MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) doi: 10.1111/j.1755-0998.2011.03049.x

Searching for evidence of selection in avian DNA barcodes

KEVIN C. R. KERR
Division of Birds, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, USA

Abstract

The barcode of life project has assembled a tremendous number of mitochondrial cytochrome *c* oxidase I (COI) sequences. Although these sequences were gathered to develop a DNA-based system for species identification, it has been suggested that further biological inferences may also be derived from this wealth of data. Recurrent selective sweeps have been invoked as an evolutionary mechanism to explain limited intraspecific COI diversity, particularly in birds, but this hypothesis has not been formally tested. In this study, I collated COI sequences from previous barcoding studies on birds and tested them for evidence of selection. Using this expanded data set, I re-examined the relationships between intraspecific diversity and interspecific divergence and sampling effort, respectively. I employed the McDonald-Kreitman test to test for neutrality in sequence evolution between closely related pairs of species. Because amino acid sequences were generally constrained between closely related pairs, I also included broader intra-order comparisons to quantify patterns of protein variation in avian COI sequences. Lastly, using 22 published whole mitochondrial genomes, I compared the evolutionary rate of COI against the other 12 protein-coding mitochondrial genes to assess intragenomic variability. I found no conclusive evidence of selective sweeps. Most evidence pointed to an overall trend of strong purifying selection and functional constraint. The COI protein did vary across the class Aves, but to a very limited extent. COI was the least variable gene in the mitochondrial genome, suggesting that other genes might be more informative for probing factors constraining mitochondrial variation within species.

Keywords: birds, cytochrome *c* oxidase, DNA barcode, mitochondrial DNA, purifying selection, selective sweep

Received 2 March 2011; revision received 16 June 2011; accepted 20 June 2011

Introduction

The role of selection in the evolution of the mitochondrial genome is the subject of ongoing debate (Gerber *et al.* 2001; Ballard & Whitlock 2004; Meiklejohn *et al.* 2007). Variation at mitochondrial genes was long regarded as largely neutral and has been frequently used to infer effective population size and historical demographies based on that assumption. Consequently, mitochondrial DNA (mtDNA) variation has become a mainstay of molecular ecology and phylogeographic studies (Ballard & Whitlock 2004). More recently, this paradigm has shifted, as an ever-increasing role of selection has been recognized in the evolution of mitochondrial genes (Gerber *et al.* 2001).

The earliest studies to test the expectations of neutrally evolving mtDNA instead found evidence of selection against mildly deleterious mutations in varied groups of animals including *Drosophila* (Rand & Kann 1996), mice (Nachman *et al.* 1994), humans (Hasegawa *et al.* 1998; Wise *et al.* 1998) and birds (Fry 1999). In such cases, the observed trend was towards an excess of amino acid polymorphisms within species as compared to amino acid substitutions between species. In contrast, more recent studies have cited evidence of positive selection in the mitochondrial genome, which has been attributed to cyto-nuclear interactions (Ballard & Whitlock 2004). Looking across 26 mammalian taxa, Schmidt *et al.* (2001) found that the nonsynonymous substitution rate was much greater in gene regions that coded for close-contact residues (i.e. those interacting with nuclear-encoded residues), suggesting positive selection acting at those sites.

The aforementioned examples of cyto-nuclear interactions also illustrate the mitochondrial genome's susceptibility to indirect selection. The mitochondrial genome generally lacks recombination and thus behaves as a single linkage group. In some cases, additional genes may also be linked, as in birds where females are the heterogametic sex and linkage extends to the W chromosome (Berlin *et al.* 2007). A new view is emerging that mitochondrial evolution could be governed largely by

Correspondence: Kevin C. R. Kerr, Fax: (202) 633 8084;
E-mail: kerrkc@si.edu

recurrent 'selective sweeps'. A selective sweep, also known as 'genetic hitchhiking', occurs when selection acting on one site results in loss of variation from linked sites (Hedrick 1980). Unfortunately, the selective sweep hypothesis is difficult to test in the mitochondrial genome *because* it forms a single linkage group; most tests depend on the comparison of multiple loci (Galtier *et al.* 2000).

Demonstration of selective sweeps occurring in mitochondrial genes has typically been indirect. For example, in a landmark study that examined nearly 3000 animal species, Bazin *et al.* (2006) concluded that the genetic diversity of mitochondrial markers was independent of population size, contradicting prior assumptions. They contended that purifying selection could not explain the observed pattern and that recurrent fixation of beneficial mutations was the most parsimonious explanation. The study generated much debate (Mulligan *et al.* 2006; Meiklejohn *et al.* 2007) and was perhaps most soundly criticized for assessing neutrality between distantly related taxa (Wares *et al.* 2006). The supposed sweeps were actually detected at deep phylogenetic levels, not necessarily between closely related species (Berry 2006).

Studies involving large-scale surveys of the mitochondrial gene cytochrome *c* oxidase I (COI) for DNA barcoding have proposed that routine selective sweeps could explain the observation of consistently low variation within species, despite the varying age of species (e.g. Kerr *et al.* 2007). This observation violates a rudimentary expectation of neutral theory that levels of intraspecific polymorphism are correlated with interspecific divergence (Langley *et al.* 1993). Simulations based on neutral models predicted that more than 4 million generations would be necessary to achieve the degree of COI differentiation observed between closely allied species (Hickerson *et al.* 2006), but barcode data suggested independence of intraspecific variation and species age (Kerr *et al.* 2007, 2009b). Baker *et al.* (2009) offered an alternative explanation, arguing that low intraspecific diversity is an artefact of the small number of individuals examined in most barcoding studies and that denser intraspecific sampling would erase this pattern. This debate is further complicated by varying methods that can be used to measure intraspecific diversity (e.g. genetic distance, haplotype number, etc.).

The number of available COI sequences has increased dramatically with the success of DNA barcoding, particularly for taxa such as birds (Frezal & Leblois 2008). In this study, I take advantage of the expanded avian COI barcode data to more rigorously test for evidence of selection. This includes a reassessment of the relationship between intraspecific variation and interspecific divergence and sampling effort, tests for neutrality and a cross-genome comparison of genetic variation.

Methods

Data collection

Published data were accessed from three public projects in the Barcode of Life Database (BOLD, Ratnasingham & Hebert 2007): 'Birds of North America—Phase II' (Kerr *et al.* 2007), 'Birds of Argentina—Phase I' (Kerr *et al.* 2009b) and 'Birds of the eastern Palearctic' (Kerr *et al.* 2009a). Public data were supplemented with previously unpublished DNA barcode sequences from 826 specimens representing 113 species of North American birds. These sequences are available from BOLD ('Birds of North America, additional sequences') and GenBank (accession numbers HM033200–HM034025). The majority of specimens (98%) were represented by feather samples collected from banding stations including several across Canada (Atlantic Bird Observatory and Brier Island Bird Migration Research Station, Nova Scotia; St. Andrews Banding Station, New Brunswick; Gros Morne National Park Migration Monitoring Station, Newfoundland; McGill Bird Observatory, Québec; Haldimand Bird Observatory, Long Point Bird Observatory, Prince Edward Point Bird Observatory, and Tommy Thompson Park Bird Research Station, Ontario; Inglewood Bird Sanctuary, Alberta; Mackenzie Nature Observatory, Rocky Point Bird Observatory, and Vaseux Lake Bird Observatory, British Columbia; Albert Creek Bird Banding Station and Teslin Lake Bird Banding Station, Yukon) and a single station in North Carolina, U.S.A. (Appalachian Highlands Science Learning Center of the US National Park Service). The remaining specimens were comprised of muscle tissue samples from curated collections (including the Canadian Wildlife Service, Royal Ontario Museum, Burke Museum of Natural History and Culture, Museum of Comparative Zoology, Museum of Southwestern Biology and the Smithsonian Institution National Museum of Natural History). All DNA extraction, PCR and sequencing methods follow those reported by Kerr *et al.* (2007) and were performed at the Biodiversity Institute of Ontario, University of Guelph.

Assessing genetic diversity

To focus on species that were sampled most comprehensively, only those that were represented in BOLD by 12 or more COI sequences were included in this analysis. Additionally, only sequences greater than 650 bp and with fewer than 7 (= approximately 1%) ambiguous base calls were included. In total, 55 species were used in the analysis and are listed in supplementary Table S1 (Supporting information). By nature of the original sampling scheme, specimens were sampled broadly from across the respective range of each species. To quantify genetic

diversity, the number of unique haplotypes (h), haplotype diversity (H) and nucleotide diversity (π) were calculated using DnaSP version 5.0 (Librado & Rozas 2009). The minimum nearest-neighbour distance was used to measure interspecific divergence and was calculated using the Kimura 2-parameter metric in the BOLD Management and Analysis System version 2.5 (Ratnasingham & Hebert 2007).

Linear regression was used to test the relationship between sampling effort (i.e. the number of specimens included in the analysis) and h, H and π, respectively, using R version 2.5.0 (R Development Core Team 2007). The Pearson product–moment correlation coefficient was used to test for a relationship between nearest-neighbour distance and h, H and π, respectively, also using R version 2.5.0 (R Development Core Team 2007).

Neutrality tests of COI variation

To test COI variation for evidence of neutrality, I sought out well-represented pairs of sister species from the BOLD database. Congeneric pairs of sister taxa were identified using a neighbour-joining tree generated from BOLD (Ratnasingham & Hebert 2007). Species pairs were selected when one member of the pair was represented by at least seven specimens (to capture intraspecific variation) and the other was represented by at least two specimens (to avoid the risk of comparison to an aberrant sequence). In total, this included 34 pairs of species

Table 1 Thirty-four species pairs included in McDonald-Kreitman tests for neutrality of COI variation in birds. Sample size for each species is indicated in parentheses. Acronyms are for nonsynonymous intraspecific polymorphisms (P_n), synonymous intraspecific polymorphisms (P_s), nonsynonymous interspecific fixed differences (D_n), synonymous interspecific fixed differences (D_s) and neutrality index (NI). P_s, D_s, P_n and D_n are uncorrected values. The P-values from Fisher's exact test are reported

Species 1 (n)	Species 2 (n)	P_n	P_s	D_n	D_s	NI	P
Lagopus muta (21)	*L. leucura* (5)	0	6	2	38	0	1.000
Phalaropus lobatus (11)	*P. fulicarius* (2)	1	2	0	34	–	0.081
Actitis macularius (8)	*A. hypoleucos* (5)	0	0	0	64	–	–
Brachyramphus brevirostris (7)	*B. marmoratus* (2)	0	7	0	31	–	–
*Gallinago gallinago** (9)	*G. paraguaiae* (3)	1	4	0	21	–	0.192
Larus ridibundus (8)	*L. philadelphia* (4)	0	1	0	15	–	–
Stercorarius longicaudus (7)	*S. parasiticus* (4)	0	5	0	37	–	–
Thalasseus sandvicensis (8)	*T. elegans* (5)	0	6	0	10	–	–
Phalacrocorax pelagicus (11)	*P. penicillatus* (5)	0	3	0	37	–	–
Puffinus pacificus (7)	*P. bulleri* (6)	0	4	0	22	–	–
Strix occidentalis (7)	*S. varia* (4)	1	3	1	51	17	0.139
Megascops asio (9)	*M. kennicottii* (5)	0	4	1	37	0	1.000
Chaetura vauxi (8)	*C. pelagica* (2)	0	3	0	14	–	–
Falco sparverius (7)	*F. tinnunculus* (3)	0	8	1	53	0	1.000
Picoides villosus (10)	*P. albolarvatus* (7)	0	14	0	18	–	–
Zenaida macroura (8)	*Z. auriculata* (8)	0	8	0	15	–	–
Columbina passerina (8)	*C. talpacoti* (4)	0	3	0	35	–	–
Empidonax alnorum (8)	*E. traillii* (4)	2	3	2	13	4.3	0.249
Leptasthenura aegithaloides (8)	*L. fuliginiceps* (3)	0	5	1	37	0	1.000
Phacellodomus ruber (8)	*P. striaticollis* (3)	0	2	0	21	–	–
Phytotoma rutila (7)	*P. rara* (3)	2	3	0	59	–	0.005
Nucifraga caryocatactes (9)	*N. columbiana* (8)	0	10	0	35	–	–
Poecile montana (20)	*P. palustris* (11)	1	12	0	38	–	0.255
Cinclus mexicanus (7)	*C. cinclus* (5)	0	7	2	51	0	1.000
Locustella certhiola (7)	*L. ochotensis* (7)	0	10	0	30	–	–
Turdus viscivorus (8)	*T. philomelos* (4)	0	14	0	54	–	–
Seiurus noveboracensis (24)	*S. aurocapilla* (14)	0	21	0	47	–	–
Sicalis luteola (7)	*S. flaveola* (6)	3	7	0	55	–	0.003
Melospiza melodia (27)	*M. lincolnii* (26)	2	7	0	18	–	0.103
Emberiza aureola (11)	*E. rustica* (9)	1	11	0	34	–	0.261
Paroaria capitata (7)	*P. coronata* (5)	0	1	1	31	0	1.000
Molothrus bonariensis (10)	*M. ater* (9)	3	4	0	16	–	0.020
Fringilla montifringilla (12)	*F. coelebs* (4)	0	7	0	49	–	–
Passer domesticus (17)	*P. montanus* (11)	1	11	0	37	–	0.245

*Includes *Gallinago delicata*.

representing 10 orders and 29 families of birds (see Table 1). The standard McDonald-Kreitman test (available at http://bioinf3.uab.cat/mkt/mkt.asp) was run on each species pair using the vertebrate mitochondrial code (Egea *et al.* 2008). This test produces a 2 × 2 contingency table of nonsynonymous intraspecific polymorphisms (P_n), synonymous intraspecific polymorphisms (P_s), nonsynonymous fixed differences (D_n) and synonymous fixed differences (D_s). The program also provides the neutrality index, or NI, which is calculated as $(P_n/P_s)/(D_n/D_s)$ (Rand & Kann 1996). Neutrality is supported when NI = 1, whereas $N < 1$ implies an excess of amino acid divergence between species and $N > 1$ implies an excess of amino acid polymorphism within species. The program employs a chi-square test to assess significance, but I instead used a Fisher's exact test using R version 2.5.0 (R Development Core Team 2007) because of the low number observed of nonsynonymous mutations.

Amino acid variation

Because amino acid variation tends to be low between pairs of avian sister species, I also examined variation at the ordinal level to assess amino acid variation in COI. I selected the 12 best-represented orders from the database (Apodiformes, Anseriformes, Charadriiformes, Ciconiiformes, Columbiformes, Coraciiformes, Falconiformes, Galliformes, Passeriformes, Piciformes, Psittaciformes and Strigiformes) and then trimmed the database to include only species with at least two full-length barcodes (i.e. 694 bp). Species with polymorphisms (n = 43) were removed from the analysis, so only species with fixed differences were included. As the remaining sequences were identical within species, a single sequence was selected randomly from each species to populate the final working data set (n = 623). Nucleotide sequences were translated to amino acid sequences using Geneious version 3.5 (Drummond *et al.* 2007). The number of amino acid sequence 'types' was tallied for each order (denoted as h', analogous to haplotype number). To capture the diversity of amino acid sequence types within each order (i.e. the frequency of each type), I calculated a diversity value (denoted H') using a modified version of Nei's haplotype diversity equation (equation 8.5, Nei 1987),

$$H' = n(1 - \sum x_i^2)/(n - 1)$$

where n equals the number of species sampled in each order and x_i represents the frequency of the ith amino acid sequence type within each order. The number of amino acid types unique to each order was identified using a neighbour-joining tree generated through BOLD (Ratnasingham & Hebert 2007). To assess the level of intra-order amino acid divergence, I calculated the mean PAM1 matrix scores for each order using MEGA version 4.0 (Tamura *et al.* 2007).

To approximate the position of amino acid substitutions within the COI protein, a consensus sequence was constructed from the 623 amino acid sequences described earlier. The consensus sequence was aligned to the bovine sequence using an ends-free local alignment and the BLOSUM62 substitution matrix. The consensus sequence was positioned on the bovine secondary structure, which has been determined via crystallography (Tsukihara *et al.* 1996), using the alignment as a guide. Residues were identified as either loop or helix sites based on their location in the bovine structure. A chi-square test with Yates correction was run in R version 2.5.0 (R Development Core Team 2007) to see whether variable sites were equally distributed between the two regions.

COI vs. other mitochondrial genes

Whole avian mitochondrial genomes published on GenBank as of 14 December 2009 were surveyed for pairs of congeneric taxa. Two genera—*Gallus* and *Syrmaticus*—were represented by more than two species, so the two most closely related species were selected for analysis. In total, 22 complete mitochondrial genomes were downloaded for the 11 available congeneric pairs (see supplementary Table S2, Supporting information). Each of the 13 protein-coding genes was segregated into a separate fasta file and aligned using ClustalW in MEGA version 4.0 (Tamura *et al.* 2007). An extra base pair was removed from the ND3 sequence in several species to maintain reading frame (Mindell *et al.* 1998). ND6 was analysed in reverse complement for all species. Pairwise d_N/d_S ratios were calculated for all genes from all congeneric pairs using the codeml package in PAML version 4.3 (Yang 2007).

The d_N/d_S ratios were transformed prior to statistical analysis using an arcsine square root transformation. A one-way ANOVA was used to test for a difference in d_N, d_S and the d_N/d_S ratios between the different genes. A Tukey's honestly significant difference (HSD) test was used subsequently to identify which genes differed significantly. Both tests were performed using R version 2.5.0 (R Development Core Team 2007).

Results

Genetic diversity

The sampling effort ranged from 12 to 34 COI sequences per species. The nearest-neighbour distance varied dramatically from 0% to 12.88% K2P corrected distance.

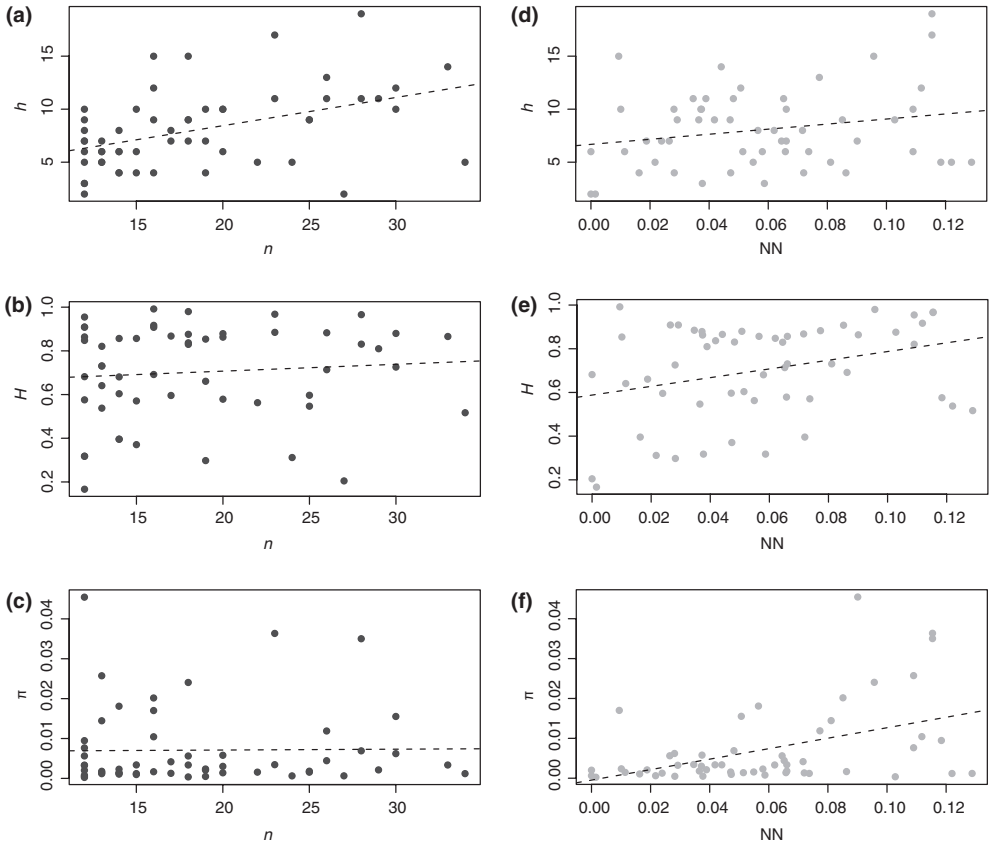

Fig. 1 Scatter plots relating sampling effort of avian COI sequences (*n*) to (a) haplotype number (h), (b) haplotype diversity (H) and c) nucleotide diversity (π), and K2P nearest-neighbour distance (NN) to d) haplotype number (h), (e) haplotype diversity (H) and (f) nucleotide diversity (π).

Figure 1 provides scatter plots for all six comparisons. Only haplotype number was significantly correlated with sampling effort, although the relationship was weak ($R^2 = 0.20$, $F_{1,\ 53} = 13.16$, $P < 0.001$). There was no significant relationship with haplotype diversity ($R^2 = 0.09$, $F_{1,\ 53} = 0.42$, $P = 0.518$) or nucleotide diversity ($R^2 = 0.01$, $F_{1,\ 53} = 0.01$, $P = 0.923$). However, there was a weak but significant correlation between nearest-neighbour distance and nucleotide diversity ($r_{53} = 0.46$, $P < 0.001$), as well as haplotype diversity ($r_{53} = 0.31$, $P = 0.019$), but there was no significant correlation to haplotype number ($r_{53} = 0.22$, $P = 0.103$).

Neutrality tests

Both polymorphic and fixed nonsynonymous differences were rare between species pairs (see Table 1). Consequently, 15% of NI values were zero and 79% were undefined. Only two pairwise comparisons (*Empidonax alnorum/E. traillii* and *Strix occidentalis/S. varia*) possessed at least one polymorphic and one fixed nonsynonymous difference, but neither pair had significantly different ratios of polymorphic to fixed differences (see Table 1). Overall, D_S was significantly greater than P_S (7.5-fold on average; $t_{39} = -10.49$, $P < 0.001$), whereas D_N did not differ significantly from P_N ($t_{59} = 1.09$, $P = 0.279$). P_S was not correlated with D_S ($r_{32} = 0.14$, $P = 0.404$) but was correlated with sample size ($r_{32} = 0.49$, $P < 0.01$).

Amino acid diversity

Table 2 summarizes the results of the diversity tests. Of the 231 residues, 41 were variable among the species examined. Amino acid sequence diversity was relatively high within orders, but the degree of variation (i.e. the number of substitutions) was limited. Highly divergent taxa often shared the same sequence. For example, the same amino acid sequence was recovered from members of the Charadriiformes, Columbiformes, Coraciiformes and Falconiformes.

The predicted secondary structure of the consensus amino acid sequence is illustrated in Fig. 2. According to the predicted structure, 78% of residues occur in helix

Table 2 Summary of COI amino acid variation for the 12 orders of birds examined. The number of amino acid sequence types (*h*′), the diversity of amino acid sequence types (*H*′) and the percentage of types unique to each order are outlined below. The mean intra-order PAM score indicates the level of amino acid divergence

Order	*n*	*h*′	*H*′	Uniqueness (%)	PAM (±SD)
Ciconiiformes	9	6	0.833	50	0.00778 (±0.00355)
Anseriformes	30	9	0.662	100	0.00507 (±0.00215)
Falconiformes	15	9	0.848	93	0.01199 (±0.00443)
Galliformes	11	8	0.891	100	0.01113 (±0.00488)
Charadriiformes	84	10	0.361	50	0.00196 (±0.00076)
Columbiformes	14	8	0.901	50	0.00689 (±0.00301)
Psittaciformes	9	7	0.944	71	0.00817 (±0.00331)
Strigiformes	9	8	0.972	100	0.01817 (±0.00562)
Apodiformes	15	9	0.924	89	0.01084 (±0.00341)
Coraciiformes	7	6	0.952	71	0.01724 (±0.00526)
Piciformes	22	4	0.403	75	0.00194 (±0.00112)
Passeriformes	398	78	0.903	95	0.01458 (±0.00490)
Total	623	148	0.944		

sites and 73% of variable positions were in that region, revealing no association between variation and position in the secondary structure ($\chi^2 = 0.46$, $P = 0.497$). Most amino acid substitutions were low-impact changes known to commonly occur (e.g. isoleucine ↔ valine, isoleucine ↔ leucine and alanine ↔ serine) (Betts & Russell 2003). Approximately 20% of the amino acid substitutions were only observed within a single species. Confidence in the accuracy of these amino acid sequences is increased by the sampling strategy, wherein only fixed amino acid substitutions were included in analysis. However, two unusual (i.e. rare) substitutions were observed, both occurring in the last residue and both in Anseriformes: leucine → serine in *Netta peposaca* and leucine → phenylalanine in *Amazonetta brasiliensis*.

Genomic comparisons

Syrmaticus ellioti and *S. humiae* were too narrowly divergent to be informative (the d_N/d_S ratio for every gene except ND5 was either 0 or 1), so this species pair was removed from further analyses. The mean values for d_N, d_S and the d_N/d_S ratio are depicted for each gene in Fig. 3. There was a significant difference in the d_N/d_S ratios recorded for the thirteen protein-coding genes ($F_{12, 117} = 6.40$, $P < 0.001$). A significant difference was also recorded for d_N ($F_{12, 117} = 3.12$, $P < 0.001$), though not for d_S ($F_{12, 117} = 0.75$, $P < 0.699$). Post hoc Tukey's HSD comparisons revealed that the ANOVA result for the d_N/d_S ratios was attributed to ATP8, which differed significantly from all other genes ($P < 0.01$), and COI, which differed from ND3 ($P < 0.05$). Similarly, the difference in d_N occurred between ATP8 and COI, COII, COIII and Cyt B, respectively ($P < 0.01$).

Discussion

The present results provide some confirmation to the prediction of Baker *et al.* (2009) that the number of rare haplotypes encountered will increase with sampling effort. However, because intraspecific divergence remains relatively low between most haplotypes, the mean intraspecific variation is nearly unaffected by sampling effort. This has also been demonstrated in human populations, where it was found that haplotype number might not reach saturation until sample size reaches 1000 individuals, yet neither haplotype diversity nor nucleotide diversity increases because of the limited divergence of those rarer haplotypes (Pereira *et al.* 2004).

There was a weak relationship between haplotype diversity and interspecific divergence and, in contrast to Kerr *et al.* (2007), also a weak relationship between intraspecific variation and interspecific divergence as well. This discrepancy may be partly due to their treatment of divergent mitochondrial lineages as 'provisional species', which would reduce both intraspecific variation and interspecific divergence in some of the species included in this study such as *Vireo gilvus* or *Troglodytes troglodytes*, among others. This difference aside the relationship is suggestive of neutral processes, rather than selection. However, it is important to note that, despite statistical significance, neither sample size nor nearest-neighbour distance explained much of the data.

Attempts to test neutrality were impeded by the lack of amino acid sequence variation both within and between species, which is a common problem of this method (Meiklejohn *et al.* 2007). The McDonald-Kreitman test is susceptible to error when taxa are distantly related and multiple substitutions at single sites are

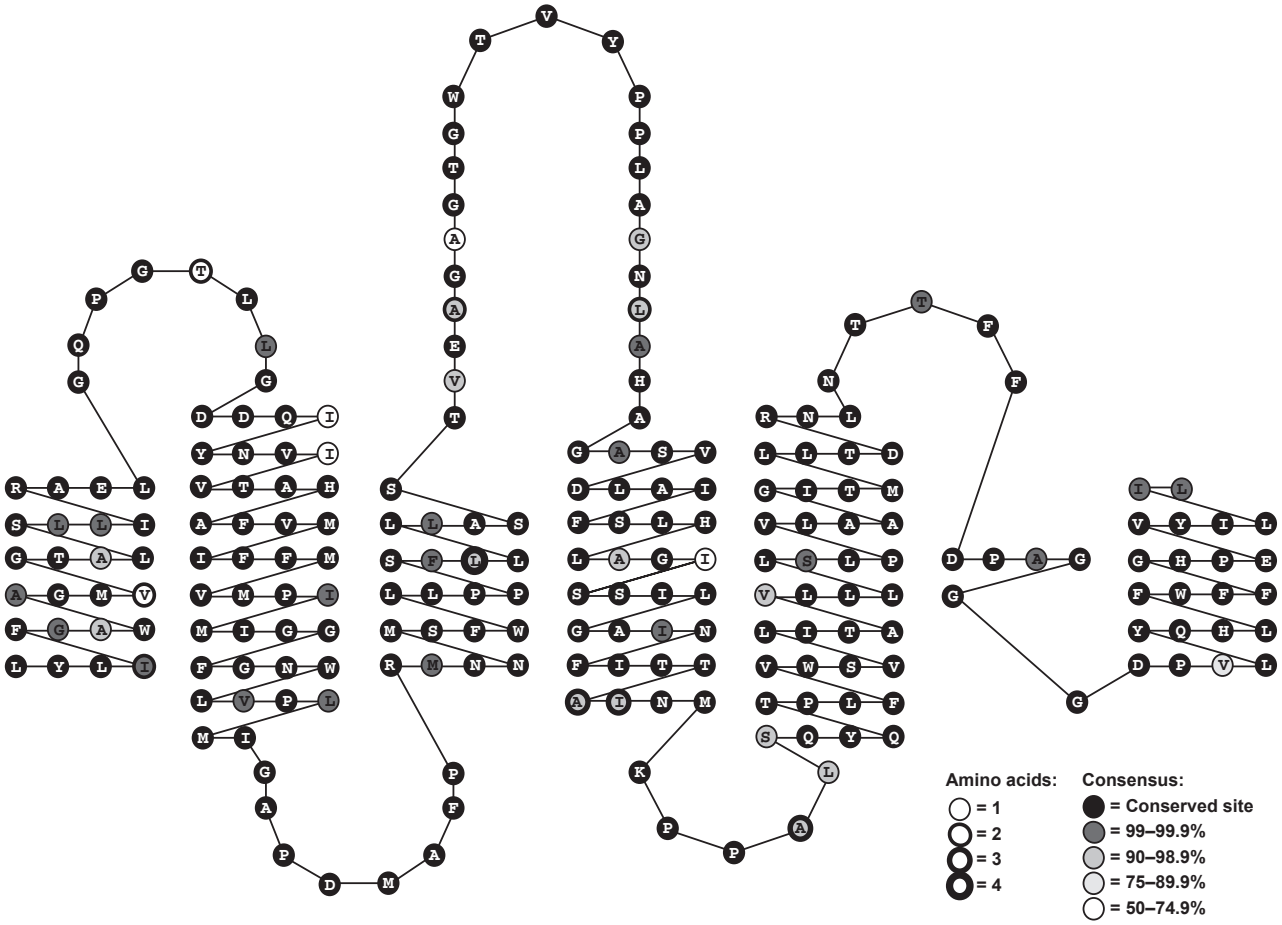

Fig. 2 Predicted secondary structure for the avian consensus sequence of the 'barcoding region' of COI based on the structure derived from bovine cytochrome *c* oxidase. Letters indicate the consensus amino acid sequence based on the one-letter code. Black circles are conserved sites. Variable sites vary from white to grey based on the percentage of sequences containing the consensus amino acid. The number of different amino acids occurring at a single site is represented by the thickness of the outline.

likely, but amino acid variation is rarer when taxa are closely related (Ballard & Whitlock 2004). Lack of amino acid variation disrupts the utility of the neutrality index, as it results in either a zero value (when polymorphisms are absent) or infinity (when divergence is absent). Other studies have circumvented this issue with a simple, yet questionable practice of substituting zeros with arbitrary values (i.e. Rand & Kann 1998; Bazin *et al.* 2006). Limited count values may also impair the power of significance tests.

Previous studies examining the neutrality of variation in avian mitochondrial genes have generally proposed a model of mildly deleterious mutations (e.g. Fry 1999). Zink (2005) found that members of the passerine genus *Parus* exhibited an excess of nonsynonymous polymorphisms in closely related species and cited purifying selection as the cause after ruling out demographic effects. However, Zink *et al.* (2006) could not reject neutrality when examining phylogroups of the polymorphic species *Sitta europaea*, suggesting that drift was largely responsible for genetic differences between budding species. The taxonomic scope in this study was much broader than its predecessors and the general pattern appeared to be one of functional constraint. While the only calculable NI values were greater than one, it would be misleading to describe the sequences as bearing excess amino acid polymorphisms because amino acid variation was generally rare. In either case, the pattern is suggestive of purifying selection.

Variation in the amino acid sequence was rare between closely related species, but there was substantial variation when broader taxonomic comparisons were made. Most positions in the amino acid sequence (82%) are conserved across all birds. Variable positions were proportionately equal between helix and loop sites. This is inconsistent with previous studies that have observed differing selective pressures on surface (primarily loops) and transmembrane sites, with substitutions in

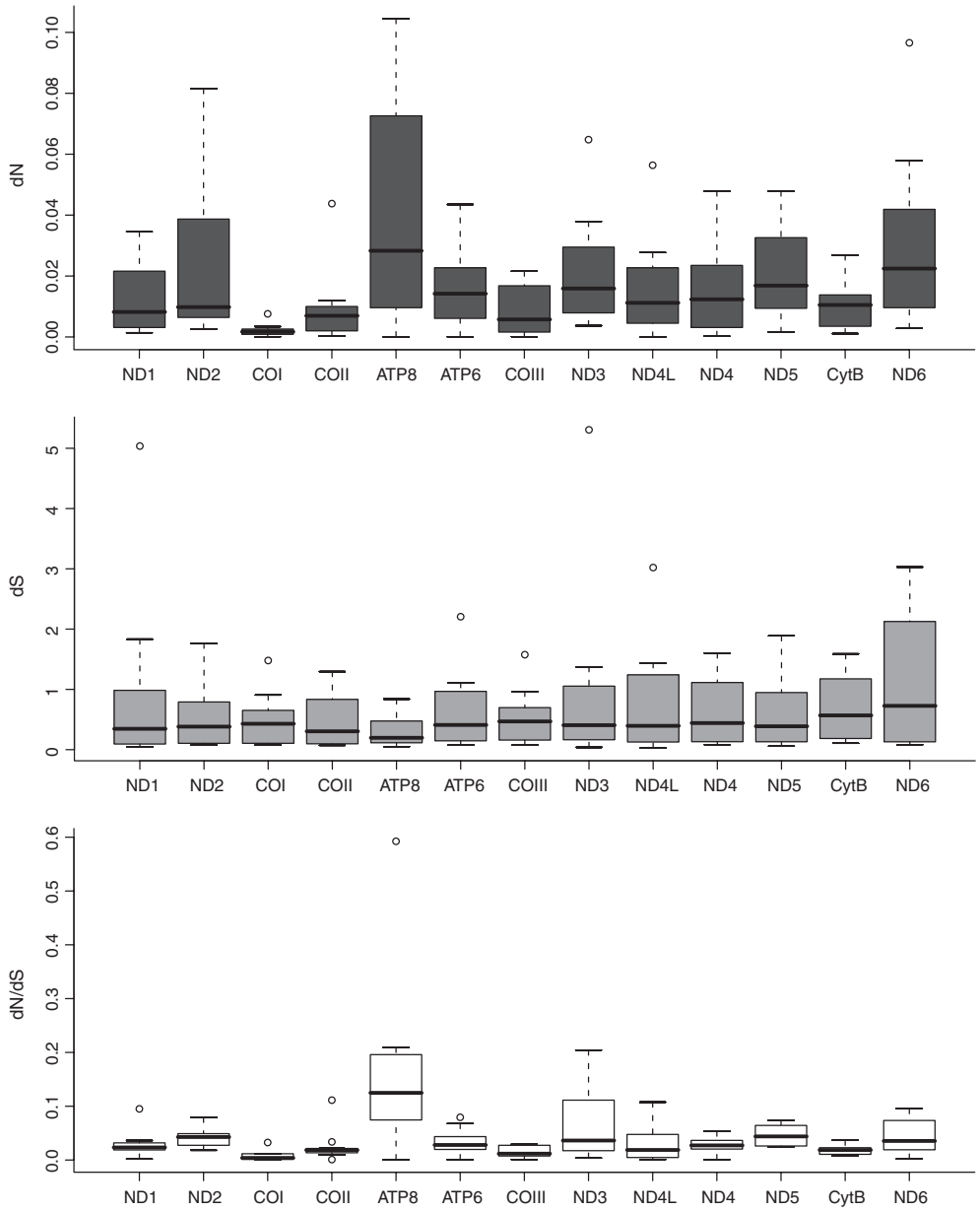

Fig. 3 Box plots of d_N (dark grey), d_S (light grey) and the d_N/d_S ratios (white) for each of the thirteen protein-coding mitochondrial genes from the 22 species listed in supplementary Table S2 (Supporting information). Genes are ordered to match the arrangement in the avian mitochondrial genome.

transmembrane sites being more heavily constrained by interaction effects with other residues (Wang & Pollock 2007) and a greater tendency towards neutral evolution within surface sites (Wise *et al.* 1998). However, these inferences suppose the accuracy of the predicted secondary structure, which unfortunately is difficult to verify.

Mapping these changes onto a phylogeny is challenging as our current understanding of evolutionary relationships between the major avian orders is in flux (Hackett *et al.* 2008). However, looking at variability within orders, it is clear that some amino acids substitutions are recurrent (e.g. isoleucine ↔ valine at position 12), whereas other substitutions have a single origin (e.g. glycine → serine at position 117 in Strigiformes). It is possible that mutations between amino acids such as valine, leucine and isoleucine escape the effects of selection because of their chemical similarity. In other cases, small changes to proteins can have an adaptive impact. For example, adaptive changes to haemoglobin proteins in high-altitude geese have been attributed to four

mutations in the bar-headed goose, *Anser indicus* (Liang *et al.* 2001), and a single mutation in the Andean goose, *Chloephaga melanoptera* (Hiebl *et al.* 1987). For the majority of the amino acid substitutions observed here, an adaptive explanation seems unlikely, especially when amino acid sequences are shared between very divergent taxa. Co-evolution within the gene has been demonstrated for proximal residues in vertebrates (Wang & Pollock 2007), but this too seems an unlikely explanation given the independent origins of the sequences. It is more likely that most of the observed amino acid substitutions have low-impact changes that escape purifying selection, particularly because linkage between genetic loci is known to reduce the effectiveness of purifying selection (Paland & Lynch 2006).

Selection in mitochondrial genes has been attributed to co-evolution with nuclear-encoded genes. In marine copepods, interpopulation hybrids have shown reduced mitochondrial function and, subsequently, reduced fitness (Burton *et al.* 2006). Reduced function has also been demonstrated in cybrid cells that cross human nuclear DNA with mtDNA from other primates (Kenyon & Moraes 1997). In birds, the fitness costs to hybrids are less clear. Empirical studies have revealed an effect on the metabolic rate of hybrids between divergent populations of an Old World passerine, *Saxicola torquata* spp. (Tieleman *et al.* 2009). Despite a measurable impact on metabolic rate, the overall fitness cost is uncertain and could still depend on the magnitude of mitochondrial mismatch between taxa. Complete mitochondrial introgression has been demonstrated between certain avian sister species, such as the Palearctic buntings *Emberiza citrinella* and *E. leucocephalos* (Irwin *et al.* 2009), which is one situation that does lend support to the selective sweep hypothesis. Introgression has also been ascribed to selective processes in other nonavian species, including salmonids (Wilson & Bernatchez 1998) and *Drosophila* (Bachtrog *et al.* 2006). Conversely, introgressed mitochondrial haplotypes from grey wolves, *Canis lupus* (Lehman *et al.* 1991), and domestic dogs, *Canis familiaris* (Adams *et al.* 2003), have been recovered from populations of coyote, *Canis latrans*, but neither of these has lead to fixation. The consequence of mitochondrial substitutions between recently diverged species appears to be unpredictable, but current data would suggest divergence is mostly spurred by drift and less occasionally by selection.

An important consideration for this study is how reflective DNA barcode data are of general trends in the mitochondrial genome. The barcode region has in fact been previously used as a predictor of variation in nucleotide composition across the mitochondrial genome (Min & Hickey 2007; Clare *et al.* 2008), but that is not to say that mutation rates cannot vary between mitochondrial genes. The number of replacement sites occurring in the 'barcoding region' of COI was not significantly different from that of the rest of the COI gene (K. C. R. Kerr, *personal observation*), which confirms that DNA barcodes provide an overall representation of COI. Across the genome, COI has been known for its conservative substitution rate (Lynch & Jarrell 1993). A comprehensive summary of substitution rates in vertebrate mitochondrial genomes suggested that the rate increases with distance from the origin (Broughton & Reneau 2006). Observing this pattern within birds is hindered because the origin of replication for the light strand is as of yet undetermined in the avian mitochondrial genome (Desjardins & Morais 1990). Regardless, this pattern would not likely be apparent in this study, given that divergence rates between most genes did not differ significantly.

Conclusions

Overall, I found no clear evidence for recurrent selective sweeps in avian DNA barcodes. While barcode sequences do not match neutral predictions, the impression is that evolution in COI is largely governed by purifying selection. Nucleotide divergence between closely related species appears mostly attributable to drift. While it is possible that positive selection could be acting on other mitochondrial genes, it is unlikely to be a major force acting on COI, at least within birds, given the rarity of amino acid substitutions. Because of the linkage of the entire mitochondrial genome, it is challenging to study its evolution without examining the entire genome. As large-scale DNA sequencing becomes more accessible, there will be growth in the number of sequenced whole mitochondrial genomes. Excepting the chicken (*Gallus gallus*), no avian species is represented by more than one mitochondrial genome in GenBank, but intraspecific mitochondrial genomic variation has yielded insights into the evolutionary process for other organisms, such as gadine fish (Marshall *et al.* 2009) and humans (Mishmar *et al.* 2003), and this should likely be viewed as a fruitful direction for future research.

Acknowledgements

This work was funded by an Ontario Graduate Scholarship and through the author's thesis advisor at the University of Guelph, Paul Hebert, by grants from NSERC and Genome Canada. I thank Justin Schonfeld for the computational support necessary to produce the secondary structure predictions. I thank Teresa Crease, Steve Lougheed and Paul Hebert for helpful input and discussion. I thank Beren Robinson and Lee-Ann Hayek for providing assistance with statistical methods. I also thank three anonymous reviewers for comments on an earlier version of this manuscript.

References

Adams JR, Leonard JA, Waits LP (2003) Widespread occurrence of a domestic dog mitochondrial DNA haplotype in southeastern US coyotes. *Molecular Ecology*, **12**, 541–546.

Bachtrog D, Thornton K, Clark A, Andolfatto P (2006) Extensive introgression of mitochondrial DNA relative to nuclear genes in the *Drosophila yakuba* species group. *Evolution*, **60**, 292–302.

Baker AJ, Tavares ES, Elbourne RF (2009) Countering criticisms of single mitochondrial DNA gene barcoding in birds. *Molecular Ecology Resources*, **9**, 257–267.

Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. *Molecular Ecology*, **13**, 729–744.

Bazin E, Glemin S, Galtier N (2006) Population size does not influence mitochondrial genetic diversity in animals. *Science*, **312**, 570–572.

Berlin S, Tomaras D, Charlesworth B (2007) Low mitochondrial variability in birds may indicate Hill-Robertson effects on the W chromosome. *Heredity*, **99**, 389–396.

Berry OF (2006) Mitochondrial DNA and population size. *Science*, **314**, 1388–1390.

Betts MJ, Russell RB (2003) Amino acid properties and consequences of substitutions. In: *Bioinformatics for Geneticists* (eds Barnes MR & Gray IC), p. 408. Wiley, Chichester.

Broughton RE, Reneau PC (2006) Spatial covariation of mutation and nonsynonymous substitution rates in vertebrate mitochondrial genomes. *Molecular Biology and Evolution*, **23**, 1516–1524.

Burton RS, Ellison CK, Harrison JS (2006) The sorry state of F-2 hybrids: consequences of rapid mitochondrial DNA evolution in allopatric populations. *American Naturalist*, **168**, S14–S24.

Clare EL, Kerr KCR, von Konigslow TE, Wilson JJ, Hebert PDN (2008) Diagnosing mitochondrial DNA diversity: applications of a sentinel gene approach. *Journal of Molecular Evolution*, **66**, 362–367.

Desjardins P, Morais R (1990) Sequence and gene organization of the chicken mitochondrial genome—a novel gene order in higher vertebrates. *Journal of Molecular Biology*, **212**, 599–634.

Drummond AJ, Ashton B, Cheung M *et al.* (2007) Geneious v3.0, Available from http://www.geneious.com/.

Egea R, Casillas S, Barbadilla A (2008) Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Research*, **36**, W157–W162.

Frezal L, Leblois R (2008) Four years of DNA barcoding: current advances and prospects. *Infection Genetics and Evolution*, **8**, 727–736.

Fry AJ (1999) Mildly deleterious mutations in avian mitochondrial dna: evidence from neutraliy tests. *Evolution*, **53**, 1617–1620.

Galtier N, Depaulis F, Barton NH (2000) Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics*, **155**, 981–987.

Gerber AS, Loggins R, Kumar S, Dowling TE (2001) Does nonneutral evolution shape observed patterns of DNA variation in animal mitochondrial genomes? *Annual Review of Genetics*, **35**, 539–566.

Hackett SJ, Kimball RT, Reddy S *et al.* (2008) A phylogenomic study of birds reveals their evolutionary history. *Science*, **320**, 1763–1768.

Hasegawa M, Cao Y, Yang ZH (1998) Preponderance of slightly deleterious polymorphism in mitochondrial DNA: nonsynonymous/synonymous rate ratio is much higher within species than between species. *Molecular Biology and Evolution*, **15**, 1499–1505.

Hedrick PW (1980) Hitchhiking: a comparison of linkage and partial selfing. *Genetics*, **94**, 791–808.

Hickerson MJ, Meyer CP, Moritz C (2006) DNA barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology*, **55**, 729–739.

Hiebl I, Braunitzer G, Schneeganss D (1987) The primary structures of the major and minor hemoglobin components of adult Andean goose (*Chloephaga melanoptera*, Anatidae)—the mutation leu-ser in position 55 of the beta-chains. *Biological Chemistry Hoppe-Seyler*, **368**, 1559–1569.

Irwin DE, Rubstov AS, Panov EV (2009) Mitochondrial introgression and replacement between yellowhammers (*Emberiza citrinella*) and pine buntings (*E. leucocephalos*; Aves, Passeriformes). *Biological Journal of the Linnean Society*, **98**, 422–438.

Kenyon L, Moraes CT (1997) Expanding the functional human mitochondrial DNA database by the establishment of primate xenomitochondrial cybrids. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 9131–9135.

Kerr KCR, Stoeckle MY, Dove CJ *et al.* (2007) Comprehensive DNA barcoding coverage of North American birds. *Molecular Ecology Notes*, **7**, 535–543.

Kerr KCR, Birks SM, Kalyakin MV *et al.* (2009a) Filling the gap—COI barcode resolution in eastern Palearctic birds. *Frontiers in Zoology*, **6**, 29. doi:10.1186/1742-9994-6-29.

Kerr KCR, Lijtmaer DA, Barreira AS, Hebert PDN, Tubaro PL (2009b) Probing evolutionary patterns in Neotropical birds through DNA barcodes. *PLoS ONE*, **4**, 6.

Langley CH, Macdonald J, Miyashita N, Aguade M (1993) Lack of correlation between interspecific divergence and intraspecific polymorphism at the suppressor of forked region in *Drosophila melanogaster* and *Drosophila simulans*. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 1800–1803.

Lehman N, Eisenhawer A, Hansen K *et al.* (1991) Introgression of coyote mitochondrial DNA into sympatric North American gray wolf populations. *Evolution*, **45**, 104–119.

Liang YH, Liu XZ, Liu SH, Lu GY (2001) The structure of greylag goose oxy haemoglobin: the roles of four mutations compared with barheaded goose haemoglobin. *Acta Crystallographica Section D-Biological Crystallography*, **57**, 1850–1856.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

Lynch M, Jarrell PE (1993) A method for calibrating molecular clocks and its application to animal mitochondrial DNA. *Genetics*, **135**, 1197–1208.

Marshall HD, Coulson MW, Carr SM (2009) Near neutrality, rate heterogeneity, and linkage govern mitochondrial genome evolution in Atlantic Cod (*Gadus morhua*) and other gadine fish. *Molecular Biology and Evolution*, **26**, 579–589.

Meiklejohn CD, Montooth KL, Rand DM (2007) Positive and negative selection on the mitochondrial genome. *Trends in Genetics*, **23**, 259–263.

Min XJ, Hickey DA (2007) DNA barcodes provide a quick preview of mitochondrial genome composition. *PLoS ONE*, **2**, 5.

Mindell DP, Sorenson MD, Dimcheff DE (1998) An extra nucleotide is not translated in mitochondrial ND3 of some birds and turtles. *Molecular Biology and Evolution*, **15**, 1568–1571.

Mishmar D, Ruiz-Pesini E, Golik P *et al.* (2003) Natural selection shaped regional mtDNA variation in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 171–176.

Mulligan CJ, Kitchen A, Miyamoto MM (2006) Comment on ''Population size does not influence mitochondrial genetic diversity in animals''. *Science*, **314**, 1390.

Nachman MW, Boyer SN, Aquadro CF (1994) Nonneutral evolution at the mitochondrial NADH dehydrogenase subunit 3-gene in mice. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 6364–6368.

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Paland S, Lynch M (2006) Transitions to asexuality result in excess amino acid substitutions. *Science*, **311**, 990–992.

Pereira L, Cunha C, Amorim A (2004) Predicting sampling saturation of mtDNA haplotypes: an application to an enlarged Portuguese database. *International Journal of Legal Medicine*, **118**, 132–136.

Rand DM, Kann LM (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. *Molecular Biology and Evolution*, **13**, 735–748.

Rand DM, Kann LM (1998) Mutation and selection at silent and replacement sites in the evolution of animal mitochondrial DNA. *Genetica*, **102–3**, 393–407.

Ratnasingham S, Hebert PDN (2007) BOLD: the barcode of life data system (http://www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364.

Schmidt TR, Wu W, Goodman M, Grossman LI (2001) Evolution of nuclear- and mitochondrial-encoded subunit interaction in cytochrome c oxidase. *Molecular Biology and Evolution*, **18**, 563–569.

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**, 1596–1599.

Team RDC (2007) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Tieleman BI, Versteegh MA, Fries A *et al*. (2009) Genetic modulation of energy metabolism in birds through mitochondrial function. *Proceedings of the Royal Society B-Biological Sciences*, **276**, 1685–1693.

Tsukihara T, Aoyama H, Yamashita E *et al*. (1996) The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 angstrom. *Science*, **272**, 1136–1144.

Wang ZO, Pollock DD (2007) Coevolutionary patterns in cytochrome c oxidase subunit I depend on structural and functional context. *Journal of Molecular Evolution*, **65**, 485–495.

Wares JP, Barber PH, Ross-Ibarra J, Sotka EE, Toonen RJ (2006) Mitochondrial DNA and population size. *Science*, **314**, 1388–1389.

Wilson CC, Bernatchez L (1998) The ghost of hybrids past: fixation of arctic charr (*Salvelinus alpinus*) mitochondrial DNA in an introgressed population of lake trout (*S. namaycush*). *Molecular Ecology*, **7**, 127–132.

Wise CA, Sraml M, Easteal S (1998) Departure from neutrality at the mitochondrial NADH dehydrogenase subunit 2 gene in humans, but not in chimpanzees. *Genetics*, **148**, 409–421.

Yang ZH (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

Zink RM (2005) Natural selection on mitochondrial DNA in *Parus* and its relevance for phylogeographic studies. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **272**, 71–78.

Zink RM, Drovetski SV, Rohwer S (2006) Selective neutrality of mitochondrial ND2 sequences, phylogeography and species limits in *Sitta europaea*. *Molecular Phylogenetics and Evolution*, **40**, 679–686.

Data Accessibility

DNA sequences new to this manuscript: GenBank accession numbers HM033200–HM034025. GenBank accessions for whole mitochondrial genomes analysed in this study are listed in Supplementary Table S2 (Supporting information). Both BOLD and GenBank accessions for COI sequences analysed in this study are listed in Supplementary Table S3 (Supporting information).

Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1 Species included in the assessment of genetic diversity.

Table S2 GenBank accession numbers for the 22 species with whole mitochondrial genomes that were included in this analysis.

Table S3 Sequences included in the assessment of genetic diversity, tests of neutrality, and amino acid variation.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.