



## 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.

© 2017 Elsevier Inc. All rights reserved.

## 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](mailto:dfnacer@gmail.com) (D.F. Nacer).

## 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 *Melospittacus 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  $\pm$  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.

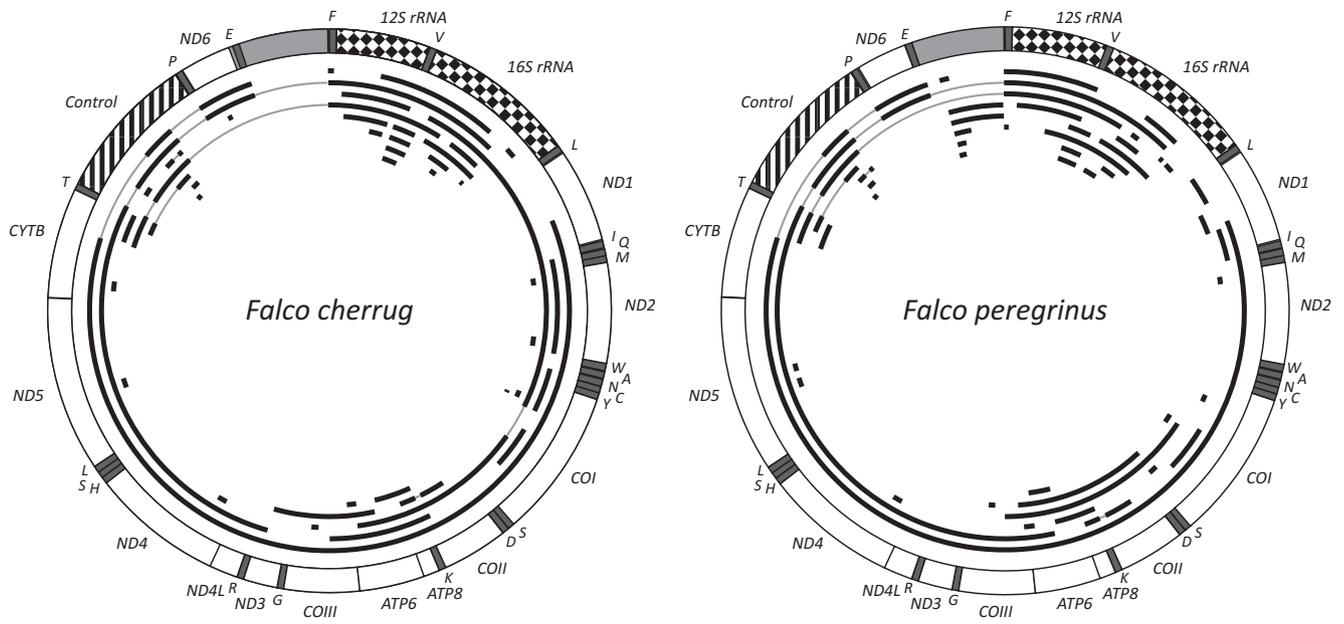
The Peregrine falcon search resulted in 102 hits arranged into 49 candidates (Table 1; Table SM2 in the Supplementary Material), all of which were considered numts, and had a mean size of 1000.22 bp  $\pm$  2387.11 (standard deviation). They ranged from 55 to 12,954 bp and only eight of them had more 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.



**Fig. 1.** Identified numts (shown in black inside the circular mitochondrial DNA) from the Saker falcon (left) and the Peregrine falcon (right). Different BLASTN hits from the same subject are connected by a fine gray line. For complete numt description, see [Tables SM1 and SM2 in the Supplementary Material](#). Acronyms stand for: NADH dehydrogenase subunits 1 to 6 (ND1–6), cytochrome *c* oxidase subunits I to III (COI–COIII), ATP synthase 6 and 8 (ATP6 and 8), and cytochrome *b* (CYTB). Other letters are amino acid abbreviations indicating tRNA. Size and position of mtDNA regions are according to [Lu et al. \(2016\)](#) and [Mindell et al. \(1998\)](#) respectively.

**Table 1**

Four longest numts identified in *Falco cherrug* (Fc) and in *F. peregrinus* (Fp) and mitochondrial regions possibly present in each of them. Sequences from GenBank that contained each numt are indicated between parentheses. Numt hits were sometimes fragmented and numbers between brackets indicate such fragments. Query orientation is always +. O = subject orientation. Lengths are in base pairs. Acronyms stand for: NADH dehydrogenase subunits 1 to 6 (ND1–6), cytochrome *c* oxidase subunits I to III (COI–COIII), ATP synthase 6 and 8 (ATP6 and 8), and cytochrome *b* (CYTB). For complete lists of numts in these species and BLASTN alignment parameters, see [Tables SM1 and SM2 in the Supplementary Material](#).

| Numt                  | Subject start | Subject end | O | Length | Aligned to which mitochondrial region   |
|-----------------------|---------------|-------------|---|--------|---|
| Fc 1 (AKMU01001195.1) | 1212          | 14,161      | + | 12,950 | [1] ND1, tRNA <sup>Ile</sup> , tRNA <sup>Gln</sup> , tRNA <sup>Met</sup> , ND2, tRNA <sup>Trp</sup> , tRNA <sup>Ala</sup> , tRNA <sup>Asn</sup> , tRNA <sup>Cys</sup> , tRNA <sup>Tyr</sup> ; [2] COI, tRNA <sup>Ser</sup> , tRNA <sup>Asp</sup> , COII, tRNA <sup>Lys</sup> , ATP8, ATP6, COIII, tRNA <sup>Gly</sup> , ND3, tRNA <sup>Arg</sup> , ND4L, ND4, tRNA <sup>His</sup> , tRNA <sup>Ser</sup> , tRNA <sup>Leu</sup> , ND5, CYTB; [3] Control, Repeat1; [4] tRNA <sup>Pro</sup> , ND6, tRNA <sup>Glu</sup> ; [5] tRNA <sup>Glu</sup> |
| Fc 2 (AKMU01057450.1) | 45,817        | 36,414      | – | 9404   | [1] ND3, tRNA <sup>Arg</sup> , ND4L, ND4, tRNA <sup>His</sup> , tRNA <sup>Ser</sup> , tRNA <sup>Leu</sup> , ND5, CYTB, tRNA <sup>Thr</sup> ; [2] Control, Repeat1; [3] tRNA <sup>Pro</sup> , ND6, tRNA <sup>Glu</sup> , Non-coding; [4] tRNA <sup>Phe</sup> , 12S rRNA, tRNA <sup>Val</sup> , 16S rRNA  |
| Fc 3 (AKMU01063999.1) | 1             | 7253        | + | 7253   | [1] 16S rRNA, tRNA <sup>Leu</sup> , ND1, tRNA <sup>Ile</sup> , tRNA <sup>Gln</sup> , tRNA <sup>Met</sup> , ND2, tRNA <sup>Trp</sup> , tRNA <sup>Ala</sup> , tRNA <sup>Asn</sup> , tRNA <sup>Cys</sup> , tRNA <sup>Tyr</sup> ; [2] COI, tRNA <sup>Ser</sup> , tRNA <sup>Asp</sup> , COII, tRNA <sup>Lys</sup> , ATP8, ATP6, tRNA <sup>Lys</sup> , ATP8, ATP6   |
| Fc 4 (AKMU01064179.1) | 946           | 4175        | + | 3230   | [1] CYTB, tRNA <sup>Thr</sup> ; [2] Control; [3] tRNA <sup>Phe</sup> , 12S rRNA, tRNA <sup>Val</sup> , 16S rRNA   |
| Fp 1 (AKMT01013886.1) | 1192          | 14,145      | + | 12,954 | [1] ND1, tRNA <sup>Ile</sup> , tRNA <sup>Gln</sup> , tRNA <sup>Met</sup> , ND2, tRNA <sup>Trp</sup> , tRNA <sup>Ala</sup> , tRNA <sup>Asn</sup> , tRNA <sup>Cys</sup> , tRNA <sup>Tyr</sup> ; [2] COI, tRNA <sup>Ser</sup> , tRNA <sup>Asp</sup> , COII, tRNA <sup>Lys</sup> , ATP8, ATP6, COIII, tRNA <sup>Gly</sup> , ND3, tRNA <sup>Arg</sup> , ND4L, ND4, tRNA <sup>His</sup> , tRNA <sup>Ser</sup> , tRNA <sup>Leu</sup> , ND5, CYTB; [3] Control; [4] tRNA <sup>Pro</sup> , ND6, tRNA <sup>Glu</sup> ; [5] tRNA <sup>Glu</sup>          |
| Fp 2 (AKMT01005544.1) | 14,070        | 3275        | – | 10,796 | [1] ATP6, COIII, tRNA <sup>Gly</sup> , ND3, tRNA <sup>Arg</sup> , ND4L, ND4, tRNA <sup>His</sup> , tRNA <sup>Ser</sup> , tRNA <sup>Leu</sup> , ND5, CYTB, tRNA <sup>Thr</sup> ; [2] Control; [3] ND6, tRNA <sup>Glu</sup> ; [4] tRNA <sup>Phe</sup> , 12S rRNA, tRNA <sup>Val</sup> , 16S rRNA  |
| Fp 3 (AKMT01034969.1) | 3750          | 6986        | + | 3237   | [1] CYTB, tRNA <sup>Thr</sup> ; [2] Control; [3] tRNA <sup>Phe</sup> , 12S rRNA, tRNA <sup>Val</sup> , 16S rRNA   |
| Fp 4 (AKMT01024124.1) | 2             | 2836        | + | 2835   | COI, tRNA <sup>Ser</sup> , tRNA <sup>Asp</sup> , COII, tRNA <sup>Lys</sup> , ATP8, ATP6, COIII  |

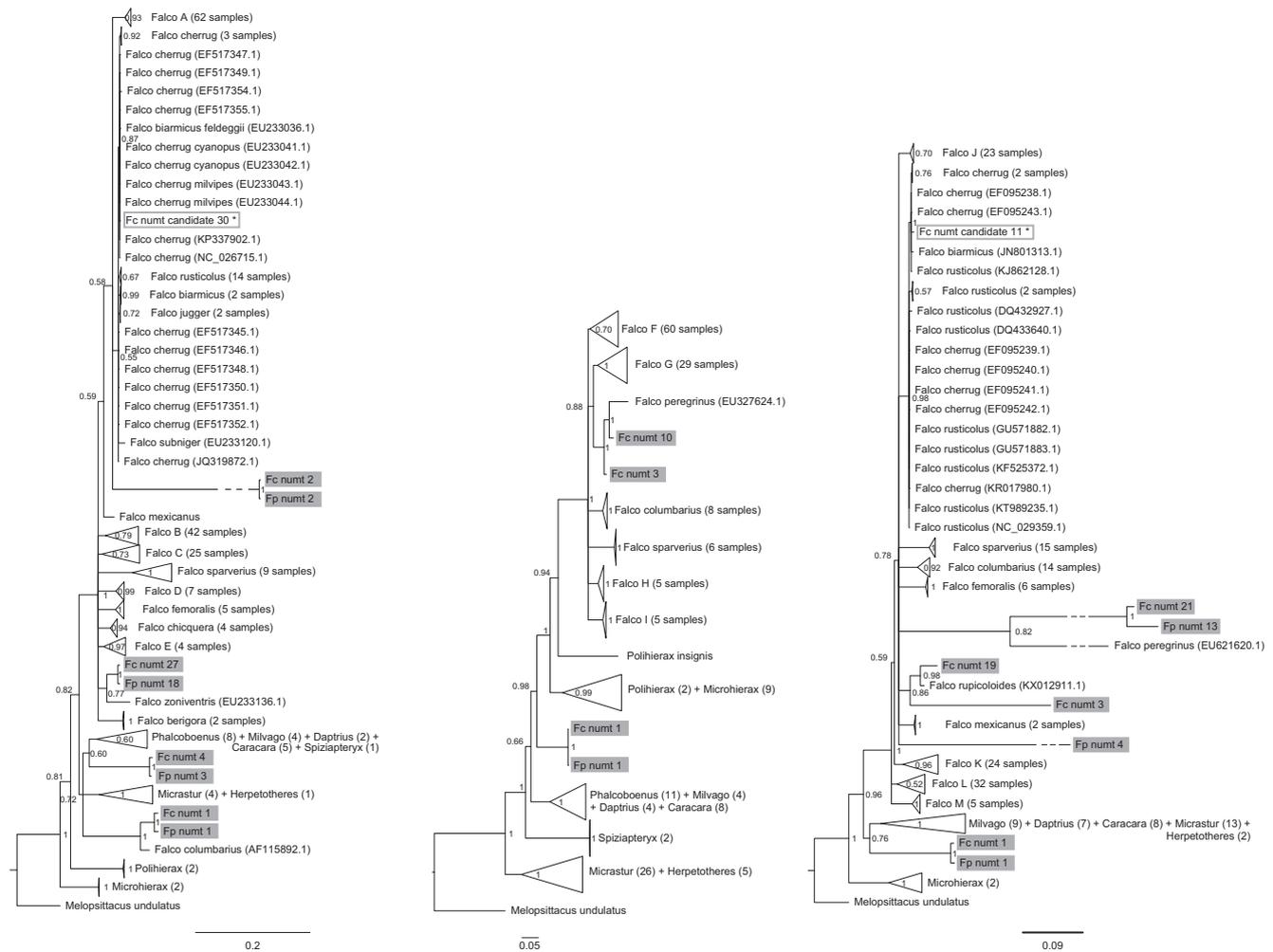
## 4. Discussion

Here we show that numts may be more pervasive among birds than previously thought, as the genomes from Falconiformes had many more nuclear insertions of mitochondrial DNA than the chicken genome - which was the first to be sequenced. We found at least three times more numts and one order of magnitude more numt base pairs in falcons than in the chicken (respectively 43–49 vs. 13 and 0.004% vs. 0.0008%; [Pereira and Baker, 2004](#)). In addition, the number of numts found is larger than in many other vertebrates, even in cases of much larger genomes as in the dog, zebrafish and rat (see [Pereira and Baker, 2004](#); [Richly and Leister,](#)

[2004](#)). Even more strikingly, we report for the first time some very large avian numts, as an instance of five merged fragments covers more than 70% of the mitogenome of this species. Furthermore, numt sequences differed from authentic mtDNA in length and divergence rate, and could be identified within public databases, which included sequences previously used in other phylogenetic studies. We discuss our most important findings in detail below.

### 4.1. Numt evolution and impact on falcon systematics

Despite the low resolution of individual gene tree reconstructions, the few strongly supported nodes (>0.95) suggest that the



**Fig. 2.** Phylogenies based respectively on CYTB, ND2 and COI sequences from Falconidae species and from numts (filled gray rectangles) found in *Falco cherrug* (Fc) and *F. peregrinus* (Fp). Fc numt candidates marked with asterisks (30 and 11) are probably real mitochondrial sequences that have been misidentified as nuclear genome in GenBank. Node numbers represent posterior probability. Numt numbers correspond to Tables SM1 and SM2 in the Supplementary Material. Clades lacking numts were collapsed and the species present in each of them can be seen in Table 2. The original trees are available in the Supplementary Material (Figs. SM1–3).

evolutionary history of falcon numts involves multiple independent integrations, possibly spread over the entire evolution of the

**Table 2**

Clades collapsed in Fig. 2 with their respective species and number of samples.

| Clade | Species (number of samples)  |
|-------|--|
| A     | <i>Falco fasciinucha</i> (3), <i>F. peregrinus</i> (53), <i>F. pelegrioides</i> (6)  |
| B     | <i>Falco tinnunculus</i> (18), <i>F. naumanni</i> (8), <i>F. cenchroides</i> (3), <i>F. rupicolus</i> (5), <i>F. newtoni</i> (1), <i>F. araeus</i> (1), <i>F. rupicoloides</i> (3), <i>F. punctatus</i> (3)  |
| C     | <i>Falco cuvierii</i> (2), <i>F. subbuteo</i> (3), <i>F. concolor</i> (2), <i>F. eleonorae</i> (7), <i>F. longipennis</i> (3), <i>F. deiroleucus</i> (1), <i>F. columbarius</i> (7)  |
| D     | <i>Falco vespertinus</i> (5), <i>F. amurensis</i> (2)  |
| E     | <i>Falco dickinsoni</i> (2), <i>F. novaeseelandiae</i> (2)   |
| F     | <i>Falco peregrinus</i> (13), <i>F. fasciinucha</i> (1), <i>F. cherrug</i> (4), <i>F. rusticolus</i> (4), <i>F. biarmicus</i> (1), <i>F. subniger</i> (3), <i>F. jagger</i> (1), <i>F. hypoleucos</i> (1), <i>F. mexicanus</i> (2), <i>F. chicquera</i> (1), <i>F. femoralis</i> (2), <i>F. novaeseelandiae</i> (1), <i>F. subbuteo</i> (5), <i>F. cuvierii</i> (1), <i>F. concolor</i> (3), <i>F. eleonorae</i> (1), <i>F. severus</i> (1), <i>F. longipennis</i> (6), <i>F. rufifigularis</i> (4), <i>F. deiroleucus</i> (2), <i>F. berigora</i> (3) |
| G     | <i>Falco naumanni</i> (4), <i>F. tinnunculus</i> (5), <i>F. cenchroides</i> (6), <i>F. moluccensis</i> (1), <i>F. rupicolus</i> (3), <i>F. newtoni</i> (2), <i>F. araeus</i> (1), <i>F. punctatus</i> (3), <i>F. rupicoloides</i> (2), <i>F. alopec</i> (1), <i>F. zoniventris</i> (1)   |
| H     | <i>Falco ardosiaceus</i> (3), <i>F. dickinsoni</i> (2)   |
| I     | <i>Falco amurensis</i> (3), <i>F. vespertinus</i> (2)  |
| J     | <i>Falco peregrinus</i> (22), <i>F. fasciinucha</i> (1)  |
| K     | <i>Falco subbuteo</i> (17), <i>F. eleonorae</i> (1), <i>F. longipennis</i> (1), <i>F. rufifigularis</i> (5)  |
| L     | <i>Falco tinnunculus</i> (26), <i>F. naumanni</i> (5), <i>F. cenchroides</i> (1)   |
| M     | <i>Falco vespertinus</i> (3), <i>F. amurensis</i> (2)  |

family. If poorly supported nodes are considered, ancient integrations are exemplified by clades branching at the base of the Falconidae tree, prior to the diversification within *Falco* (e.g. numts Fc 1, Fp 1, Fc 4, Fp 3), and there are also very recent numts nested within *Falco* (e.g. numts Fc 3, Fc 10). Thus, some of those nuclear copies may be readily identified in routine phylogenetic studies by their high divergence from authentic mtDNA, while other more recent integrations are possibly the most difficult to identify. In fact, recently derived numt sequences identified as mtDNA are present in public databases, including CYTB (AF115892, *Falco columbarius*; EU233136, *Falco zoniventris*), ND2 (EU327624, *Falco peregrinus*) and COI (EU621620, *Falco peregrinus*; KX012911, *Falco rupicoloides*).

Furthermore, some numts are shared between the two falcon species studied here and appear as sisters in the phylogenies (e.g. numt 1). They likely represent a single integration that occurred at the base of those clades - and, in this case, it is reasonable to predict that they can also be found in other members of the clade. The impact of numts in previous phylogenetic studies, however, seems to be surprisingly low despite the abundance of numts in the falconid genomes: based on the phylogenetic inferences, we identified only one published study that included numts - that of Wright et al. (2008), which used a falcon numt to root a Psittacidae phylogeny.

#### 4.2. Numts as possible witnesses of mitochondrial rearrangements

The mitochondrial gene order varies extensively among birds, and repeated rearrangements may occur even among closely related species and genera (Eberhard and Wright, 2016). The possible long timescale of integrations among falcons may provide hints about the evolution of gene order in the Falconiformes. Mitogenomes of *Falco* species are well known for having a second, degenerated control region (Mindell et al., 1998) that totals, for example, 780 bp in the Peregrine falcon (NC\_000878). Forest-falcons (*Micrastur gilvicolis*, NC\_008548.1), however, are sister to the rest of the Falconiformes radiation (Fig. 2), and show a much shorter number of repetitions in that region (58 bp, ~14 CAAA repetitions). The second longest numt identified in this study (Fc 2/Fp 2) does not include that second control region (CR). Thus, numts could be witnesses to understanding mitogenome rearrangements in this group. Future analyses including numts and complete mitogenomes may help in reconstructing the origin of the CR duplication among falconids.

#### 4.3. Conclusions

Here we found that avian numts may be common and very long - they can cover most of the mitogenome - contrary to previous evidence from the chicken genome. We found numts of varying sizes and levels of divergence that represent opportunities for PCR co-amplification using universal and/or more specific primers of both small and large fragments. The use of off-target sequencing to assemble mitogenomes (e.g. Amaral et al., 2015) is becoming popular, but little is known about how commonly numts are captured using such approach. Based on our results, we conclude that comparisons between putative mtDNA sequences and published nuclear genomes are especially important for future studies of avian systematics and phylogeography, and should be done as soon as data becomes available. New sequencing technologies and the ongoing efforts for whole genome sequencing provide great opportunities for numt discovery, and the recent accumulation of 48 complete avian nuclear genomes (Zhang et al., 2014) offer an exciting opportunity to identify and trace the history of major events of nuclear insertion of mitochondrial fragments in the main branches of the avian tree of life.

#### Acknowledgments

This work was supported by FAPESP and the Biota Program [FRA: 2011/50143-7 and 2011/23155-4], a joint NSF/Biota FAPESP/NASA Dimensions of Biodiversity Grant [FRA: FAPESP 2013/50297-0 and DOB 1343578], CNPq/MCT and CAPES. We thank editor Scott V. Edwards and three anonymous reviewers for their excellent comments and suggestions.

#### Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.ympev.2017.07.002>.

#### References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3403.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. *Int. Symp. Inf. theory*, 267–281.

Amaral, F.R., Neves, L.G., Resende Jr, M.F.R., Mobili, F., Miyaki, C.Y., Pellegrino, K.C.M., Biondo, C., 2015. Ultraconserved elements sequencing as a low-cost source of complete mitochondrial genomes and microsatellite markers in non-model amniotes. *PLoS ONE* 10, e0138446.

Bell, D.A., Griffiths, C.S., Caballero, I.C., Hartley, R.R., Lawson, R.H., 2014. Genetic evidence for global dispersal in the Peregrine Falcon (*Falco peregrinus*) and affinity with the Taita Falcon (*Falco fasciinuca*). *J. Raptor Res.* 48, 44–53.

Bensasson, D., Zhang, D., Hartl, D.L., Hewitt, G.M., 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol. Evol.* 16, 314–321.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsche, G., Pütz, J., Middendorf, M., Stadler, P.F., 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol. Phylogenet. Evol.* 69, 313–319.

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods.* <http://dx.doi.org/10.1038/nmeth.2109>.

Eberhard, J.R., Wright, T.F., 2016. Rearrangement and evolution of mitochondrial genomes in parrots. *Mol. Phylogenet. Evol.* 94, 34–46.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res* 32, 1792–1797.

Edwards, S.V., 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63, 1–19.

Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* 61, 717–726.

Fuchs, J., Johnson, J.A., Mindell, D.P., 2015. Rapid diversification of falcons (Aves: Falconidae) due to expansion of open habitats in the Late Miocene. *Mol. Phylogenet. Evol.* 82 (Part A), 166–182.

Galtier, N., Nabholz, B., Glémin, S., Hurst, G.D.D., 2009. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Mol. Ecol.* 18, 4541–4550.

Griffiths, C.S., Barrowclough, G.F., Groth, J.G., Mertz, L., 2004. Phylogeny of the Falconidae (Aves): a comparison of the efficacy of morphological, mitochondrial, and nuclear data. *Mol. Phylogenet. Evol.* 32, 101–109.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704.

Hackett, S.J., Kimball, R.T., Reddy, S., Bowie, R.C.K., Braun, E.L., Braun, M.J., Chojnowski, J.L., Cox, W.A., Han, K.-L., Harshman, J., Huddleston, C.J., Marks, B. D., Miglia, K.J., Moore, W.S., Sheldon, F.H., Steadman, D.W., Witt, C.C., Yuri, T., 2008. A phylogenomic study of birds reveals their evolutionary history. *Science* 320, 1763–1768 (80).

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649.

Leister, D., 2005. Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends Genet.* 21, 655–663.

Lu, J., Lu, J., Li, X.-F., Jiang, H., 2016. Complete mitochondrial genome of the Saker falcon, *Falco cherrug* (Falco, Falconidae). *Mitochondrial DNA Part A* 27, 3226–3227.

Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: 2010 Gateway Computing Environments Workshop, GCE 2010.

Mindell, D.P., Sorenson, M.D., Dimcheff, D.E., 1998. Multiple independent origins of mitochondrial gene order in birds. *Proc. Natl. Acad. Sci.* 95, 10693–10697.

Pereira, S.L., Baker, A.J., 2004. Low number of mitochondrial pseudogenes in the chicken (*Gallus gallus*) nuclear genome: implications for molecular inference of population history and phylogenetics. *BMC Evol. Biol.* 4, 17.

Richly, E., Leister, D., 2004. NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.* 21, 1081–1084.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542.

Song, H., Buhay, J.E., Whiting, M.F., Crandall, K.A., 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proc. Natl. Acad. Sci.* 105, 13486–13491. <http://dx.doi.org/10.1073/pnas.0803076105>.

Sorenson, M.D., Quinn, T.W., 1998. Numts: a challenge for avian systematics and population biology. *Auk* 115, 214–221.

White, C.M., Sonsthagen, S.A., Sage, G.K., Anderson, C., Talbot, S.L., 2013. Genetic relationships among some subspecies of the Peregrine Falcon (*Falco peregrinus* L.), inferred from mitochondrial DNA control-region sequences. *Auk* 130, 78–87.

Wright, T.F., Schirtzinger, E.E., Matsumoto, T., Eberhard, J.R., Graves, G.R., Sanchez, J. J., Capelli, S., Müller, H., Scharpegge, J., Chambers, G.K., Fleischer, R.C., 2008. A multilocus molecular phylogeny of the parrots (Psittaciformes): support for a Gondwanan origin during the Cretaceous. *Mol. Biol. Evol.* 25, 2141–2156.

Zhan, X., Pan, S., Wang, J., Dixon, A., He, J., Muller, M.G., Ni, P., Hu, L., Liu, Y., Hou, H., Chen, Y., Xia, J., Luo, Q., Xu, P., Chen, Y., Liao, S., Cao, C., Gao, S., Wang, Z., Yue, Z., Li, G., Yin, Y., Fox, N.C., Wang, J., Bruford, M.W., 2013. Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nat. Genet.* 45, 563–566.

Zhang, G., Li, C., Li, Q., Li, B., Larkin, D.M., Lee, C., Storz, J.F., Antunes, A., Greenwold, M.J., Meredith, R.W., Ödeen, A., Cui, J., Zhou, Q., Xu, L., Pan, H., Wang, Z., Jin, L., Zhang, P., Hu, H., Yang, W., Hu, J., Xiao, J., Yang, Z., Liu, Y., Xie, Q., Yu, H., Lian, J., Wen, P., Zhang, F., Li, H., Zeng, Y., Xiong, Z., Liu, S., Zhou, L., Huang, Z., An, N., Wang, J., Zheng, Q., Xiong, Y., Wang, G., Wang, B., Wang, J., Fan, Y., da Fonseca, R. R., Alfaro-Núñez, A., Schubert, M., Orlando, L., Mourier, T., Howard, J.T., Ganapathy, G., Pfenning, A., Whitney, O., Rivas, M.V., Hara, E., Smith, J., Farré, M., Narayan, J., Slavov, G., Romanov, M.N., Borges, R., Machado, J.P., Khan, I., Springer, M.S., Gatesy, J., Hoffmann, F.G., Opazo, J.C., Håstad, O., Sawyer, R.H.,

Kim, K.-W., Kim, H.J., Cho, S., Li, N., Huang, Y., Bruford, M.W., Zhan, X., Dixon, A., Bertelsen, M.F., Derryberry, E., Warren, W., Wilson, R.K., Li, S., Ray, D. A., Green, R.E., O'Brien, S.J., Griffin, D., Johnson, W.E., Haussler, D., Ryder, O.A., Willerslev, E., Graves, G.R., Alström, P., Fjeldså, J., Mindell, D.P., Edwards, S.V.,

Braun, E.L., Rahbek, C., Burt, D.W., Houde, P., Zhang, Y., Yang, H., Wang, J., Jarvis, E.D., Gilbert, M.T.P., Wang, J., . Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* 346, 1311–1320 (80).