoologie VII*. Walter Gruyer, Berlin.

- TAJIMA, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- TAJIMA, F., and M. NEI. 1984. Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.* **1**: 269–285.
- TAYLOR, M. F. J., S. W. MCKECHNIE, N. PIERCE, and M. KREITMAN. 1993. The Lepidopteran mitochondrial control region: structure and evolution. *Mol. Biol. Evol.* **10**:1259–1272.
- WALBERG, M. W., and D. A. CLAYTON. 1981. Sequences and properties of the human KB cell and mouse L cell D-loop regions of mitochondrial DNA. *Nucleic Acids Res.* **9**:5411–5421.
- WENINK, P. W., A. J. BAKER, H.-U. RÖSNER, and M. G. J. TILANUS. 1996. Global mitochondrial DNA phylogeography of holarctic breeding dunlins (*Calidris alpina*). *Evolution* **50**:318–330.
- WENINK, P. W., A. J. BAKER, and M. G. J. TILANUS. 1993. Hypervariable-control-region sequences reveal global population structuring in a long-distance migrant shorebird, the dunlin (*Calidris alpina*). *Proc. Natl. Acad. Sci. USA.* **90**: 94–98.
- . 1994. Mitochondrial control-region sequences in two shorebird species, the turnstone and the dunlin, and their utility in population genetic studies. *Mol. Biol. Evol.* **11**: 22–31.
- YANG, Z. 1995. Phylogenetic analysis by maximum likelihood (PAML). Version 1.1. Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park, Pa.
- YANG, Z., and S. KUMAR. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Mol. Biol. Evol.* **13**: 650–659.
- ZHU, D., B. G. M. JAMIESON, A. HUGALL, and C. MORITZ. 1994. Sequence evolution and phylogenetic signal in control-region and cytochrome b sequences of rainbow fishes (Melanotaeniidae). *Mol. Biol. Evol.* **11**:672–683.
- ZUKER, M. 1989. Computer prediction of RNA structure. *Methods Enzymol.* **180**:262–288.

CRAIG MORITZ, reviewing editor

Accepted November 4, 1996