MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) **15**, 921–931 doi: 10.1111/1755-0998.12344

Barcoding Neotropical birds: assessing the impact of nonmonophyly in a highly diverse group

BÁRBARA R. N. CHAVES,* ANDERSON V. CHAVES,*† AUGUSTO C. A. NASCIMENTO,* JULIANA CHEVITARESE,* MARCELO F. VASCONCELOS‡ and FABRÍCIO R. SANTOS*

**Laboratório de Biodiversidade e Evolução Molecular, Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas, Avenida Antônio Carlos, 6627, Bloco L3, Sala 244, Belo Horizonte, CEP: 31270-901 Minas Gerais, Brazil, †Programa de Pós-Graduação em Manejo e Conservação de Ecossistemas Naturais e Agrários, Universidade Federal de Viçosa, Instituto de Ciências Biológicas e da Saúde, campus Florestal, Rodovia LMG-818, km 6, Florestal, CEP: 35690-000 Minas Gerais, Brazil, ‡Museu de Ciências Naturais, Pontifícia Universidade Católica de Minas Gerais, Avenida Dom José Gaspar, 290, Belo Horizonte, CEP: 30535-610 Minas Gerais, Brazil*

Abstract

In this study, we verified the power of DNA barcodes to discriminate Neotropical birds using Bayesian tree reconstructions of a total of 7404 COI sequences from 1521 species, including 55 Brazilian species with no previous barcode data. We found that 10.4% of species were nonmonophyletic, most likely due to inaccurate taxonomy, incomplete lineage sorting or hybridization. At least 0.5% of the sequences (2.5% of the sampled species) retrieved from GenBank were associated with database errors (poor-quality sequences, NuMTs, misidentification or unnoticed hybridization). Paraphyletic species (5.8% of the total) can be related to rapid speciation events leading to nonreciprocal monophyly between recently diverged sister species, or to absence of synapomorphies in the small COI region analysed. We also performed two series of genetic distance calculations under the K2P model for intraspecific and interspecific comparisons: the first included all COI sequences, and the second included only monophyletic taxa observed in the Bayesian trees. As expected, the mean and median pairwise distances were smaller for intraspecific than for interspecific comparisons. However, there was no precise 'barcode gap', which was shown to be larger in the monophyletic taxon data set than for the data from all species, as expected. Our results indicated that although database errors may explain some of the difficulties in the species discrimination of Neotropical birds, distance-based barcode assignment may also be compromised because of the high diversity of bird species and more complex speciation events in the Neotropics.

Keywords: birds, DNA barcoding, neotropical fauna, passerines, phylogenetic methods, speciation

Received 23 May 2014; revision received 31 October 2014; accepted 3 November 2014

Introduction

Initial DNA barcoding studies suggested the existence of a 'barcoding gap' between intra- and interspecific variation (Hebert *et al.* 2003, 2004), but recent studies attributed it to insufficient sampling by showing a significant overlap for many taxa (Goldstein *et al.* 2000; Moritz & Cicero 2004; Meyer & Paulay 2005; Baker *et al.* 2009; Kerr *et al.* 2009a,b; Tavares *et al.* 2011).

However, the validity of clusters obtained using genetic distances has been criticized because of insufficient taxonomic knowledge and geographical sampling (Moritz & Cicero 2004; Prendini 2005), the use of a single locus (Mallet & Willmott 2003; Will & Rubinoff 2004; Knowles & Carstens 2007) and the frequentist framework employed for the assignment of individuals to species (Nielsen & Matz 2006). It has been suggested that DNA barcoding methods should consider phylogenetic information regarding the evolutionary relationships among species for the assignment of individuals to their corresponding species (Nielsen & Matz 2006; Vogler & Monaghan 2007).

The delimitation of species is of central importance in barcoding studies, as in biology in general, but this issue is generally confused with the problem of species concepts. There is a range of existing species concepts, but they all include a common element, which refers to the possibility of being diagnosable because of evolutionary independence (Goldstein & DeSalle 2000; De Queiroz 2007; Aleixo 2007). However, barcoding approaches frequently ignore the evolutionary diversification of species, prioritizing the use of genetic distances to delimit taxa (Vogler & Monaghan 2007).

Correspondence: Fabrício R. Santos, Fax: +55-31-3409-2570; E-mail: fsantos@icb.ufmg.br

A phylogenetic tree is a graphic representation of common ancestry, where the concept of monophyly, that is groups that include a common ancestor and all of their descendants, is critical (Farris 1974). Monophyly has long (but not always) been used as a criterion for species delimitation (phylogenetic concept of species; Donoghue 1985; Mishler 1985; De Queiroz 2007), and nonmonophyly can sometimes result from inaccurate taxonomy when the phenotypic boundaries of nominal species do not reflect the history of evolutionary entities, as in oversplitting or overlumping (Funk & Omland 2003). In fact, Meyer & Paulay (2005) showed that DNA barcodes can provide robust specimen assignment only for taxa whose taxonomy is well understood and when representative specimens are thoroughly sampled. Apart from problems in taxonomy, a general source of nonmonophyly is the incomplete lineage sorting of allelic lineages, which is a nonintraspecific coalescence of DNA lineages, usually due to recent speciation (Neigel & Avise 1986; Avise 2000; Knowles & Carstens 2007), such as through peripatric events (Losos & Glor 2003). Disregarding issues related to tree branching resolution, another cause of nonmonophyly is occasional mating between distinct species (hybridization), which can generate individuals having the morphology of one species but the mitochondria of another one. Thus, appropriate identification of hybrids requires comparison between mitochondrial DNA and morphological data or nuclear genes (Funk & Omland 2003).

If left undetected, nonmonophyly compromises evolutionary inferences based on trees that are erroneously assumed to accurately depict species trees (Funk & Omland 2003). Based on the analyses of data on 331 bird species retrieved from 74 studies, Funk & Omland (2003) estimated a proportion of 16.7% of nonmonophyletic species. However, the above-cited sources of nonmonophyly (inaccurate taxonomy, incomplete lineage sorting and hybridization) are not the only ones. Indeed, even if all species analysed are in fact monophyletic, a reconstructed gene tree may exhibit false nonmonophyletic groupings, which do not represent their real history. Harris (2003) warns scientists to be aware that the quality of sequences in molecular banks is not always optimal and suggests that sequences that are phylogenetically unusual should be checked because errors in published raw data are extremely widespread. There are three main problems in deposited DNA data reflecting unusual tree topologies: misidentified samples; poor-quality sequences or nuclear mtDNA insertions (NuMTs); and absent or insufficiently represented species. If disregarded, these issues may affect calculations of intra- and interspecific genetic distances in the barcode methodology.

The final possible sources of nonmonophyly are the phylogenetic methods of tree reconstruction and the genetic marker itself. A small gene fragment may provide too few synapomorphies to recover a robust gene tree, and moreover, a robust gene tree may not match the species tree. In fact, gene tree/species tree problems represent major limitations on evolutionary inferences made from single loci, such as under the barcoding approach (Funk & Omland 2003). Concerning tree reconstruction methods, the choice of more reliable approaches, such as maximum likelihood or Bayesian algorithms with a realistic mutational model, represents an attempt to obtain a more consistent gene tree (Barton *et al.* 2010).

An underestimated number of bird species is expected in early studies because there is a great diversity and strong geographic structure presenting highly divergent clades indicating the existence of likely unnamed taxa in the Neotropics (Chaves *et al.* 2008; Kerr *et al.* 2009b; Tavares *et al.* 2011; Milá *et al.* 2012). In this study, we increased the representation of Neotropical bird species in the DNA barcode database by adding samples from Brazil, particularly from Cerrado and Atlantic Forest biomes, which are two important biodiversity hotspots with few barcode sequences so far (Vilaça *et al.* 2006; Chaves *et al.* 2008; Tavares *et al.* 2011). It can be expected that denser geographical and taxonomical sampling may result in the discovery of new clusters and perhaps in the reduction of divergence between them (Vogler & Monaghan 2007). Rather than accepting arbitrary groupings, such as species boundaries based on genetic divergences, we took into account groupings that correspond to collections of reproductively coherent individuals, which are considered to represent a more likely approximation of true species in nature, as suggested by Vogler & Monaghan (2007). To delimitate those groupings, we performed Bayesian tree reconstructions and evaluated the species monophyly.

Methods

Our sample consisted of 515 individuals from 305 Brazilian bird species deposited in the tissue collection of the *Laboratório de Biodiversidade e Evolução Molecular* (LBEM) in the *Universidade Federal de Minas Gerais* (UFMG), including 55 species never barcoded previously and 221 vouchers from 170 species (Table S1). The vouchers are deposited in one of two zoological collections: the *Centro de Coleções Taxonômicas da UFMG* or the *Museu de Ciências Naturais da Pontifícia Universidade Católica de Minas Gerais* (MCNA). The samples mainly came from passerines, mostly Tyrannidae (44 species), Furnariidae (30), Thamnophilidae (28), Thraupidae (24) and Emberizidae (20) (Table 2). The examined specimens were collected in the Amazon, Caatinga, Pantanal, Cerrado and Atlantic Forest, mainly in the last two biomes (Fig. 1).

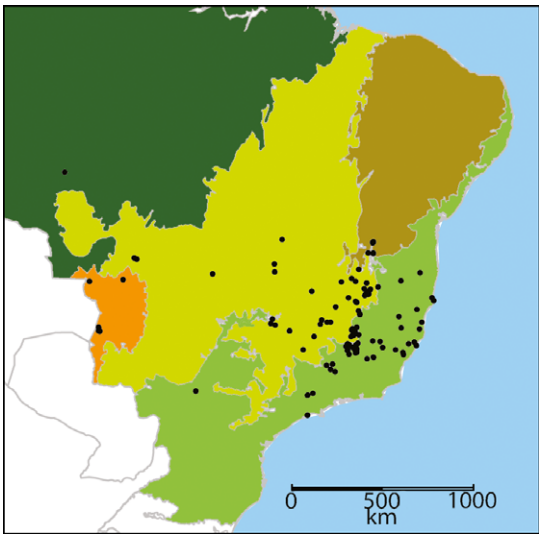

Fig. 1 Collection points for the Brazilian samples from Brazil added by this study to the DNA barcode database. *Dark green*: Amazon; *Yellow*: Cerrado; *Brown*: Caatinga; *Orange*: Pantanal; *Light green*: Atlantic Forest; *Purple*: Grassland.

Total genomic DNA was extracted from tissues using the phenol–chloroform protocol (Sambrook & Russell 2001) and quantified in a NanoDrop 2000c (Thermo Scientific). Polymerase chain reaction (PCR) amplification with universal primers and sequencing of the standardized DNA barcode from 5′ end of the COI gene were performed (Table 1). High-quality chromatograms were obtained using an automatic MegaBACE 1000 sequencer (GE Healthcare Life Sciences), and a consensus sequence for each individual was generated with the Phred-Phrap and Consed packages (Ewing & Green 1998; Ewing *et al.* 1998; Gordon *et al.* 1998). Multiple alignments and visual editing of COI sequences (using translated amino acids) were performed using the Clustal W alignment algorithm (Thompson *et al.* 1994) in MEGA software v. 5 (Tamura *et al.* 2011). The new COI sequences produced in this study and details on the analysed specimens are deposited under the 'Barcoding of Brazilian Birds' (BBB) project, with the tag 'Brazilian Barcode of Life' (BrBOL), in the Barcode of Life Data System (BOLD Accession nos BBB319-13 to BBB421-13, GenBank Accession nos KM896216–KM896655).

To verify the power of DNA barcoding to identify specimens correctly, COI sequences from Neotropical bird species were obtained from GenBank (Table S1) to allow phylogenetic comparisons. The sequences extracted from GenBank were selected according to the keywords 'coi[gene] AND *family*[Organism]', where *family* was the name of avian families occurring in South America (according to BirdLife International & NatureServe 2013). We adopted the nomenclature of the Brazilian Committee of Ornithological Records available in 2013 (CBRO, http://www.cbro.org.br). Only sequences from the 5′ end of the COI gene were used, but sequences with too many nucleotide ambiguities were excluded from the analysis. To include all sequences from this study and as many as possible from GenBank, we used 452-bp sequence-length alignments. A total of 6894 sequences from 1466 species from GenBank were analysed in combination with our sample (Table 2).

Because mtDNA monophyly is an expected outcome of a reproductively coherent species, we used a detailed phylogenetic approach to analyse COI alignments. Bayesian tree reconstructions were performed in MrBayes v3.2.1 (Ronquist *et al.* 2012) at the Cipres Science Gateway (Miller *et al.* 2010). The best-fit model (General Time Reversible with the proportion of invariable sites and gamma, GTR+I+G) was selected using a sample from the original data set including 1–5 representative samples from each bird family and applying the Akaike information criterion (AIC) (Posada & Buckley 2004) based on likelihood scores from PAUP* (Swofford 1998),

Table 1 Universal primers used in this study to amplify the 5′ end of the COI gene

Primer	3′–5′ Sequence	Reference
L6615	CCYCTGTAAAAAGGWCTACAGCC	Sorenson (2003)
H8121	GGGCAGCCRTGRATTCAYTC	Sorenson (2003)
H6035	CCTCCTGCAGGGTCAAAGAATGT	Chaves *et al.* (2008)
socoiF1	TTCTACAAACCATAAAGATATTGGCA	Chaves *et al.* (2008)
LCO1490	GGTCAACAAATCATAAAGATATTGG	Folmer *et al.* (1994)
HCO2198	TAAACTTCAGGGTGACCAAAAAATCA	Folmer *et al.* (1994)
VF1_t1	TGTAAAACGACGGCCAGTTCTCAACCAACCACAAAGACATTGG	Ivanova *et al.* (2007)
VF1d_t1	TGTAAAACGACGGCCAGTTCTCAACCAACCACAARGAYATYGG	Ivanova *et al.* (2007)
VF1i_t1	TGTAAAACGACGGCCAGTTCTCAACCAACCAIAAIGAIATIGG	Ivanova *et al.* (2007)
VR1_t1	CAGGAAACAGCTATGACTAGACTTCTGGGTGGCCAAAGAATCA	Ivanova *et al.* (2007)
VR1d_t1	CAGGAAACAGCTATGACTAGACTTCTGGGTGGCCRAARAAYCA	Ivanova *et al.* (2007)
VR1i_t1	CAGGAAACAGCTATGACTAGACTTCTGGGTGICCIAAIAAICA	Ivanova *et al.* (2007)
M13F(-21)	TGTAAAACGACGGCCAGT	Messing (1983)
M13R(-27)	CAGGAAACAGCTATGAC	Messing (1983)

Table 2 Sample size included in the analysis, distributed by bird groups

Bird Group	Total species	New species	Total specimens	Specimens per species	Single-specimen species	Multispecimen species
Aves	1521	55	7409	4.87	257	1264
Nonpasserines	651	16	2585	3.97	133	518
Passerines	870	39	4824	5.54	124	746
Oscines	342	13	1934	5.65	41	301
Suboscines	528	26	2890	5.47	83	445

in the program MRMODELTEST v2.3 (Nylander 2004). The species were classified as monophyletic or nonmonophyletic, according to the coalescence of their COI lineages, verified by visual inspection of Bayesian trees. We followed Meyer & Paulay's (2005) phylogenetic terminology to discriminate nonmonophyly in paraphyly and polyphyly, but low branching resolution prevented the characterization of these trees in some cases.

Genetic distances were calculated under the Kimura 2-parameter model (K2P) for all pairwise comparisons using MEGA software. We wrote R scripts to separate, summarize and compare the mean and median of intra- and interspecific genetic distances and to calculate the mean barcoding gap using the smallest interspecific and the largest intraspecific distances for each species (Meier *et al.* 2008). To verify interferences caused by the occurrence of nonmonophyly (inaccurate taxonomy, incomplete lineage sorting, hybridization, misidentification of samples, poor-quality sequences or NuMts), we conducted two different calculations of intra- and interspecific genetic distances: one including all-sequence data, and another including only monophyletic species. The interspecific genetic distances were calculated only between species of the same genus. We performed two series of genetic distance calculations: the first encompassed all sequences with their original identification (all-sequence data set), while the second encompassed only monophyletic species (only-monophyletic data set). In the second calculation, we excluded any likely problematic sequences, one-specimen species making another species paraphyletic, and nonmonophyletic species.

Results

From the total of 1521 bird species analysed here, 1264 were represented by multiple individuals and 257 (16.9%) by a single individual (Table 2). Based on the examination of trees, the vast majority (89.6%) of multi-individual species were classified as monophyletic, while 73 were paraphyletic, 51 were polyphyletic, and 17 presented a not discriminated nonmonophyly (Tables 3 and S5), with some species showing both paraphyly and polyphyly. Forty sequences (0.5%) from 32 species (2.5%) were diagnosed as data source errors in the GenBank sample (due to misidentification, hybridization, poor-quality sequences or NuMts) based on noncoalescence in the Bayesian tree (Tables 3 and S4). Among the 257 single-specimen species analysed, the vast majority (87.2%) formed a new branch in the Bayesian tree, appearing to be phylogenetically independent lineages in relation to other species, but 33 nested within another species clade (Table 4).

As expected, the mean and median Kimura 2-parameter (K2P) pairwise distances were smaller for intraspecific than for interspecific comparisons (Table 5, Fig. 3). There was a marked difference between the mean and median intraspecific distances (1.8% and 0.4%, respectively), indicating the effect of outliers, probably because of sequencing errors, inaccurate taxonomy, incomplete lineage sorting, hybridization, misidentification of samples, poor-quality sequences or NuMts. Indeed, there were more species showing intraspecific distances below than above 0.4%. Moreover, species presenting intraspecific distances above 0.4%

Table 3 Analysis of monophyly in each bird group according to Bayesian trees built from 452 bp of the 5′ COI gene. The phylogenetic terminology (monophyly, paraphyly and polyphyly) followed Meyer & Paulay (2005), but low resolutions prevented discrimination in some cases (other). Possible GenBank errors comprise misidentifications, poor-quality sequences, NUMTs and introgressions. See also Fig. 5

Bird Group	Monophyletic	Paraphyletic	Polyphyletic	Other	Total Nonmonophyletic	GENBANK Error
Aves	1133 (89.6%)	73 (5.8%)	51 (4.0%)	17 (1.3%)	131 (10.4%)	32 (2.5%)
Nonpasserines	471 (90.9%)	31 (6.0%)	17 (3.3%)	1 (0.2%)	47 (9.1%)	6 (1.2%)
Passerines	662 (88.7%)	42 (5.6%)	34 (4.6%)	16 (2.1%)	84 (11.3%)	26 (3.5%)
Oscines	266 (88.4%)	16 (5.3%)	18 (6.0%)	5 (1.7%)	35 (11.6%)	7 (2.3%)
Suboscines	396 (89.0%)	26 (5.8%)	16 (3.6%)	11 (2.5%)	49 (11.0%)	19 (4.3%)

were represented by larger samples on average (Table 6). There were no differences between the mean and median interspecific distances (8.1%) obtained for the birds. The distribution histograms (Fig. 2) of genetic distances showed that intraspecific differences generally tended to be smaller than interspecific distances; however, there was no precise and marked 'barcoding gap' between them. Rather, there was a quite substantial overlap between the two levels.

For the only-monophyletic distance calculation, 40 problematic sequences and 983 sequences from 121 non-monophyletic species were excluded (Tables S3 and S4). In this calculation, the average barcoding gap was shown to be 1.8% larger; interspecific distances of 0–1% were less frequent; and the intraspecies diversity histogram showed a shorter tail of large distances compared with the all-sequence calculation (Table 5, Figs 2 and 3).

In the all-sequence calculation, we detected 54 species (6.8%) sharing haplotypes with another species and 147 (11.6%) species that showed intraspecific distances that were greater than interspecific distances (overdivergent species). However, in the only-monophyletic calculation, no single shared haplotype was detected, and only 14 (1.2%) of the species were overdivergent (Table 7).

When the pairwise genetic distances among different bird groups were compared, different patterns could be detected. Passerine birds (order Passeriformes) showed higher average intraspecific genetic divergences (2.0%) compared with nonpasserines (0.9%). On the other hand, nonpasserines showed a larger average barcode gap (4.0%) in relation to passerines (2.3%), indicating more informative barcode identification in nonpasserines (Table 5 and Fig. 4). Indeed, passerines presented more nonmonophyletic species (11.3%, against 9.1% in non-passerines), more species sharing haplotypes (4.6%, against 2.2% in nonpasserines) and more overdivergent species (12.1%, against 6.5% in nonpasserines) compared

Table 4 Branching of single-specimen species in Bayesian trees built from 452 bp of the 5′ COI gene

Bird Group	Single-specimen species	New branch	Other species' branch
Aves	257	224 (87.2%)	33 (12.8%)
Nonpasserines	133	114 (85.7%)	19 (14.3%)
Passerines	124	110 (88.7%)	14 (11.3%)
Oscines	41	36 (87.8%)	5 (12.2%)
Suboscines	83	74 (89.2%)	9 (10.8%)

Table 5 Mean and median Kimura 2-parameter model (K2P) pairwise genetic distances and average barcode gaps (using the smallest interspecific and bigger intraspecific distances for each species) for bird groups. See also Figs 2, 3 and 4. (A) All-sequences calculation and (B) only-monophyletic species calculation

(A) Bird Clade	Mean intraspecies	Median intraspecies	Mean interspecies	Median interspecies	Average barcode gap
Aves	1.8%	0.4%	8.1%	8.1%	2.9%
Nonpasserines	0.9%	0.2%	8.5%	8.7%	4.0%
Passerines	2.0%	0.7%	7.9%	7.8%	2.3%
Oscines	1.4%	0.4%	7.0%	6.8%	2.5%
Suboscines	2.4%	0.9%	9.0%	9.2%	2.2%

(B) Bird Group	Mean intraspecies	Median intraspecies	Mean interspecies	Median interspecies	Mean barcode gap
Aves	1.7%	0.4%	8.7%	8.6%	4.7%
Nonpasserines	0.6%	0.2%	9.2%	9.1%	5.5%
Passerines	1.9%	0.4%	8.5%	8.1%	4.3%
Oscines	0.9%	0.2%	7.3%	6.6%	4.0%
Suboscines	2.4%	0.7%	9.7%	10.3%	4.5%

Table 6 Proportion of pairwise intraspecific Kimura 2-parameter model (K2P) distances above and below 0.4% (0.4% corresponds to the median intraspecific genetic distance). (A) All-sequence calculation and (B) only-monophyletic calculation

	Species		Individuals per species	
	A	B	A	B
Intraspecific Distance < 0.4%	1138 (62%)	1051 (64%)	5.88	5.70
Intraspecific Distance >0.4%	693 (38%)	568 (36%)	7.31	7.31

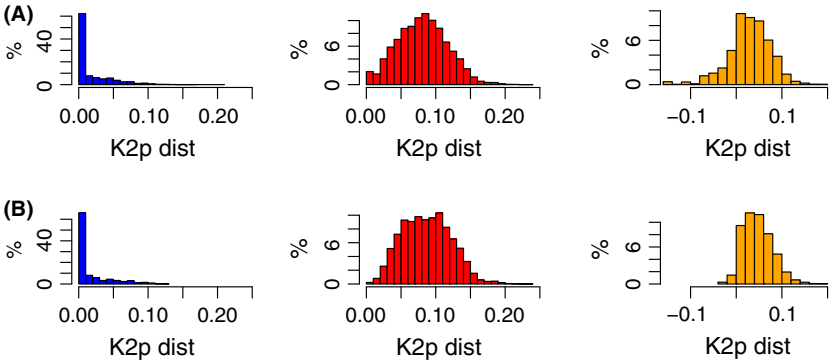

Fig. 2 Histograms of the distribution of Kimura 2-parameter model (K2P) distances and barcoding gaps. (A) All-sequence calculation and (B) only-monophyletic calculation. *Blue:* Intraspecific distances; *Red*: Interspecific distances; *Orange*: Barcoding gaps.

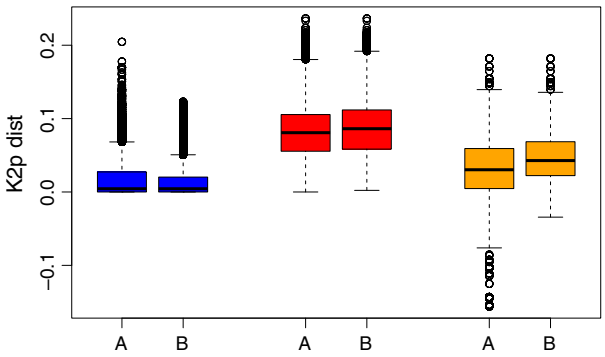

Fig. 3 Boxplots of Kimura 2-parameter model (K2P) pairwise genetic distances and barcoding gaps. See also Table 5. (A) All-sequence calculation and (B) only-monophyletic calculation. *Blue:* Intraspecific distances; *Red*: Interspecific distances; *Orange*: Barcoding gaps.

with nonpasserines. Additionally, the sequence errors were more overspread throughout passerines (3.5%) than nonpasserines (1.2%) (Table 7 and Figs 5 and 6).

Some passerine taxa also presented different patterns. Suboscines birds (suborder Tyranii) exhibited greater average genetic distances in the two levels of comparison (2.4% of intraspecific distance and 9.0% of interspecific distance) compared with Oscines birds (suborder Passeri) (1.4% and 7.0%, respectively, Table 5 and Fig. 4). Moreover, Suboscines presented fewer species sharing haplotypes (3.6%, against 6.1% in Oscines) and fewer overdivergent species (11.7%, against 12.6% in Oscines, Table 7 and Fig. 6), indicating a better power of barcode assignment in this clade.

Discussion

The observed proportion of 10.4% of nonmonophyletic species of Neotropical birds was smaller than the estimation reported by Funk & Omland (2003) (16.7%). This difference may be a reflection of better sampling, as we used approximately four times the number of bird species. Zwickl & Hillis (2002) and Pollock *et al.* (2002) found that an increase in taxon sampling results in a greatly reduced phylogenetic estimation error.

Table 7 Numbers of nonmonophyletic, haplotype-sharing and overdivergent species and species with one or more erroneous sequence (misidentifications, poor-quality sequences, NUMTs or hybridization) among bird groups. See also Figs 5 and 6. (A) All-sequence calculation and (B) only-monophyletic calculation

(A) Bird group	Total species (>1 ind)	Nonmonophyletic species	Sharing haplotype species	Overdivergent species	Errors
Aves	1521	131 (10.4%)	54 (3.6%)	147 (9.7%)	32 (2.5%)
Nonpasserines	651	47 (9.1%)	14 (2.2%)	42 (6.5%)	6 (1.2%)
Passerines	870	84 (11.3%)	40 (4.6%)	105 (12.1%)	26 (3.5%)
Oscines	342	35 (11.6%)	21 (6.1%)	43 (12.6%)	7 (2.3%)
Suboscines	528	49 (11.0%)	19 (3.6%)	62 (11.7%)	19 (4.3%)

(B) Bird group	Total species (>1 ind)	Nonmonophyletic species	Sharing haplotype species	Overdivergent species	Errors
Aves	1379	0	0	14 (1.0%)	0
Nonpasserines	597	0	0	2 (0.3%)	0
Passerines	782	0	0	12 (1.5%)	0
Oscines	305	0	0	7 (2.3%)	0
Suboscines	477	0	0	5 (1.0%)	0

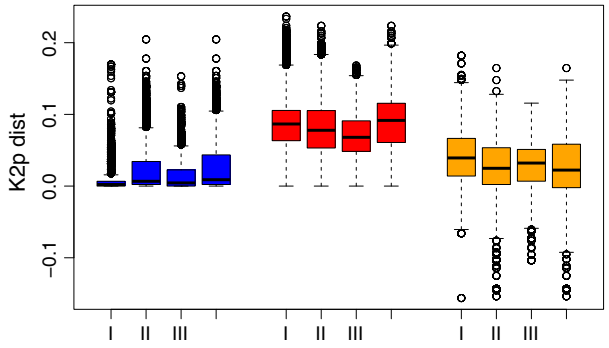

Fig. 4 Boxplots of Kimura 2-parameter model (K2P) pairwise genetic distances and barcoding gaps for each bird group – extracted from the boxplots in Fig. 3. See also Table 5. *Blue:* Intraspecific distances; *Red*: Interspecific distances; *Orange*: Barcoding gaps.

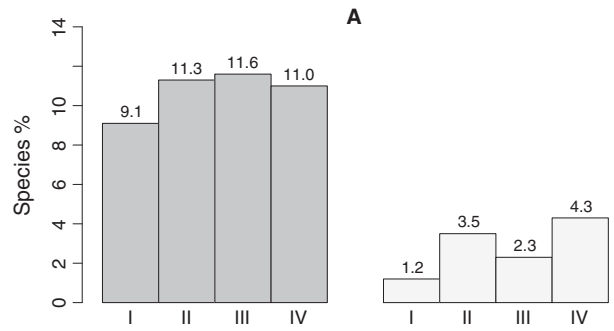

Fig. 5 Proportion of nonmonophyletic species (*grey bars*) and species with one or more erroneous sequence (misidentifications, poor-quality sequences, NUMTs or hybridization – *white bars*) in each bird group, based on Bayesian trees built from 452 bp of the 5′ COI. These sequences were removed to perform the only-monophyletic calculation (see text). See also Table 7. (I) Nonpasserines; (II) Passerines; (III) Oscines; (IV) Suboscines. (A) All sequence calculation.

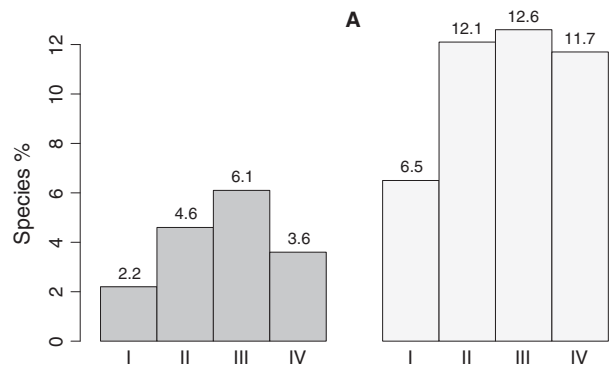

Fig. 6 Proportion of species that were undistinguishable through genetic distances, due to sharing of COI haplotypes (*grey bars*) or to presenting larger intraspecific than interspecific distances (*white bars*). See also Table 7. (I) Nonpasserines; (II) Passerines; (III) Oscines; (IV) Suboscines. (A) All sequence calculation.

The nonmonophyly of some species can be explained by their geographical distributions indicating likely taxonomic uncertainty. The species *Cantorchilus leucotis* (Troglodytidae) and *Polioptila plumbea* (Polioptilidae) were shown to be paraphyletic, with each being composed of distinct lineages occurring in different biomes and separated by wide geographical distances (Fig. 7). These species could correspond to more than one species combined and possibly require taxonomic revision, but additional analyses, for example, examining other loci (e.g. nuclear loci) or morphological traits, are necessary to confirm this.

The nonmonophyly of another species (Table S2) can also be explained by recent speciation leading to incomplete lineage sorting. In peripatric speciation events in particular, when a small population is isolated from a larger ancestral stock, accelerated gene tree progression is expected, bypassing the polyphyletic phase, leading directly to paraphyly (Avise 2000; Losos & Glor 2003). Thus, the existence of monophyletic species with a small geographic range that render a geographically widespread species paraphyletic might indicate the occurrence of recent peripatric speciation (Harrison 1991; Losos & Glor 2003). This seems to be the case for *Manacus vitellinus/M. manacus* (Pipridae) and *Rhytipterna holerythra/R. simplex* (Tyrannidae) (Fig. 8).

Although Bayesian algorithms are known to obtain consistent gene trees, it is possible that some cases of nonmonophyly occurred because of the lack of synapomorphies provided by the small gene fragment analysed, especially in taxa with taxonomy uncertainty.

We found a minimum of 0.5% of sequences (from 2.5% of the total species) associated with database errors (due to misidentification, poor-quality sequences, NuMts or hybridization), a value that is much lower than the 20% observed by Bridge *et al.* (2003) in fungal sequences, which may be a reflection of greater facility to correctly identify birds through morphology (Table 3). As noted by Bock (1969), bird species are better known than many other groups of organisms and are easily observed and studied. However, it is important to investigate, curate and annotate any possible errors as well as to review the taxonomy of some species, to improve the reliability of the DNA barcode database. Maintenance of voucher specimens in public-access collections, linking them with corresponding DNA barcodes through photodocumentation, will allow future morphological re-examination (Agerer *et al.* 2000; Hunter *et al.* 2008). As observed by Meyer & Paulay (2005), evolutionarily significant units (ESUs, population groupings that take into account combined genetic and morphological data) present a reduced number of polyphyletic species compared with traditional, morphological species. Therefore, even well-studied groups such as birds are in need of detailed

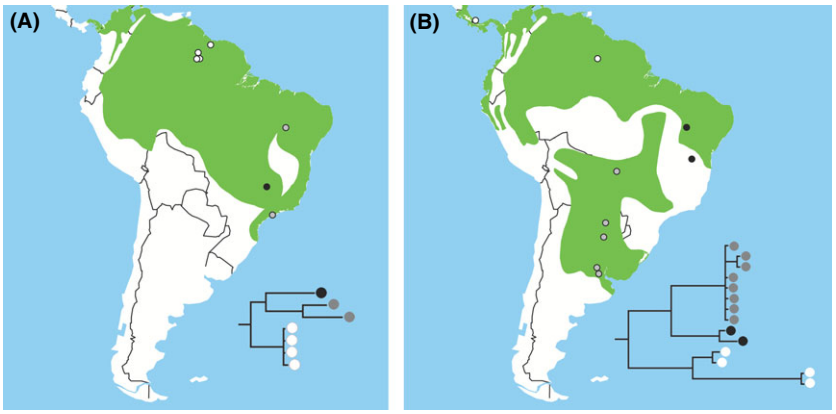

Fig. 7 Maps showing the distribution patterns of two wide-ranging species that exhibited paraphyletic COI lineages. Species ranges are presented in green, and circles indicate collection sites. Black and white circles correspond to the distinct separated lineages of a given species, while grey circles correspond to another species. (A) *Cantorchilus leucotis* (grey circles: *Cantorchilus longirostris*). (B) *Polioptila plumbea* (grey circles: *Polioptila dumicola*).

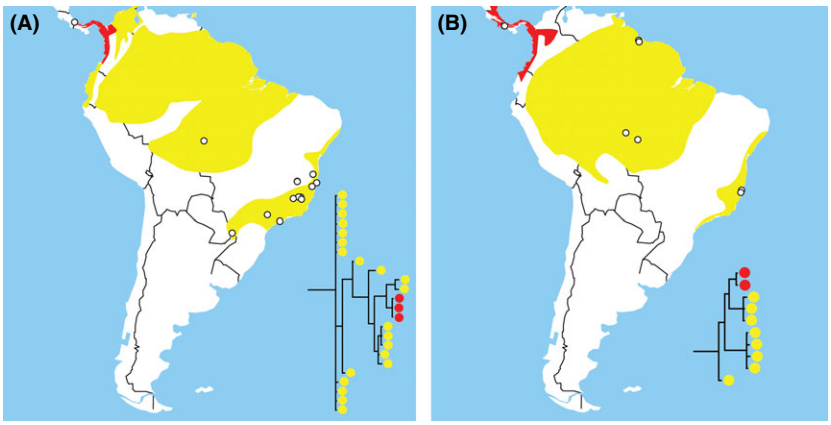

Fig. 8 Maps showing the distribution patterns of two sister species pairs that exhibited paraphyletic COI lineages indicating the occurrence of recent peripatric speciation. The species ranges highlighted in yellow correspond to paraphyletic widespread species, while those in red correspond to monophyletic peripatric species. Circles indicate the collection sites for the two pairs of species. (A) *Manacus vitellinus* (red) and *M. manacus* (yellow). (B) *Rhytipterna holerythra* (red) and *R. simplex* (yellow).

taxonomic revisions before accurate tests of incongruence can be performed.

The incomplete representation of species in the database has been considered to be a likely explanation for incorrect assignments of DNA barcodes (Nielsen & Matz 2006). However, our results showed that the vast majority (87.2%) of single-specimen species nested in a distinct branch in Bayesian trees, separated from all other species (Table 4). Among the 33 single-specimen species that nested in another species' branch, approximately one-third shared their haplotype with those species, suggesting problems of inaccurate taxonomy or incomplete lineage sorting, but only by increasing the numbers of samples and sequences examined may their real relationships be elucidated. In any case, the Bayesian method of tree reconstruction was shown to be informative in identifying sequences that may correspond to new taxa in the database.

The majority of multi-individual species presented intraspecific distances below 0.4%. Some species showed much larger intraspecific distances, probably because of sequencing errors, inaccurate taxonomy, incomplete lineage sorting, hybridization, misidentification of samples, poor-quality sequences or NuMts. The results also indicated that the assessment of a larger sample should lead to larger intraspecific distances (Table 6). This improved genetic assessment would most likely be accompanied by broad geographic sampling, which allows a more accurate estimation of genetic diversity within species. However, the correspondence between sampling size and geographic distribution should be still verified in the database. Kerr *et al.* (2009b) found an underestimation of species diversity and deep intraspecific divergence in birds in Argentina, and the same was observed more recently for Amazonian birds (Tavares *et al.* 2011; Milá *et al.* 2012). These authors also suggest that complex patterns of speciation and regional divergence may have been responsible for more restricted endemic birds and the high diversity of Neotropical bird species. Deep intraspecific divergence, strong geographic structuring and underestimated regional diversity are also present in our data when we examine the Bayesian phylogenetic trees (Data S1) considering the increased sampling in Atlantic Forest and Cerrado biomes. According to Milá *et al.* (2012), phylogeographic splits between ecoregions or biomes were the most divergent within species and varied in genetic distances in different species.

Our results also showed an overlap between intra- and interspecific distances, rather than a marked

barcoding gap, as found by other studies (Meyer & Paulay 2005; Tavares *et al.* 2011). Tavares *et al.* (2011) and Milá *et al.* (2012) also found several species that recovered as nonmonophyletic between closely related species in sympatry. However, the average barcoding gap was larger in the only-monophyletic species calculation, which took into account the monophyly of species and database errors (Figs 2 and 3, Table 5). Furthermore, our results indicate that nonmonophyly and database errors are the main causes of both overdivergent species and species sharing haplotypes. In other words, nonmonophyly of taxa (resulting from inaccurate taxonomy, incomplete lineage sorting or hybridization) and database errors (misidentification or poor-quality sequences) are the main explanations for the problems associated with distance-based barcode assignment (Table 7).

Passerines, or songbirds (order Passeriformes), comprise the largest order of birds, with the highest species richness (Sibley & Monroe 1990) and diversification rates (Jetz *et al.* 2012) in the entire class of birds. Generally, comparisons of the results for passerine and nonpasserine birds must be made with caution because of an overall better sampling of passerine species (5.54 individuals per species) in relation to nonpasserines (3.97 individuals per species, Table 2).

Regarding comparisons of major taxa within passerines, our results showed that the genetic divergences of the Suboscines (in both intraspecific and interspecific levels) were far larger than those of the Oscines (Table 5). Still, the Oscines showed a poorer barcode performance, mainly due to haplotype sharing among species (Table 7). This is most likely a reflection of a higher *in situ* diversification rate of Suboscines, resulting in the current greater number of deeply divergent lineages. These results are in accord with hypotheses suggested for passerine birds (Baker *et al.* 2009), indicating that New World Suboscines lineages have evolved in South America since at least 40 Mya, while some New World Oscines did not begin to diversify until the Miocene, approximately 20 Mya, although many Oscines lineages only arrived from North America in the last 5 Mya (Vilaça & Santos 2010).

Conclusions

Our findings suggest that the genetic distances between many Neotropical birds are not sufficient to correctly assign individuals to species through a DNA barcoding approach. The results also show that genetic distance values vary between different groups of Neotropical birds, reflecting their peculiar history of colonizing the South American continent. Disregarding errors caused by poor-quality sequences, NuMts, misidentified samples or hybridization issues, we have shown that checking for monophyletic status could improve the barcode identification of Neotropical birds. Approximately 10.4% of bird species were considered nonmonophyletic, some of which were explained by inaccurate taxonomy or recent speciation events (such as peripatric speciation) causing incomplete lineage sorting. This finding indicates that the DNA barcoding approach can be refined through a detailed phylogenetic analysis based on the criterion of monophyly, but some Neotropical taxa will require a full taxonomic review.

Acknowledgements

We thank Eloisa Helena Reis Sari, Frederico Queiroga do Amaral, Guilherme Henrique Silva de Freitas, Leonardo Esteves Lopes, Lilian Mariana Costa, Lucas Aguiar Carrara de Melo, Luciene de Paula Faria, Marcelo de Campos Cordeiro Malta, Miguel Ângelo Marini and Marcos Rodrigues for curating the collections and species identification of most samples deposited in the Centro de Coleções Taxonômicas of UFMG. Thanks also to people from LBEM-UFMG for DNA extraction and help in other laboratory procedures. Additional thanks to Eloisa Helena Reis Sari for helping in research design in the final version of the manuscript. Financial support was provided by CNPq and FAPEMIG, particularly through grants to the Brazilian Barcode of Life (BrBOL) initiative.

References

Agerer R, Ammirati J, Blanz P *et al.* (2000) Always deposit vouchers. *Mycological Research*, **104**, 642–644.

Aleixo A (2007) Conceitos de espécie e o eterno conflito entre continuidade e operacionalidade: uma proposta de normatização de critérios para o reconhecimento de espécies pelo ComitêBrasileiro de Registros Ornitológicos. *Revista Brasileira de Ornitologia*, **15**, 297–310.

Avise JC (2000) *Phylogeography: The History and Formation of Species*. Harvard University Press, Cambridge.

Baker AJ, Tavares ES, Elbourne RF (2009) Countering criticisms of single mitochondrial DNA gene barcoding in birds. *Molecular Ecology Resources*, **9**, 257–268.

Barton NH, Briggs DEG, Eisen JA, Goldstein DB, Patel NH (2010) Phylogenetic reconstruction. In: *Evolution* (eds Barton NH, Briggs DEG, Eisen JA, Goldstein DB, Patel NH). Cold Spring Harbor Laboratory Press, New York.

BirdLife International and NatureServe (2013) *Bird Species Distribution Maps of the World*. BirdLife International, Cambridge, UK and NatureServe, Arlington, TX, USA.

Bock WJ (1969) The origin and radiation of birds. *Annals of the New York Academy of Sciences*, **167**, 147–155.

Bridge PD, Roberts PJ, Spooner BM, Panchal G (2003) On the unreliability of published DNA sequences. *New Phytologist*, **160**, 43–48.

Chaves AV, Clozato CL, Lacerda DR, Sari EHR, Santos FR (2008) Molecular taxonomy of Brazilian tyrant-flycatchers (Passeriformes: Tyrannidae). *Molecular Ecology Resources*, **8**, 1169–1677.

De Queiroz K (2007) Species concepts and species delimitation. *Systematic Biology*, **56**, 879–886.

Donoghue MJ (1985) A critique of the biological species concept and recommendations for a phylogenetic alternative. *Bryologist*, **88**, 172–181.

Ewing B, Green P (1998) Base-calling of automated sequencer traces using Phred. II. Error Probabilities. *Genome Research*, **8**, 186–194.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using Phred. I. Accuracy Assessment. *Genome Research*, **8**, 175–185.
Farris J (1974) Formal definitions of paraphyly and polyphyly. *Systematic Zoology*, **23**, 548–554.
Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, **3**, 294–299.
Funk DJ, Omland KE (2003) Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics*, **34**, 397–423.
Goldstein PZ, DeSalle R (2000) Phylogenetic species, nested hierarchies, and character fixation. *Cladistics*, **16**, 364–384.
Goldstein PZ, DeSalle R, Amato G, Vogler AP (2000) Conservation genetics at the species boundary. *Conservation Biology*, **14**, 120–131.
Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Research*, **8**, 195–202.
Harris JD (2003) Can you bank on GenBank? *Trends in Ecology & Evolution*, **18**, 317–319.
Harrison RG (1991) Molecular changes at speciation. *Annual Review of Ecology and Systematics*, **22**, 281–308.
Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, **270**, 313–321.
Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of Birds through DNA barcodes. *PLoS Biology*, **2**, e312.
Hunter SJ, Goodall TI, Walsh KA, Owen R, Day JC (2008) Nondestructive DNA extraction from blackflies (Diptera: Simuliidae): retaining voucher specimens for DNA barcoding projects. *Molecular Ecology Resources*, **8**, 56–61.
Ivanova NV, Zemlak TS, Hanner RH, Hebert PDN (2007) Universal primer cocktails for fish DNA barcoding. *Molecular Ecology Notes*, **7**, 544–548.
Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO (2012) The global diversity of birds in space and time. *Nature*, **491**, 444–448.
Kerr KCR, Birks SM, Kalyakin MV *et al*. (2009a) Filling the gap - COI barcode resolution in eastern Palearctic birds. *Frontiers in Zoology*, **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**, e4379.
Knowles LL, Carstens BC (2007) Delimiting species without monophyletic gene trees. *Systematic Biology*, **56**, 887–895.
Losos JB, Glor RE (2003) Phylogenetic comparative methods and the geography of speciation. *Trends in Ecology & Evolution*, **18**, 220–227.
Mallet J, Willmott K (2003) Taxonomy: renaissance or Tower of Babel? *Trends in Ecology & Evolution*, **18**, 57–59.
Meier R, Zhang G, Ali F (2008) The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. *Systematic Biology*, **57**, 809–813.
Messing J (1983) New M13 vectors for cloning. *Methods in Enzymology*, **101**, 20–78.
Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology*, **3**, e422.
Milá B, Tavares ES, Smith TB, Baker AJ (2012) A trans-amazonian screening of mtDNA reveals deep intraspecific divergence in forest birds and suggests a vast underestimation of species diversity. *PLoS ONE*, **7**, e40541.
Miller MA, Pfeiffer W, Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: *Gateway Computing Environments Workshop (GCE)*, pp. 1–8. New Orleans, LA.
Mishler BD (1985) The morphological, developmental, and phylogenetic basis of species concepts in bryophytes. *Bryologist*, **88**, 207–214.
Moritz C, Cicero C (2004) DNA barcoding: promise and pitfalls. *PLoS Biology*, **2**, e354.
Neigel JE, Avise JC (1986) Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In: *Evolutionary Processes and Theory* (eds Nevo E, Karlin S), pp. 515–534. Academic Press, New York.
Nielsen R, Matz M (2006) Statistical approaches for DNA barcoding. *Systematic Biology*, **55**, 162–169.
Nylander J (2004) MRMODELTEST *v2. 3. Program Distributed by the Author*. Evolutionary Biology Centre, Uppsala University, Uppsala.
Pollock DD, Zwickl DJ, Mcguire JA, Hillis DM (2002) Increased taxon sampling is advantageous for phylogenetic. *Systematic Biology*, **51**, 664–671.
Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, **53**, 793–808.
Prendini L (2005) Comment on "identifying spiders through DNA barcodes". *Canadian Journal of Zoology*, **83**, 498–504.
Ronquist F, Teslenko M, van der Mark P *et al*. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, **61**, 539–542.
Sambrook J, Russell DW (2001) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, New York.
Sibley CG, Monroe BL Jr (1990) *Distribution and Taxonomy of Birds of the World*. Yale University Press, New Haven.
Sorenson MD (2003) *Avian mtDNA Primers*. http://people.bu.edu/msoren/Bird.mt.Primers.pdf
Swofford D (1998) *PAUP 4.0: Phylogenetic Analysis Using Parsimony*. Smithsonian Institution, Washington.
Tamura K, Peterson D, Peterson N *et al*. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, **28**, 2731–2739.
Tavares ES, Gonçalves P, Miyaki CY, Baker AJ (2011) DNA barcode detects high genetic structure within neotropical bird species. *PLoS ONE*, **6**, e28543.
Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.
Vilaça ST, Santos FR (2010) Biogeographic history of the species complex Basileuterus culicivorus (Aves, Parulidae) in the Neotropics. *Molecular Phylogenetics and Evolution*, **57**, 585–597.
Vilaça ST, Lacerda DR, Sari HER, Santos FR (2006) DNA-based identification to Thamnophilidae (Passeriformes) species: the first barcodes of Neotropical birds. *Revista Brasileira de Ornitologia*, **14**, 7–13.
Vogler AP, Monaghan MT (2007) Recent advances in DNA taxonomy. *Journal of Zoological Systematics and Evolutionary Research*, **45**, 1–10.
Will KW, Rubinoff D (2004) Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics*, **20**, 47–55.
Zwickl DJ, Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology*, **51**, 588–598.

B.R.N.C. performed laboratory procedures, analysed data and wrote the study. A.V.C. contributed with research design and collection of samples, performed laboratory procedures, helped data analysis and contributed with suggestions and revision of study. M.F.V. collected and performed taxonomic identification of most samples, and contributed with suggestions and revision of study A.C.A.N. and J.C. performed laboratory procedures. F.R.S. contributed with research design, supervised laboratory procedures and data analysis, and contributed with suggestions and revision of study.

Data Accessibility

DNA sequences: BOLD project BBB, Accession nos BBB319-13 to BBB421-13; DNA sequences: GenBank Accession nos KM896216-KM896655; Final DNA sequence assembly: online supplementary material; Phylogenetic data: online supplementary material; R Scripts: online supplementary material.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 List of the species and respective numbers of individuals analyzed in this study.

Table S2 Family groupings used in this study for genetic distance calculations and Bayesian tree reconstructions.

Table S3 List of erroneous sequences from GenBank, comprising misidentifications, poor quality sequences, NuMts, and hybridization between species (data accessed in March 2013).

Table S4 List of non-monophyletic species according Bayesian trees built from 452 bp of the 5′ the COI gene.

Table S5 Mean and median Kimura 2-Parameter model (K2P) pairwise genetic distances and average barcoding gap (using the smallest interspecific and larger intraspecific distances for each species) for all bird orders.

Table S6 List of species that were unidentifiable through Kimura 2-Parameter model (K2P) genetic distances due to sharing haplotypes or showing larger intraspecific than interspecific distances (overdivergent species).

Data S1. All Bayesian trees built from 452 bp of the 5′ end of the COI gene.

Data S2. Scripts for R used in the analyses.