Molecular Phylogenetics and Evolution 115 (2017) 1–6

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Short Communication

Striking pseudogenization in avian phylogenetics: Numts are large and common in falcons

Deborah F. Nacer *, Fabio Raposo do Amaral

Departamento de Ecologia e Biologia Evolutiva, Universidade Federal de São Paulo, Rua Professor Artur Riedel, 275, Jardim Eldorado, Diadema, SP CEP 09972-270, Brazil

ARTICLE INFO

Article history:
Received 28 March 2017
Revised 23 June 2017
Accepted 5 July 2017
Available online 6 July 2017

Keywords:
Pseudogenes
Synology
Aves
Falconidae
Falco cherrug
Falco peregrinus

ABSTRACT

Nuclear copies of mitochondrial genes (numts) are a well-known feature of eukaryotic genomes and a concern in systematics, as they can mislead phylogenetic inferences when inadvertently used. Studies on avian numts initially based on the chicken genome suggest that numts may be uncommon and relatively short among birds. Here we ask how common numts are in falcons, based on recently sequenced genomes of the Saker falcon (*Falco cherrug*) and Peregrine falcon (*F. peregrinus*). We identified numts by BLASTN searches and then extracted CYTB, ND2 and COI sequences from them, which were then used for phylogeny inference along with several sequences from other species in Falconiformes. Our results indicate that avian numts may be much more frequent and longer than previously thought. Phylogenetic inferences revealed multiple independent nuclear insertions throughout the history of the Falconiformes, including cases of sequences available in public databases and wrongly identified as authentic mtDNA. New sequencing technologies and ongoing efforts for whole genome sequencing will provide exciting opportunities for avian numt research in the near future.

1. Introduction

Reconstructing the complete tree of life is the ultimate goal of systematics, and mitochondrial DNA (mtDNA) has been an important tool in that endeavor for decades. The recent astonishing increase in the sequencing of nuclear data, along with theoretical and methodological advances in phylogenetic inferences have paved the way to an era of species trees based on hundreds or even thousands of independent loci (Edwards, 2009; Faircloth et al., 2012). Notwithstanding its well-known limitations (Galtier et al., 2009), mtDNA will probably continue to help providing hints about many organisms' histories for years to come, especially because it is easily obtained as a subproduct in next-generation sequencing of nuclear datasets (e.g. Amaral et al., 2015).

Despite being routine lab work around the world, generating a set of mtDNA orthologous sequences may not be always straightforward. Nuclear copies of mitochondrial loci - also known as numts - are a well-known feature of eukaryotic genomes (Richly and Leister, 2004). Numts have been found in many taxonomic groups, and are a concern among systematists since they often retain high sequence similarity to their mitochondrial counterparts, and may mislead barcoding, phylogenetic and phylogeographic inferences when they are regarded as true mitochondrial sequences (Bensasson et al., 2001; Song et al., 2008). When properly identified, numts can be very useful as roots for phylogenies, as genetic markers or even as a source of data from which nuclear mutation rates can be inferred (Bensasson et al., 2001; Leister, 2005). Avian numts in particular seem to be especially prone to amplification when universal primers are used with tissues that are rich in nuclear DNA such as avian blood samples (Sorenson and Quinn, 1998). The sequencing of both the mitochondrial and nuclear genomes of the chicken provided the first opportunity for mining nuclear mtDNA copies in birds (Pereira and Baker, 2004), leading to the idea that avian numts are mostly small and not abundant in avian genomes.

Here we use recently sequenced complete nuclear genomes (Zhan et al., 2013) to ask how common and how large are numts among falcons, a group of birds whose phylogeographic and phylogenetic history has been explored mostly using mtDNA-based datasets (e.g. Bell et al., 2014; Fuchs et al., 2015; Griffiths et al., 2004; White et al., 2013). Contrary to expectations based on the chicken genome, we found that numts in falcons are abundant and can be quite long, including an instance of integration that spans most of the avian mitogenome. We also evaluate the pervasiveness of numts in previous historical inferences within falcons, and discuss strategies for future studies including mtDNA.

* Corresponding author.
E-mail address: dfnacer@gmail.com (D.F. Nacer).

http://dx.doi.org/10.1016/j.ympev.2017.07.002
1055-7903/

2. Material and methods

2.1. Numt identification

We identified candidate sequences that could be numts by performing BLASTN v2.6.0 (Altschul et al., 1997) searches of complete mitochondrial genomes against complete nuclear genomes of the same species in January 13th 2017. Two *Falco* species were used: the Saker falcon (*Falco cherrug*, WGS AKMU01, mitochondrion NC_026715.1) and the Peregrine falcon (*Falco peregrinus*, WGS AKMT01, mitochondrion NC_000878.1). Sequence accession numbers in this article all refer to GenBank (https://www.ncbi.nlm.nih.gov/genbank/). Searches were conducted on the online version of BLASTN using parameters as in Pereira and Baker (2004). Only hits with an Expect Value (e-value) equal to or smaller than 10^{-4} were considered, and no filters (e.g. low complexity) were used. Two or more identical matches were counted as one. In some cases, multiple BLASTN hits were found in the same subject sequence. When hits' sequences overlapped or complemented each other in mitochondrial genes, a single insertion event was considered and those hits were merged into a single candidate.

Candidates had then their identity percentage analyzed before they were considered numts. Hits that had identities >98.3% were discarded as false positives. This threshold was established by calculating pairwise identities between every complete mtDNA from *Falco* species available at the time of this study (*Falco cherrug*, *F. peregrinus*, *F. columbarius*, *F. naumanni*, *F. rusticolus*, *F. sparverius*, and *F. tinnunculus*). The largest identity obtained was 98.3% between *Falco cherrug* and *F. rusticolus*, the two closest species between all seven according to Fuchs et al. (2015). Anything above this probably meant that the subject sequence was actually mtDNA misidentified as nuclear, which lead to those candidates being discarded.

2.2. Numt annotation and phylogenetic inference

All aligned candidates were compared to the original mitogenome for loci identification and we marked all regions possibly present in a candidate. We selected three genes that have been important in evolutionary studies and/or in DNA barcoding to explore the possible impact of pseudogenization in falcon systematics: cytochrome *b* (CYTB), NADH dehydrogenase subunit 2 (ND2), and cytochrome *c* oxidase subunit I (COI). Candidate sequences that possibly included these genes were analyzed using the MITOS WebServer (Bernt et al., 2013) and genes were annotated. Relevant regions were extracted with Geneious R6.1 (Kearse et al., 2012) and then checked for frameshift mutations and aberrant stop codons.

Numt sequences that matched CYTB, ND2 or COI and were longer than 300 bp were included in phylogenetic reconstructions along with sequences available on GenBank from over 65 species of Falconidae. Sequences were downloaded in January 25th 2017 using three different queries: "Falconidae AND cytochrome *b*", "Falconidae AND (NADH dehydrogenase subunit 2 OR ND2)" and "Falconidae AND (cytochrome oxidase subunit 1 OR cytochrome *c* oxidase subunit I OR COI)". Sequences from other orders or not from the intended genes were removed from the resulting fasta. We used *Melopsittacus undulatus* (NC_009134.1) as outgroup based on the sister relationship between falconids and a clade containing passerines and parrots (Hackett et al., 2008). Raw sequences are available on Mendeley Data (doi:http://dx.doi.org/10.17632/4mhd4v4ssc.1).

Gene sequences were aligned using MUSCLE (Edgar, 2004) as implemented in Geneious. The best model of nucleotide evolution for a set of mitochondrial and nuclear genes was determined by jModelTest2 (Darriba et al., 2012; Guindon and Gascuel, 2003) based on the Akaike Information Criterion (Akaike, 1973), as implemented on the CIPRES Science Gateway (Miller et al., 2010). Bayesian phylogenetic inferences were conducted in MrBayes 3.2.6 (Ronquist et al., 2012) based on two independent runs of 15 million generations. Sampling was performed every 1000, and the first 25% trees were discarded as burn in. This was also computed on CIPRES Science Gateway. All resulting trees were visualized and edited on FigTree 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

3. Results

3.1. Numt identification

We identified nuclear copies of mtDNA covering 91.2% of the mitogenome of the Saker falcon and 93.6% of the Peregrine falcon (Fig. 1). The Saker falcon BLASTN search resulted in 115 hits. These hits were arranged into 67 candidates that corresponded to 43 numts (Table 1; Table SM1 in the Supplementary Material) with a mean size of 1135.16 bp ± 2566.11 (standard deviation). They ranged from 37 to 12,950 bp and only eight of them had more than 1000 bp. Numt sequences totaled 48,812 bp in 1,174,811,715 bp, or 0.004% of the nuclear genome.

The Peregrine falcon search resulted in 102 hits arranged into 49 candidates (Table 1; Table SM2 in the Supplementary Material), all of which were considered numts, and had a mean size of 1000.22 bp ± 2387.11 (standard deviation). They ranged from 55 to 12,954 bp and only eight of them had more than 1000 bp. Numt sequences totaled 49,011 bp in 1,171,973,431 bp, or 0.004% of the nuclear genome. Search results were identified as numts for their identity to the original mitochondrial sequence (between 64.5% and 98.0%) and presence in nuclear sequences.

Of all numt candidates, 11 had one or more discontinuities ranging from 65 to 456 bp between same subject hits. The merging of these hits revealed some very long numts in both species (Table 1): 12,950 bp (Saker falcon) and 12,954 bp (Peregrine falcon), which correspond to the insertion of a sequence of 72% the size of the mtDNA of these falcons.

3.2. Numt annotation and phylogenetic inferences

Twenty-three numt candidates from *F. cherrug* and 15 from *F. peregrinus* were annotated and had multiple copies of mitochondrial genes identified. As an example, numt 1 from both species comprised at least partial sequences for all 13 typical vertebrate protein-coding genes and at least 18 tRNAs. In total, we retrieved 15 CYTB, 7 ND2 and 15 COI sequences from those candidates (Table SM3 in the Supplementary Material). Of these sequences, respectively nine, four and eight had more than 300 bp and were used in the phylogenetic reconstructions. All gene sequences extracted from numts presented at least one aberrant stop codon. A total of 242 CYTB, 191 ND2 and 194 COI sequences were used (see Table SM4 in the Supplementary Material for accession numbers). Gene alignments resulted in GTR + I + G for CYTB and ND2, and GTR + G for COI as the best models of nucleotide evolution.

Numts did not form a single monophyletic group in any tree. Instead, multiple nuclear insertions of mitochondrial genes can be seen throughout the history of the Falconiformes (Fig. 2; see Figs. SM1–3 in the Supplementary Material for the complete trees). Integration events may predate the diversification of the genus *Falco* - as it seems to be the case for the longest numts found, Fc 1 and Fp 1 - or be nested in the diversification of the genus.