Molecular Phylogenetics and Evolution 181 (2023) 107722

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Bringing to light nuclear-mitochondrial insertions in the genomes of nocturnal predatory birds

Miguel Baltazar-Soares [a,*], Patrik Karell [b,c], Dominic Wright [d], Jan-Åke Nilsson [e], Jon E. Brommer [a]

[a] Department of Biology, University of Turku, Turku 20500, Finland
[b] Bioeconomy Research Team, Novia University of Applied Sciences, Raseborgsvägen 9, FI-10600 Raseborg, Finland
[c] Evolutionary Ecology Unit, Department of Biology, Lund University, Sölvegatan 39 (Ecology Building), SE-223 62 Lund, Sweden
[d] IFM Biology, Linköping University, Linköping 58183, Sweden
[e] Department of Biology, Section of Evolutionary Ecology, Lund University, Ecology Building, 223 62 Lund, Sweden

ARTICLE INFO

Keywords:
NUMT
Strigiformes
Comparative genomics
Mitochondrial genome

ABSTRACT

Mito-nuclear insertions, or *NUMTs*, relate to genetic material of mitochondrial origin that have been transferred to the nuclear DNA molecule. The increasing amounts of genomic data currently being produced presents an opportunity to investigate this type of patterns in genome evolution of non-model organisms. Identifying *NUMTs* across a range of closely related taxa allows one to generalize patterns of insertion and maintenance in autosomes, which is ultimately relevant to the understanding of genome biology and evolution. Here we collected existing pairwise genome-mitogenome data of the order Strigiformes, a group that includes all the nocturnal bird predators. We identified *NUMTs* by applying percent similarity thresholds after blasting mitochondrial genomes against nuclear genome assemblies. We identified *NUMTs* in all genomes with numbers ranging from 4 in *Bubo bubo* to 24 in *Ciccaba nigrolineata*. Statistical analyses revealed *NUMT* size to negatively correlate with *NUMT*'s sequence similarity to with original mtDNA region. Lastly, characterizing these nuclear insertions of mitochondrial origin in a comparative genomics framework produced variable phylogenetic patterns, suggesting in some cases that insertions might pre-date speciation events within Strigiformes.

1. Introduction

For decades, mitochondrial analysis has been considered the molecular tool of choice for phylogeography and historic demographic inferences of natural populations (Ballard & Pichaud, 2014). Views over mitochondrial evolution have been recently challenged by new perspectives, largely supporting its role as a key player in adaptive evolution (Ballard & Pichaud, 2014). In eukaryotes, the mitogenome is a circular molecule whose size varies between 14 k and 20 k base pairs. It comprises 37 highly conserved genes that encode for 13 proteins, 22 transfer RNAs (tRNAs) and 2 ribosomal RNAs (rRNA) and a control region (CR) (Formenti et al., 2021). The mitochondrial genome however, differs structurally across taxa through order re-arrangements, duplications, and single nucleotide polymorphisms (Formenti et al., 2021). Mitochondriás genetic content is not restricted to the cytoplasm, as the occurrence of mitochondrial inserts into the nuclear genome, Nuclear DNA of MiTochondrial origin or *NUMTs*, have been extensively reported in eukaryotes (Formenti et al., 2021; Hazkani-Covo, Zeller, & Martin, 2010; Kleine, Maier, & Leister, 2009; Richly & Leister, 2004). *NUMTs* are thought to originate either during DNA replication or when V(D)J recombination occurs in lymphocyte early development via non-homologous end joining at double-strand breaks during DNA repair processes (Gaziev & Shaikhaev, 2010; Ricchetti, Fairhead, & Dujon, 1999; Schatz & Swanson, 2011). These insertions might attain sizes up to that of a full mitogenome of 14 k base-pairs, allegedly reflecting a co-evolutionary process intrinsically connected to the incorporation of the organelle into eukaryotic cells. The major evidence for co-evolution is the fact that main protein complexes involved in the OXPHOS pathway are encoded both by nuclear and mitochondrial subunits (Puertas & González-Sánchez, 2020). Nevertheless, most of the nuclear insertions of mitochondrial origin are commonly characterized as pseudogenes.

The generation of *NUMTs* is presumably ongoing and suspected to be shaping the size and functionality of genomic regions, though these assumptions are based on observations gathered from a handful of

* Corresponding author.
E-mail address: miguelalexsoares@gmail.com (M. Baltazar-Soares).

https://doi.org/10.1016/j.ympev.2023.107722
Received 30 September 2022; Received in revised form 24 January 2023; Accepted 24 January 2023
Available online 30 January 2023

model species genomes (Hazkani-Covo et al., 2010; Kleine et al., 2009; Puertas & González-Sánchez, 2020). Expanding the characterization of nuclear insertions of mitochondrial origin to as many taxa as possible is critical to understand whether patterns of *NUMT* evolution can be generalized across eukaryote genomes. For instances, It might assist in the detection of recombination and DNA loss, given that the detection of NUMTs is intrinsically related to whether an identifiable sequence is maintained in the genome (Richly & Leister, 2004). Exploring *NUMT* characterization in a comparative genomics framework might also clarify whether the frequency of DNA transfer from mitochondria to the nucleus is a taxonomic characteristic and thus a genomes evolutionary trait (Richly & Leister, 2004). Additionally, and given the extensive utilization of single-mitochondrial genes to construct phylogeographic patterns and more recently in NGS-based DNA barcoding, identifying *NUMTs* is essential to build databases free of these mitochondrial-like fragments (Bertheau, Schuler, Krumboeck, Arthofer, & Stauffer, 2011; Jordal & Kambestad, 2014). Dedicated methodologies to detect *NUMTs* consist of the isolated sequencing of mito- and nuclear genome, as performed by Bertheau et al (2011), or utilizing mapping/aligning algorithms to bioinformatically screen available genome assemblies using mitogenomes as a template and deduce *NUMTs* as a function of sequence similarity percentages. The later strategy allows one to expand *NUMT*-screening to entire phylogenies while utilizing the wealth of genomic resources currently made available with next-generation sequencing (Liang, Wang, Li, Kimball, & Braun, 2018).

Identification of *NUMTs* in avian genomes had its onset after the sequencing of the chicken genome (Pereira & Baker, 2004). Recent comparative genomics approaches – making use of the availability of avian genomes – has extended the search across entire phylogenies and revealed a high diversity in terms of size and numbers (Liang et al., 2018; Nacer and do Amaral, 2017). Here we explore the occurrence of NUMTs in the order *Strigiformes*, which encompasses all the emblematic nocturnal birds of prey commonly known as owls. Its phylogeny is apparently well resolved (Salter et al., 2020), several genomes are available for organisms of this clade but mito-nuclear insertions are yet to be categorized within it. The aim of this study is therefore to improve the current knowledge on mito-nuclear insertions among *Strigiformes*, while investigating aspects related to their detection and evolution. Namely, we explored the relationship between *NUMT* size and detection probability, as well as characterized *NUMT* frequency and distribution within closely related species (Strigiformes). We utilized Pacific Biosciences (PacBio) subreads obtained from a whole-genome sequencing experiment of a tawny owl (*Strix aluco*) individual to assemble a more complete version of this species' mitochondrial genome. Then, we collected available genomes and mitogenomes to interpret the presence of *NUMTs* within Strigiformes.

2. Methods

2.1. Assembly mitochondrial genome from PacBio reads

Data parsing and curation was performed with customized bash scripts unless stated otherwise. Plots and statistical analyses were performed in R (Core 2021). We first re-assembled the tawny owl mitogenome by blasting a dataset of PacBio subreads obtained with continuous long-read (CLR) technology of a tawny owl male individual (see Table S1 for subread information) against complete or nearly complete, mitochondrial genomes of owl species' that are closely related to the tawny owl. Specifically, we downloaded mitochondrial genome of the Spotted owl (*Strix occidentalis caurina*), the Barred owl (*Strix varia*), the Ural owl (*Strix uralensis*), the barn owl (Tyto alba) from the family Tytonidae, and we also included the available partial mitogenome of the tawny owl (*Strix aluco*) (Table S2). This task was performed in *blasr* (Chaisson & Tesler, 2012) with the objective of identifying subreads matching mitochondrial DNA with default max output score of −200 and a percent similarity corresponding to at least 75 %. Identified reads were retrieved with *subset* function in *seqkt* (Li, 2012) utilizing scripts provided by Kovar et al, 2018 (Kovar et al., 2018). The next step included assembly and circularization of the mitochondrial genome. For that purpose, we utilized Trycycler (Wick et al., 2021), which is a pipeline developed to assemble circular genomes from long-read data by combining multiple assemblies of the same dataset (Wick et al., 2021). Within Trycycler, we performed 5 assemblies using Canu 2.1.1 (Koren et al., 2017), which were fed into the Trycycler's pipeline. Canu ran with default parameters and assemblies were polished with PacBio's® tools *pbmm2* and *GCpp VERSION*. The new mitochondrial assembly was annotated with Mitos2 server (Donath et al., 2019) and manually curated by blasting unannotated regions to NCBI database in order to confirm that the circularization of the molecule did not created artificial repeated regions. This new assembly was further validated by comparing percentage of similarity to genome and NCBI *S. aluco* mitogenome and inspecting e-values or quality scores of tRNA, rRNA and mitochondrial genes provided by Mitos2 upon annotation (Donath et al., 2019).

2.2. Exploring NUMT's copy number variation in available Strigiformes nuclear genomes

In order to search for *NUMT's* copy number variation across Strigiformes phylogeny, we collected genome/mitogenome pairs currently available for this order. Both the nuclear genome and the mitochondrial genome are available for the barn owl, the northern spotted owl, the Eurasian eagle owl (*Bubo bubo*), the oriental scops-owl (*Otus stunia*), the ferruginous pigmy owl (*Glaucidium brasilianum*), and the black-and-white owl (*Strix nigrolineata*, formerly known as *Ciccaba nigrolineata*) (Table S2). We also utilized nuclear and the mitochondrial genome of the tawny owl *S. aluco*, even though its nuclear genome is yet unpublished. Our strategy consisted in identifying *NUMTs* with *blasr* – by utilizing the genomes as the reference and mitogenomes as the query – which outputs alignment scores, percent of identity, as well as the start and the end (coordinates) of query matches with a maximum score of −200. To devise filters for false positives, we adopted a strategy similar to the one applied by Nacer and do Amaral (2017), which considered the percent of similarity of the entire mitochondrial DNA molecule within the genus *Falco*. Here we calculated average percent of similarity within the *Strigidae*. Because only one representative of the family Tytonidae (*Tyto alba*) has both a nuclear and a mito- genome available, we applied the same threshold to define *NUMTs* in the single representative of the family. This allowed us to accommodate thresholds based on evolutionary constraints specific to the family. In addition, and to further accommodate the possibility that mitogenomes were not assembled in single contigs corresponding to the molecule, we further filtered out presumptive nuclear contigs with high number of *NUMTs* (>3) (Hazkani-Covo et al., 2010). The overall objective of utilizing a phylogenetic threshold (consequently filtering out sequences with high percent of similarity) as well as removing contigs with high number of putative *NUMTs* from downstream phylogenetic analyses was to avoid the mistake of identifying unassembled or partially assembled mitogenomic sequences as *NUMT* containing nuclear regions. We acknowledge the chosen strategy to be conservative but still robust.

In order to extract the sequence corresponding to a *NUMT* from assembled nuclear genomes, we utilized the functions *seq*, *subseq* and *grep* from *seqkit* (Shen, Le, Li, & Hu, 2016). *Seqkit* is a toolkit to manipulate fasta/q files. Here we utilized the coordinates identified with *blasr* in the previous step as an input to extract genomic regions corresponding to a *NUMT* from each respective genome assembly. To explore whether the probability to detect a *NUMT* relates to fragment length, taxonomy or from which mitochondrial gene it originated from, we tested what could explain percent of similarity with the following linear model: *aov(% similarity ~ length + mtDNAregion + species)* and corrected for the use of sequential sum squares and model selection with *drop1 (NUMT, ~., test = "Chisq")*. Lastly, we performed a Pearsońs correlation

to explore whether larger *NUMTs* possess high similarity with the mitochondrial region of origin, which could be expected if de-functionalization would not occur.

2.3. Phylogenetic relationships of nuclear insertions of mitochondrial origin

We also collected complete mitochondrial genomes of diurnal birds of prey to serve as outgroup to contextualize *NUMT* evolution. Specifically, the Golden eagle (*Aquila chrysaetos*), the Northern goshawk (*Accipiter gentilis*), the white-tailed eagle (*Haliaeetus albicilla*), and the common buzzard (*Buteo buteo*) of the order Acciptriformes, and the Eurasian hobby (*Falco subbuteo*) and the Peregrine falcon (*Falco peregrinus*) of the order Falconiforms. All phylogenetic relationships were explored in MEGA7 (Kumar, Stecher, & Tamura, 2016). First, we aligned all the mitochondrial genomes to build a Maximum Likelihood tree (ML). We estimated nucleotide substitution model, which was then inserted as a parameter in the ML tree construction. The bootstrap consensus was inferred with 1000 replicates, and trees to initiate the algorithm were obtained aby applying Neighbor-Join and BioNJ algorithms, utilizing all sites, prior to selecting the topology with superior log likelihood value. The final ML tree was estimated based ona nucleotide alignment where nucleotide positions with more than 30 % missing data were deleted, and rooted at the divergence between Falconiformes and the all other birds of prey (*F. pererginus* and *F. subbuteo* as outgroup). We also constructed the phylogeny of specific mitochondrial regions where *NUMTs* have originated, by utilizing both the original mitochondrial gene and the respective *NUMT*.

3. Results

3.1. Mitochondrial DNA assembly and annotation

A total of 836 reads were collected via *blasr* of PacBio subreads against the mitogenomes of 5 owl species. The assembly of mitochondrial genome from PacBio subreads revealed three mitochondrial genes that were absent in previous version of the mitogenome of the Tawny owl, transfer RNA-Proline (*trnP*), transfer RNA-Glutamate (*trnE*) and the NADH-ubiquinone oxidoreductase chain 6 (*nad6*), which are located downstream the control region. We reported two putative duplicated regions, namely the gene encoding the *nad3*, and a portion of the *dloop* (Table S3). While the quality score and e-value reported for these regions suggests those being true copies, we observed that all mitochondrial genomes collected from NCBI and re-annotated revealed duplications of the exact same regions.

3.2. Identification of nuclear-inserted mitochondrial fragments

Mitogenomés similarity range within the order Strigidae varied between 80.1 % for the *Otus sunia-Strix occidentalis caurina* comparison and 91.8 % of the *Strix aluco- Strix uralensis* comparison, resulting in an average of 84.5 %(SD=±3.3 %). The original NUMT list included 327 candidates among which 205 were removed due to similarity higher than the calculated phylogenetic threshold and 47 due to over-representation in a single contig, resulting in putative 75 NUMTs. Among these, we report extensive copying of *nad2*, specifically 3 copies in both *Otus sunia* and *Strix aluco* genome, 2 copies in *Glaucidium brasilianum* and 5 copies in *Ciccaba nigrolineata*. The high number of *NUMTs* in *Ciccaba nigrolineata* genome relate nevertheless to several copies of the same mitochondrial genes: 4 copies of *cyt-b*, 3 copies of *cox1*, 2 copies of *nad1*, 3 copies of *nad4*, 4 of *nad5* and 3 of the large subunits of the mitochondrial ribosome. In addition, we also found 2 copies of *nad4* in *Otus sunia* and *Tyto alba,* 2 copies of *cyt-b* in *Strix aluco* (Table 1). All other mito-nuclear inserts were single copies. Insertion sizes varied between tens and thousands of base pairs. There was a substantially higher representativity of protein-coding genes among *NUMTs* (n = 69) than tRNA (n = 4), especially considering that mitochondrial genomes comprise almost 2-fold higher number of tRNA-coding than protein-coding regions (Table S5). The full list of candidates NUMTs and respective distribution across species is presented in Table S5. We found mtDNA region to be the most relevant factor amongst the ones analyzed (after model selection) in explaining similarity percentage (*% similarity ~ mtDNAregion*; F = 2,31; df = 20, p = 0.006) (Table S4). Lastly, we found a negative correlation between *NUMT*s fragment size and similarity percentage (Pearsońs correlation = -0.32, p = 0.001), supporting the hypothesis that large *NUMTs* tend to become unrecognizable after being subjected to recombination or the accumulation of mutations – both of which are suggestive of pseudogenization in the nuclear genome.

Table 1
Progression of the number of candidates *NUMTs* across filtering steps.

Species	Original list	After Phylogenetic threshold	After Overrepresented contig removal
Bubo bubo	65	4	4
Otus sunia	57	32	10
Strix occidentalis	57	5	5
Tyto alba	17	12	6
Glaucidium brasilianum	18	7	7
Ciccaba nigrolineata	81	44	28
Strix aluco	32	18	11

Original list corresponds to all the hits reported by *blasr*; Phylogenetic threshold was set to 84.5% similarity for Strigiformes (all species except *Tyto alba*). A full list of NUMTs and associated mitochondrial region – including copies – can be found in Table S5.

3.3. Comparing mitochondrial phylogenies: Full molecule, single genes and NUMTs.

Full mitochondrial phylogeny of available mitochondrial genomes served as a template to contextualize phylogenetic trees of individual mitochondrial regions and specific *NUMTs*. We captured putative *NUMTs* resembling *cox1*, *cyt-b*, *nad6* and *nad2* in 4 owl species. Mitochondrial gene-specific phylogenies stood in general agreement with the one built with the full molecule and with the currently accepted phylogenetic relationships between nocturnal and diurnal birds of prey. Embedding *NUMTs* with the original mtDNA gene phylogeny revealed a pattern where *NUMTs* rarely cluster with the respective mtDNA genes, suggestive of evolutionary histories distinct from their mitochondrial counterparts (Fig. 1). Specifically, gene-specific phylogenies showed *NUMT*s to generally cluster in relatively younger branches of each respective phylogenetic tree. This is observable in the phylogenies of *nad6*, *cyt-b*, *cox2*, and *NUMT* cluster A of *nad2*. The *NUMT* cluster B on *nad2*'s phylogenetic tree – composed of 4 copies found on the *C. nigrolineata* genome, 1 on *O. sunia*, and 1 on *G. brasilianum* – shows a different pattern, as its divergence pre-dated the Strigiformes diversification. (Fig. 2).

4. Discussion

High throughput sequencing facilitates the generation of large amounts of genomic data, from which a multitude of information can be mined to better understand genome and organismal evolution. Here we have mined PacBio reads, obtained for a genome assembly that is currently at a draft stage, with the objective of searching for mitochondrial insertions in the tawny owl nuclear genome. We first report the extension of the tawny owĺs mitogenome assembly with the inclusion of three previously undetected tRNAs and the mtDNA-*nad6*. We further discovered a diversified scenario of *NUMT* insertions in the *Strigidae* phylogeny with numbers of putative *NUMTs* varying among species, several copies of *nad2* present in almost all genomes, striking

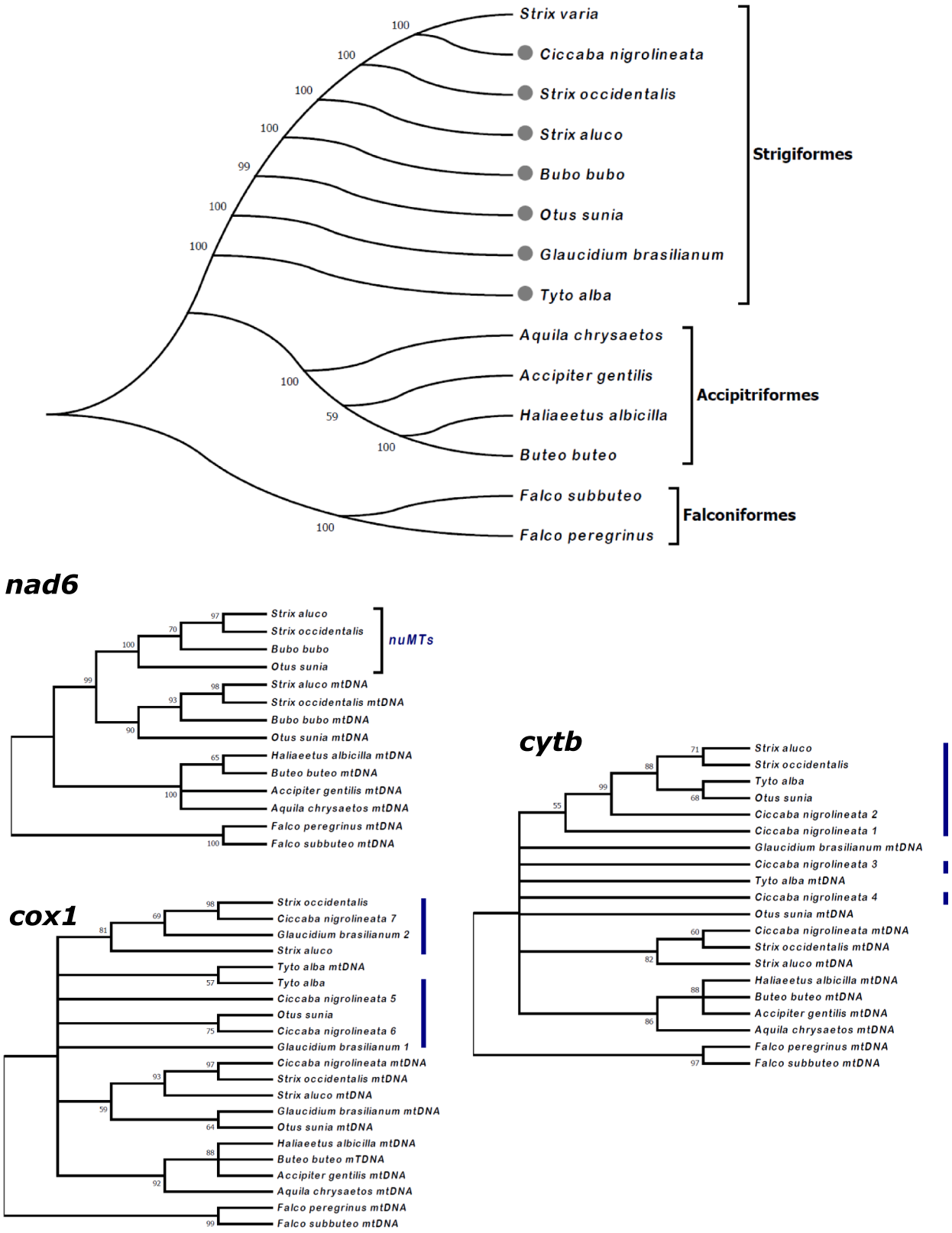

Fig. 1. Phylogenetic relationships. Here represented are the two-level phylogenetic relationships inspected in this manuscript. On the top is represented a ML-tree based on full mitochondrial molecules where grey circles represent Strigiformes with an available genome. On the bottom are the gene-specific phylogenies for mitochondrial regions (*cox1, cytb, nad6*) whose *NUMTs* have been reported for several species. Mitochondrial genes are labeled with "mtDNA" after the species name while *NUMTs* are numbered and identified with a blue bar. Bootstrap values are present for all branches to assess statistical support. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

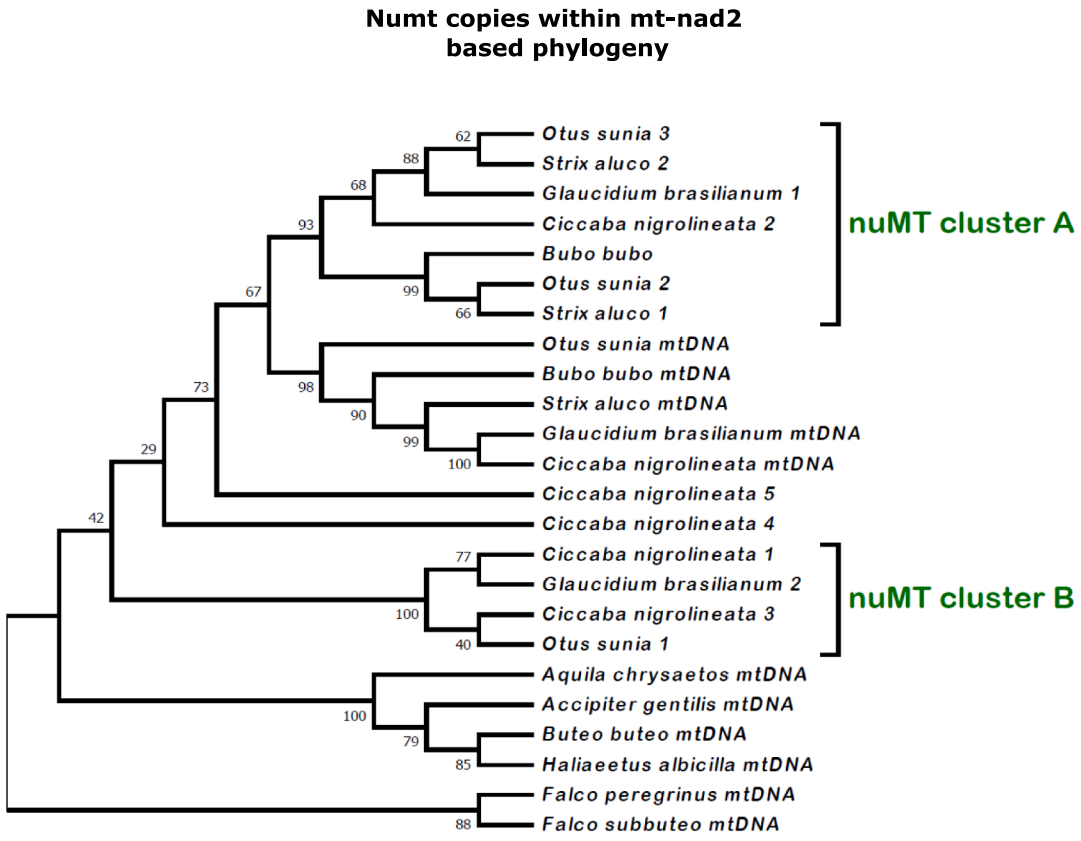

Fig. 2. *nad2* based phylogeny. Comparison of nad2 copies identified in owl' genomes. Similar to Fig. 1, true mitochondrial genes are labeled with "mtDNA", while *NUMTs* are labeled now with darker green and numbered. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

phylogenetic patterns where *NUMTs* and original mtDNA regions do not cluster together, and some deeper *NUMTs* divergences that pre-date speciation events.

4.1. Tawny owl's mitogenome improved assembly

The process of assembling and annotating the Tawny owl mitogenome led us to re-analyze the resources available for some members of the *Strigidae* family. The fact that we found multiple duplications/insertions of *nad3* and/or *dloop* in all other owl species might be attributed to errors propagated by an analytical process relying on assisted assembly to construct the circular mitochondrial genome from NGS reads, though we cannot strongly argue for this to be the case. Whether duplicated pairs are a true characteristic of owl mitogenomes remains to be proved. Nevertheless, the utilization of long-reads (such as PacBio or Nanopore) greatly improves the completeness of mitochondrial assemblies (Formenti et al., 2021) which is why we are able to complement the previous assembly of tawny owl mitogenome by filling a missing section containing *trnP*, *nad6*, and *trnE*.

4.2. Characterization of mito-nuclear insertions: protein-coding genes, copy number variation and phylogenetic relationships

In general, our results did not agree with the extraordinarily high number of putative *NUMTs* reported among the phylogeny of diurnal birds of prey (Falconidae = 43 < n < 49) (Nacer and do Amaral, 2017). Instead, our numbers are in line with those found so far in the chicken genome (n = 13) (Nacer and do Amaral, 2017). Exception is made for *Ciccaba nigrolineata*, though the number of *NUMTs* relate to several copies of a handful of mitochondrial regions. This uncommon pattern might be due to duplication events in the genome of this species, though the scarce information available for the assembly reports an estimated genome size of 1.2 Gb, which is in line with the other owls. Large variation in *NUMT* frequency has been detected across the broader scale of avian taxonomy, as shown by the high numbers observed in Tanagers and New World sparrows contrasting with other bird species (Liang et al., 2018). Our results suggest that high variation might also occur at finer phylogenetic scale, i.e., within the order of Strigiformes in this case. However, it is critical to stress that the threshold parameters we utilized in this study are more conservative than those employed by Liang et al (2018) and Nacer and do Amaral (2017). The fact that we applied a phylogeny-based threshold removed nuclear sequences which were highly similar to those of the original mitochondrial region. Furthermore, we were only able to work with a single species of the family Tytonidae and as such we cannot guarantee that the phylogenetic threshold based on Strigidae alone is not conservative for *Tyto alba*. Considering the loss of function of mitochondrial material upon its insertion in the nuclei and consequent free accumulation of mutations, it is safe to assume that our analytical pipeline filtered out recent insertions and rather captured molecular fossils (Hazkani-Covo et al., 2010). It is thus not surprising that the phylogenetic relationships between *NUMTs* and mtDNA genes estimated here for *cox1*, *cytb* and *nad2* suggest owl *NUMTs* to be older than speciation events, as it has been reported in insects of the order Orthoptera and resembling some sort of incomplete lineage sorting (Song, Moulton, & Whiting, 2014). Alternatively, in the light of the current knowledge gaps on *NUMTs* insertion, evolution, and maintenance, we cannot disregard the fact that mutations experienced by the inserts in the nuclear genome after pseudogenization may re-shape sequences to states recognized as ancestral by the maximum likelihood algorithm.

Similarly, our conservative thresholds might also help to explain why the large majority of the *NUMTs* identified in our study belong to former

mtDNA protein-coding genes, despite transfer-RNA regions are more represented in the original mitochondrial DNA molecule. The small size of tRNAs - which are often in the dozens of base-pairs in contrast with protein-codings hundreds or thousands - might render tRNA *NUMTs* unrecognizable after the free accumulation of mutations. The effect of mitochondrial region in explaining percentage of similarity is suggestive that some regions are more likely to be recognized than others after insertion and evolution in the nuclear genome. Assuming loss of function, we can hypothesize that structural stability of the DNA molecule is facilitated by features associated with nucleotide sequences and some inserts maintain the original sequence for longer than others. Alternatively, could be the case that some inserts have been utilized to generate novel protein sequences and are thus maintained in a re-shaped fashion (Noutsos, Kleine, Armbruster, DalCorso, & Leister, 2007). The negative correlation between *NUMT* size and percentage of identity might perhaps reinforce the challenges in directly detecting the full extent of the timing and size of *NUMT* insertion: recognizable regions labelled as *NUMTs* are but a fraction of the original inserts. While future research might indeed be directed towards the detection of insertions with increasing accuracy, it will benefit from a detailed documented database to develop machine learning or Bayesian algorithms.

4.3. *NUMT-embedded phylogenies suggest some insertions might pre-date speciation events*

Our reconstruction of phylogenies embedding *NUMTs* with their mitochondrial counterparts exposed scenarios that strikingly contradicted expectations: except for *cox1*, *NUMTs* rarely branch with the original mitochondrial region, particularly the different copies of *nad2*. In the context of phylogeographic or DNA barcoding, this study serves primarily to reinforce the growing concerns of *NUMT* amplification with mtDNA-designed primers (Bensasson, Zhang, Hartl, & Hewitt, 2001; Yao, Kong, Salas, & Bandelt, 2008). For instances, in species with high genetic diversity, a researcher might accept similarity percentage thresholds in the order of those we here utilize to define *NUMTs* and thus report larger numbers of mitochondrial haplotypes (Bertheau et al., 2011). Structural reasons for the detection ofhigher *NUMT* copies in some genomes in relation to others, i.e., the case in point being the extreme number of *NUMTs* in *Cicabba nigrolineata*, remain to be explained as there is not much available information on all species genomes aside from their assemblies. Noteworthily, factors such as the number and/or stability of mitochondria in the germline, species-specific mechanisms controlling accumulation/loss of nuclear DNA, or genome assembly quality might be key players in shaping interspecific diversity of *NUMTs* (Richly & Leister, 2004).

Overall, the identification of *NUMTs* is pivotal to the curation of both nuclear and mitochondrial assemblies which are increasingly becoming the targets of datamining studies. The capacity to characterize a multitude of patterns of molecular evolution is also inherent to advances in NGS and as such opens research avenues to better understand evolution and maintenance of mito-nuclear insertions. Though *NUMTs* are known to be involved in the occurrence of certain human diseases and aging (Reynolds, Bwiza, & Lee, 2020), not much else is known regarding functionality and impacts to organismal fitness in other organisms. Cleary, the knowledge gap will only be filled by exhaustive characterization of *NUMT* occurrence across taxa, which will allow us to build evolutionary models to theoretically predict insertion rates or life expectancy of mitochondrial insertions, and ultimately devise search strategies standardized across taxa to prevent sampling bias when screening whole genomes. Ultimately, *NUMT*-driven research may also expand in a population genetic direction and widening investigations towards characterizing variation at individual level. Nevertheless, the quality of genome databases is a critical consideration. A recent *meta*-analyses of the NCBI's avian mitogenome collection indeed showed that insertions/duplications are most likely assembly errors or unreported *NUMTs* (Sangster & Luksenburg, 2021). On the other hand, Formenti et al (2021) analyzed 100 vertebrate mitogenomes from long- and short-read data to show that duplications of mtDNA regions might otherwise be pervasive. To conclude, insertions reported here represent a starting point for the investigation of the existence of this specific type of mobile elements that will shed light on their evolutionary roles in genome biology.

5. Funding

The work developed in this manuscript was supported with the Academy of Finland funding decision 321417.

6. Data statement

Newly discovered *NUMT* sequences utilized to build maximum likelihood trees are available as a supplementary data (all_numts.fasta) and respective accession numbers of deposited genomes and mitogenomes utilized in their identification are available in supplementary tables. The new mitogenome is also submitted as a supplementary file (tawny_owl_mitogenome.faa) and we will provide accession number as soon as it is attributed. Scripts utilized for this study are available as supplementary text and deposited in https://github.com/Miguel-BSoares/ mito-nuclear-scripts.

CRediT authorship contribution statement

Miguel Baltazar-Soares: Conceptualization, Methodology, Data curation, Writing – original draft. **Patrik Karell:** Conceptualization, Writing – review & editing. **Dominic Wright:** Conceptualization, Writing – review & editing. **Jan-Åke Nilsson:** Conceptualization, Writing – review & editing. **Jon E. Brommer:** Conceptualization, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Tawny owl genome assembly is still under construction but a draft version can be made available upon request

Acknowledgements

We would like to thank the two anonymous reviewers and the members of the editorial board for the comments and suggestions on the manuscript. The authors also wish to acknowledge CSC – IT Center for Science, Finland, for computational resources and Academy of Finland Funding decision 321471.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2023.107722.

References

Ballard, J.W.O., Pichaud, N., 2014. Mitochondrial DNA: more than an evolutionary bystander. Funct. Ecol. 28 (1), 218–231.

Bensasson, D., Zhang, D.-X., Hartl, D.L., Hewitt, G.M., 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol. Evol. 16 (6), 314–321.

Bertheau, C., Schuler, H., Krumboeck, S., Arthofer, W., Stauffer, C., 2011. Hit or miss in phylogeographic analyses: the case of the cryptic NUMTs. Mol. Ecol. Resour. 11 (6), 1056–1059.

Chaisson, M.J., Tesler, G., 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinf. 13 (1), 1–18.
Donath, A., Jühling, F., Al-Arab, M., Bernhart, S.H., Reinhardt, F., Stadler, P.F., Bernt, M., 2019. Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. Nucleic Acids Res. 47 (20), 10543–10552.
Formenti, G., Rhie, A., Balacco, J., Haase, B., Mountcastle, J., Fedrigo, O., Ambrosini, R., 2021. Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. Genome Biol. 22 (1), 1–22.
Gaziev, A., Shaikhaev, G., 2010. Nuclear mitochondrial pseudogenes. Mol. Biol. 44 (3), 358–368.
Hazkani-Covo, E., Zeller, R.M., Martin, W., 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS genetics 6 (2), e1000834.
Jordal, B.H., Kambestad, M., 2014. DNA barcoding of bark and ambrosia beetles reveals excessive NUMT s and consistent east-west divergence across P alearctic forests. Mol. Ecol. Resour. 14 (1), 7–17.
Kleine, T., Maier, U.G., Leister, D., 2009. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. Annu. Rev. Plant Biol. 60 (1), 115–138.
Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27 (5), 722–736.
Kovar, L., Nageswara-Rao, M., Ortega-Rodriguez, S., Dugas, D.V., Straub, S., Cronn, R., Rodriguez, D.N., 2018. PacBio-based mitochondrial genome assembly of Leucaena trichandra (Leguminosae) and an intrageneric assessment of mitochondrial RNA editing. Genome Biol. Evol. 10 (9), 2501–2517.
Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33 (7), 1870–1874.
Li, H., 2012. seqtk Toolkit for processing sequences in FASTA/Q formats. GitHub 767, 69.
Liang, B., Wang, N., Li, N., Kimball, R.T., Braun, E.L., 2018. Comparative Genomics Reveals a Burst of Homoplasy-Free Numt Insertions. Mol. Biol. Evol. 35 (8), 2060–2064. https://doi.org/10.1093/molbev/msy112.
Nacer, D.F., do Amaral, F.R., 2017. Striking pseudogenization in avian phylogenetics: numts are large and common in falcons. Mol. Phylogenet. Evol. 115, 1–6.
Noutsos, C., Kleine, T., Armbruster, U., DalCorso, G., Leister, D., 2007. Nuclear insertions of organellar DNA can create novel patches of functional exon sequences. Trends Genet. 23 (12), 597–601.
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 (1), 1–8.
Puertas, M.J., González-Sánchez, M., 2020. Insertions of mitochondrial DNA into the nucleus—effects and role in cell evolution. Genome 63 (8), 365–374.
R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.
Reynolds, J.C., Bwiza, C.P., Lee, C., 2020. Mitonuclear genomics and aging. Hum. Genet. 139 (3), 381–399.
Ricchetti, M., Fairhead, C., Dujon, B., 1999. Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. Nature 402 (6757), 96–100.
Richly, E., Leister, D., 2004. NUMTs in sequenced eukaryotic genomes. Mol. Biol. Evol. 21 (6), 1081–1084.
Salter, J.F., Oliveros, C.H., Hosner, P.A., Manthey, J.D., Robbins, M.B., Moyle, R.G., Faircloth, B.C., 2020. Extensive paraphyly in the typical owl family (Strigidae). The Auk 137 (1), ukz070.
Sangster, G., Luksenburg, J.A., 2021. Sharp Increase of Problematic Mitogenomes of Birds: Causes, Consequences, and Remedies. Genome Biol. Evol. 13 (9), evab210.
Schatz, D.G., Swanson, P.C., 2011. V (D) J recombination: mechanisms of initiation. Annu. Rev. Genet. 45, 167–202.
Shen, W., Le, S., Li, Y., Hu, F., 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One 11 (10), e0163962.
Song, H., Moulton, M.J., Whiting, M.F., 2014. Rampant Nuclear Insertion of mtDNA across Diverse Lineages within Orthoptera (Insecta). PLoS One 9 (10), e110508.
Wick, R.R., Judd, L.M., Cerdeira, L.T., Hawkey, J., Méric, G., Vezina, B., Holt, K.E., 2021. Trycycler: consensus long-read assemblies for bacterial genomes. Genome Biol. 22 (1), 1–17.
Yao, Y.-G., Kong, Q.-P., Salas, A., Bandelt, H.-J., 2008. Pseudomitochondrial genome haunts disease studies. J. Med. Genet. 45 (12), 769–772.