Diverse evolutionary trajectories of mitocoding DNA in mammalian and avian nuclear genomes

Yu-Chi Chen, David LJ Vendrami, Maximilian L Huber, et al.

Genome Res. published online March 31, 2025
Access the most recent version at doi:10.1101/gr.279428.124

Diverse evolutionary trajectories of mitocoding DNA in mammalian and avian nuclear genomes

Yu-Chi Chen[1,2], David L. J. Vendrami[3,5,8], Maximilian L. Huber[2], Luisa E. Y. Handel[2], Christopher R. Cooney[4], Joseph I. Hoffman[3,5,6,7,8], Toni I. Gossmann[*,1,2,3,8]

[1] TU Dortmund University, Computational Systems Biology, Faculty of Biochemical and Chemical Engineering, 44227 Dortmund, Germany

[2] Bielefeld University, Department Evolutionary Genetics, Faculty of Biology, 33615 Bielefeld, Germany

[3] Bielefeld University, Department of Animal Behaviour, Faculty of Biology, 33615 Bielefeld, Germany

[4] University of Sheffield, Ecology and Evolutionary Biology, School of Biosciences, Sheffield S3 8ER, UK

[5] Bielefeld University, Department of Evolutionary Population Genetics, Faculty of Biology, 33501 Bielefeld, Germany

[6] Bielefeld University, Center for Biotechnology (CeBiTec), Faculty of Biology, 33615 Bielefeld, Germany

[7] British Antarctic Survey, High Cross, Madingley Road, Cambridge CB3 OET, UK

[8] Joint Institute for Individualisation in a Changing Environment (JICE), Bielefeld University and University of Münster, Bielefeld, Germany

[*]**Contact:** toni.gossmann@tu-dortmund.de

Abstract

Sporadically genetic material that originates from an organelle genome integrates into the nuclear genome. However it is unclear what processes maintain such integrations over evolutionary time. Recently it was shown that nuclear DNA of mitochondrial origin (NUMTs) may harbour genes with intact mitochondrial reading frames despite the fact that they are highly divergent from the host's mitochondrial genome. Two major hypotheses have been put forward to explain the existence of such mitocoding nuclear genes: (i) recent introgression from another species and (ii) long-term selection. To investigate whether these intriguing possibilities play a role we scanned the genomes of more than 1,000 avian and mammalian species for NUMTs. We show that the subclass of divergent NUMTs harbouring mitogenes with intact reading frames are widespread across mammals and birds. We show that some of these NUMTs appear to have similarity across species. We also demonstrate that many mitochondrial-coding NUMTs exhibit signs of long-term selection. In a subset of these NUMT genes, we detected evolutionary signals consistent with adaptive evolution, including one human NUMT shared among seven ape species. These findings suggest that NUMT insertions may occasionally be functional.

Keywords: Mitochondrial genomes, lateral gene transfer, NUMT, nuclear genome integrations, positive selection, introgression

Introduction

A key question in evolutionary biology is how genetic novelties arise and contribute to adaptation (Sangster and Luksenburg 2021; Bomblies and Peichel 2022). Mutations are the ultimate source of genetic novelties that are subsequently subject to evolutionary forces such as selection and drift (Huang et al. 2016). There is compelling evidence that most mutations do not show a signature of positive selection on the molecular level, but instead are subject to random genetic drift and purifying selection (Bank et al. 2014). As novel mutations may be insufficient in many cases to generate adaptive variation, alternative mechanisms may explain the generation and maintenance of novel beneficial genetic elements.

Mutations that have already been exposed to evolutionary forces may provide a means of generating adaptive variation. For example, the acquisition of novel, pre-adapted genetic elements can occur via the lateral transfer of DNA. There is considerable evidence for such exchanges of genetic material among bacteria and even in eukaryotes (Danchin 2016; Dunning et al. 2019). For example, the substantial exchange of genetic material between eukaryotic species occurs through hybridisation and introgression (Gabaldón 2020; Setter et al. 2020; Popadin et al. 2022). While it is known that an organism's genomic architecture is fundamental to whether or not hybridization/introgression occurs successfully, we have limited knowledge of the extent to which lateral gene transfer, hybridisation and introgression contribute to genomic diversification driven by the acquisition of adaptive elements across species.

An intriguing possibility for the acquisition of novel, genetically adapted elements is the transfer of organelle DNA (Popadin et al. 2022; Butenko et al. 2024). In many eukaryotic species, organelles such as mitochondria or plastids possess their own genomes. Mitochondrial DNA integration into the nuclear genome (**NUMT**) is a continuous and ongoing mutagenic process, which occurs across different species and tissues (Ricchetti, Tekaia, and Dujon 2004). After NUMTs were first observed in the domestic cat (*Felis catus*) (Lopez et al. 1994) and in humans (Hu and Thilly 1994; Lopez et al. 1997) they were then found in many different eukaryotic species from yeast (Ricchetti, Fairhead, and Dujon 1999) to many other mammals (Tsuji et al. 2012; Calabrese et al. 2017; Gossmann et al. 2019; Biró et al. 2022; Uvizl et al. 2023) and birds (Nacer and Amaral 2017; Liang et al. 2018; Lucas, Vincent, and Eric 2022; Baltazar-Soares et al. 2023). NUMT insertion number, size and sequence composition vary not only across species (Richly and Leister 2022; Hazkani-Covo 2022) but also within populations (Soto-Calderón et al. 2014; Lucas, Vincent, and Eric 2022; Wei Wei et al. 2022). In humans, around 14% of individuals possess NUMTs that are present in less than 0.1% of the broader population and some of these NUMTs have been shown to be passed on from parents to their offspring (Wei Wei et al. 2022). This suggests that NUMT integration is a frequent mutagenic event that produces inheritable changes, potentially relevant to future evolution. The germline NUMT mutation rate in humans is estimated at 2.44×10^{-4} mutations per generation (Wei Wei et al. 2022) illustrating that novel NUMT genetic variation is frequently and constantly being generated.

While NUMTs are widespread in the animal kingdom, there is little evidence of functional integrations

of mitochondrial DNA (Noutsos et al. 2007; Pozzi and Dowling 2019; W. Wei and Chinnery 2020). Instead, it is generally assumed they are exposed to genetic drift and will quickly accumulate novel disruptive mutations. In line with this, it is frequently observed that mitocoding genes integrated into the nuclear genome accumulate missense mutations that ultimately lead to a broken reading frame. However, if the integration is very recent, the local mutation rate is very low, or selection actively maintains the coding properties of the integration, the former mitochondrial coding genes may retain intact reading frames over longer periods of evolutionary time. Such NUMTs containing genes with intact mitochondrial reading frames can be considered as **coding NUMTs (cNUMTs)**.

Recently in the genome of the Antarctic fur seal (*Arctocephalus gazella*) several cNUMTs were identified (Vendrami et al. 2022). The coding genes of three of the cNUMTs contained numerous silent mutations (i.e. showed substantial divergence at neutral sites), but very few amino-acid changing mutations relative to fur seal mitochondrial genes. The authors concluded that the signature of coding NUMTs that are divergent (**dcNUMTs**) at synonymous sites was consistent with purifying selection and hinted at functionality and introgression. As the dcNUMTs observed in the Antartic fur seal genome cannot be explained by a recent nuclear integration from its own mitochondrial genome, alternative scenarios of non-canonical integrations might explain the existence of dcNUMTs (Figure 1). Currently, it is unclear how common and frequent dcNUMTs are across larger taxonomic groups and whether they harbour signatures of functionality and introgression. Hence their potential role as adaptive elements of the genome has yet to be thoroughly investigated. In this study we aimed to unravel the role of organelle DNA in creating novel and potentially adaptive variation in nuclear genomes. For this we investigate whether (1) mitocoding DNA may stem from lateral gene transfer/introgression and (2) whether we find any evidence of selection on nuclear mitocoding DNA.

Results

Class-wide NUMT Identification

To examine the evolutionary impact of NUMT integration, we identified mitochondrial insertions in the nuclear genomes of over a thousand mammalian and avian species. Specifically, we analyzed 680 mammalian and 458 avian genomes with both nuclear and mitochondrial assemblies available from public databases. Using a previously published bioinformatic pipeline (Vendrami et al. 2022), we detected 85,100 NUMTs across 435 mammalian species (66.62%) and 28,947 NUMTs across 458 avian species (100%). These results align with earlier studies showing that NUMT integration is a frequent, ongoing process in vertebrates (Hazkani-Covo 2022). On average, mammals had more NUMTs than birds (≈ 125 vs. ≈ 63 per species), with NUMT count positively correlated with genome size in mammals ($P < 2.2 \times 10^{-16}$) but not in birds ($P = 0.79$, phylogenetic generalized least-squares regression).

Phylogenetic patterns of NUMTs in mammals and birds

Given the contrasting patterns in NUMT abundance between mammals and birds, we investigated the phylogenetic distribution of NUMTs. Established species phylogenies for mammals and birds (Kumar et al. (2017), Timetree.org) were used to reconstruct NUMT counts as a phylogenetic trait through a Bayesian approach implemented in BayesTraits V4.0.1 (Pagel and Meade 2022). Several clear phylogenetic patterns were observed.

First, NUMT counts showed substantial variation across clades (Figure 2). Among mammals, most marsupials exhibited either an absence or low numbers of NUMTs, with the exception of the Tasmanian devil (*Sarcophilus harrisii*) (Hazkani-Covo 2022). In contrast, clades such as Old World monkeys and cetaceans displayed elevated NUMT counts. In birds, NUMT enrichment was particularly pronounced in three clades within the order Passeriformes: (1) warblers, cowbirds, and blackbirds; (2) sparrows (e.g., *Zonotrichia albicollis*), juncos, seedeaters, tanagers, and finches (e.g., *Geospiza fortis*); and (3) cardinals (*Cardinalis cardinalis*; Sin, Lu, and Edwards (2020)), buntings, grosbeaks, longspurs, and grackles.

Second, NUMT counts varied significantly among closely related species. For example, the highly endangered spoon-billed sandpiper (*Calidris pygmaea*) exhibited extensive NUMT integration, suggesting a high level of genome degeneration (Chowdhury, Foysal, and Green 2022). Similarly, among mammals, the Tasmanian devil (*Sarcophilus harrisii*, 1,546 NUMTs), gray short-tailed opossum (*Monodelphis domestica*, 811 NUMTs), and yellow-footed antechinus (*Antechinus flavipes*, 792 NUMTs) showed elevated NUMT counts. In birds, in addition to *Calidris pygmaea* (13,170 NUMTs), the island finch (*Nesospiza acunhae*, 403 NUMTs) and northern cardinal (*Cardinalis cardinalis*, 373 NUMTs) also exhibited significant NUMT accumulation.

Distinct distributions of coding NUMTs

While mitochondrial integration into the nuclear genome is frequent and ongoing across many species, the origins and evolutionary fate of NUMTs, particularly those retaining intact mitochondrial coding frames over long timescales, remain unclear. To address this, we focused on a subset of NUMTs with at least one intact coding gene (coding NUMTs, cNUMTs) identified using the vertebrate mitochondrial genetic code. Of the 114,047 NUMTs identified, 6,429 (5,229 in 379 mammalian species and 1,200 in 319 avian species) retained at least one intact mitochondrial gene. In mammals, the number of cNUMTs is positively correlated with non-coding NUMTs ($P < 2.2 \times 10^{-16}$), whereas no such correlation exists in birds ($P = 0.9$), as shown by phylogenetic generalized least-squares regression. This suggests that cNUMTs may arise as a by-product of NUMT integration in mammals but not in birds. To estimate the evolutionary age of cNUMTs, we calculated pairwise synonymous divergence (dS) between cNUMT genes and their mitochondrial counterparts (Wang et al. 2010). The distribution of dS values provides insight into integration times across species (Figure 3). We observed a decline in cNUMT abundance with increasing dS values in mammals, while birds exhibited a pronounced peak around $dS \approx 1$.

Noncanonically integrated NUMTs

As recently integrated NUMTs will show little to no sequence divergence from their mitochondrial counterparts, they cannot tell us very much about the evolutionary consequences of NUMT integration. We hence focus on those cNUMTs that show substantial synonymous divergence from the respective gene in the host mitochondrial genome (dcNUMTs, divergent coding NUMTs with at least one gene with $dS > 0.1$ relative to the respective host mitochondrial gene), because these are likely the consequence of noncanonical NUMT integrations (Figure 1). We identify 4,182 and 917 NUMTs that are dcNUMTs with 8,406 and 2,599 coding genes with $dS > 0.1$ in mammals and birds respectively, meaning that some dcNUMTs contain multiple divergent mitocoding genes.

We identify 148 dcNUMTs (3.54 %) in mammals and 249 dcNUMTs (27.15 %) in birds that we can assign to a non-host mitochondrial genome with high confidence (Table 1, Figures 4 and S1). We find that 77.19% and 40.46% of the dcNUMTs in mammals and birds, respectively, have no hit with any of the available mitochondrial genomes at the 95% identity level. To test whether these NUMT perhaps originate from species outside of the two taxonomic clades, we conducted further BLAST analyses but did not identify any further hits. The phylogenetic origins of these dcNUMTs are therefore located in their respective taxonomic groups but their exact origins remain obscure.

dcNUMT gene analysis

Because dcNUMTS contain at least one gene with an intact reading frame, it is possible to investigate the evolutionary pressures that prevailed since the split from the last common ancestor using DNA codon models. For example, if genetic drift was the predominating force since the split from the last common ancestor of the NUMT and the host' mitochondrial genome, one would expect the ratio of nonsynonymous to synonymous substitutions (d_N/d_S) to be approximately one. It is important to note that in our analysis, we do not assume that NUMTs are transcribed or that their gene products are imported into mitochondria. The evolutionary patterns we observe do not depend on whether NUMTs are expressed or functional in a traditional sense. Instead, the signatures of selection may reflect broader genomic processes, such as linkage to functional genomic regions under selection or historical constraints on these sequences prior to their nuclear integration.

To test selection on dcNUMTs, we divided our dataset into: (A) Single NUMT genes (Figure 5A), which are highly divergent from the other NUMT genes and therefore seem to be isolated ("single"); and (B) NUMT gene clusters (Figure 5B) that are highly similar, for instance due to shared speciation events (Vendrami et al. 2022). We then aligned the single NUMTs and NUMT gene clusters together with available mitochondrial genes of related species to calculate dcNUMT gene (Figure 5A) and gene cluster specific d_N/d_S ratios (Figure 5B) using codeml (Yang 2007). Codeml does not incorporate heterogeneity in codon composition or transition-to-transversion ratio (κ) which may differ between the nuclear and mitochondrial genome (Belle et al. 2005). However, we performed a series of simulations (Figure S2 and Table S1) that incorporate such heterogeneity (Fletcher and Yang 2009) which show little evidence for selection under a neutral model. We therefore consider our approach suitable to infer selection for nuclear integrated mitochondrial sequences.

Selection analysis on isolated dcNUMT genes imply noncanonical NUMT integrations and "missing" mitochondrial genomes. We first conducted a codon analysis of the "single" dcNUMT genes. In mammals, we identified 3,549 single dcNUMT genes (42.2% of total dcNUMT genes) and 2,246 (86.4% of total dcNUMT genes) in birds. The differences in these proportions between the two taxonomic groups is likely because dcNUMTs genes in birds have greater synonymous divergence and are therefore generally more ancient (Figure 3). We conducted branch-specific analyses to test for positive and negative selection, as well as drift (Table 2). For the majority of single dcNUMT genes (Table 2), we identify sequence evolutionary patterns consistent with neutral evolution (62.4% for mammals and 83 % for birds). We also observe a smaller number of cases (37.6% for mammals and 17% for birds) that are consistent with purifying selection, and five cases of positive selection.

Selection analysis on NUMT gene clusters suggests positive and negative selection on NUMTs. A proper evolutionary analysis of a single phylogenetically isolated NUMT gene is limited by the availability of the last mitochondrial lineage from which the mitochondrial DNA was transferred into a nuclear genome. To circumvent this limitation, we clustered the NUMT genes according to their similarity to each other independent of the host species. We then aligned these groups of gene sequences together with the most closely related mitochondrial genes available in our database. Based on these alignments, we then created gene trees and conducted clade-specific tests of selection (Table 3). We excluded gene trees that did not form monophyletic NUMT clusters. We then evaluated the selection pressures acting on the clade representing the diversification of the NUMTs (i.e. a node-based last common ancestor approach (Lee 1998), in contrast to branch-based last common ancestor, e.g. excluding the grey branch in Figure 5B). It is therefore likely that the considered clade reflects an evolutionary time when the NUMT was truly integrated into the nuclear genome, e.g. a pattern one would observe due to speciation events (Vendrami et al. 2022). Consequently the estimated selective pressure analysis should reflect selection on the NUMT, not the mitochondrial DNA.

We identified over ten times more of these NUMT gene clusters for mammalian species than for birds (996 versus 98). The average number of species with NUMT genes per cluster was three for both taxonomic groups (2.98 and 3.01, for mammals and birds respectively). While mammalian dcNUMTs are generally less divergent from the host's mitogenome than avian dcNUMTs (Figure 3), they are as frequently shared between species. Based on the evolutionary rate analysis we identify more than 250 gene clusters that show evidence of purifying selection, most of them specific to mammals. Moreover, we identify an additional 24 gene clusters, 23 for mammals and one for birds, that show evolutionary signatures consistent with positive selection.

Evidence of positive selection in a human NUMT. We identified one NUMT on Chromosome 2 in humans that is part of a ND4L gene cluster with evidence for positive selection (Table 4, one out of 23 NUMT clusters for mammals). This particular NUMT is shared among seven ape species (*Nomascus leucogenys*, *Pan troglodytes*, *Gorilla gorilla gorilla*, *Pan paniscus*, *Homo sapiens*, *Hylobates moloch*, *Pongo abelii*) but the NUMT itself clusters with mitochondrial sequences of species of the family Callitrichidae (Figure 6). In the human genome, this NUMT is flanked by genes encoding

for ARHGAP15, GTDC1, ZEB2, KYNU and LRP1B. We also conducted an AlphaFold structural prediction (Mirdita et al. 2022) using the NUMT's gene sequence translated into a protein with the vertebrate mitochondrial code (Figure S3) and find very little 3D structural variation between the human ND4L protein and the 3D protein prediction of the NUMT gene.

Discussion

Here, we present a comprehensive analysis of mitochondrial DNA that has been integrated into the nuclear genomes (NUMTs) of more than 1,000 mammalian and avian species and identify more than 80,000 NUMTs in mammals and more than 29,000 NUMTs in birds (Table 5). Generally, there are more NUMTs in mammalian genomes, which is not unexpected given that mammalian genomes tend to be larger (Kapusta, Suh, and Feschotte 2017). We also find a positive relationship between genome size and the number of mitochondrial integrations in mammals but not birds. A possible explanation for this is that mammals contain NUMT copies resulting from intragenomic duplications. Birds are known to exhibit very little variation in their genome sizes in general, which might also explain the lack of intragenomic copies, such as segmental duplications. For many of the identified NUMTs, it is unclear on which chromosome in the genome they are located. Instead they are mainly located on small scaffolds that are often not much bigger than a mitochondrial genome (e.g. <20kb, Figure S4). However, large-scale application of long-range haplotype-based genomic sequencing (Rhoads and Au 2015; Lu, Giordano, and Ning 2016) will overcome the current limitations of identifying the precise genomic locations of nuclear DNA of organelle origin in the near future (Wilcox et al. 2022).

Some of the identified NUMTs are very distantly related to the host's mitochondrial DNA. As our approach of identifying NUMTs tends to detect NUMTs that show a reasonable degree of similarity between the NUMT and the host's mitochondrial genome, it is likely that we did not identify all distantly related NUMTs. Therefore, our estimates of the number of NUMTs per species are potentially underestimates. For example, we do not find any NUMTs that are highly similar to a species outside of the respective animal class. Because we do not attempt to identify NUMT integration events as such, some of the NUMTs may result from structural mutations, such as segmental genome copies (Wilcox et al. 2022). Hence our NUMT numbers are likely an overestimate of the number of NUMT integration events from nonnuclear DNA. However, because we focus on dcNUMTs for our evolutionary analysis it is irrelevant whether a NUMT has undergone intragenomic copying or is the result of a "original" genomic integration.

Our method also depends on an accurate database assignment of the mitochondrial genome to the respective species. However, due to biological mitochondrial diversity (Clark et al. 2023) or simply erroneous entries in databases (Laine et al. 2019; Caswell et al. 2019; Sangster and Luksenburg 2021), mitochondrial DNA and the NUMT DNA could appear more divergent. To limit the impact of this in our dcNUMT analysis, we focused on coding NUMTs with at least 10% synonymous divergence to the respective mitochondrial gene. To investigate the possibility of erroneous database entries we checked the mitochondrial genomes of 22 bird species that show a large number of highly dissimilar NUMTs. For this, we assembled the mitochondrial genomes of these species from raw sequencing data

using an established pipeline (Jin et al. 2020). While we observed some genetic variation between the assembled and the canonical mitochondrial genomes in the database (Table S4), we did not find any evidence for an incorrect assignment of mitochondrial genomes. We conclude that falsely assigned mitochondrial genomes play a very limited role in our analysis.

dcNUMTs We find more than 5,000 NUMTs in mammals and birds (Table 5) that contain genes with an intact mitochondrial reading frame and that are highly divergent from the respective hosts mitochondrial encoding genes (dcNUMTs with genes with $dS > 0.1$, i.e. a synonymous per site divergence of more than 10%). In mammals we find that the number of dcNUMTs becomes lower for increasing d_S values but in birds there is a clear peak at dS$\approx$ 1. This might reflect a historic episode of high transposable element activity common to all birds. However, as the nuclear and mitochondrial mutation rates may vary substantially across species, we speculate that the dS estimates should be considered only as a rough proxy of the integration time of the NUMT.

It is possible that dcNUMTs are transcribed (Cohen, Levin, and Mishmar 2016) and may have a function on the RNA level (Plazzi et al. 2023), or that the DNA itself has some regulatory function as observed for mitochondrial DNA (Noutsos et al. 2007; Acker et al. 2023). For example, the transcription of noncoding genomic DNA is very common (Gao et al. 2020) and active retroviruses (Bolisetty et al. 2012) may lead to the NUMTs being expressed as well. However, this remains speculative and would need to be supported by genomic signatures consistent with functionality. Further experimental validation, such as RNA-seq analysis of these genomic regions, is needed to confirm whether NUMTs in bird species are indeed transcribed.

Noncanonical integration To identify potential origins of noncanonically integrated NUMTs we compared dcNUMTs against available mitogenomes that also contain mitochondrial genomes that were not in our initial analysis (e.g. because there was no nuclear genome available for the species). The excess of synonymous mutations for these NUMTs likely reflect purifying selection in the original mtDNA, therefore represent these patterns likely evolutionary signatures acquired prior to integration rather than resulting from ongoing selection in the nuclear genome. These NUMT insertions may appear relatively intact because they have had less time to accumulate mutations. We observe a greater number of distant hits in birds than mammals. This is possibly due to the fact that the bird genomes in our dataset diverged earlier than the mammal genomes - which means that their mitogenomes are more diverged, allowing for better discrimination of noncanonical NUMT origins. In mammals, the majority of the noncanonically integrated NUMTs occur within the ungulates (Figure 4A and Table S2) - a taxonomic group in which hybridization is frequently occurring (Iacolina et al. 2018).

In birds, we find several examples of noncanonical integration events where NUMTs are similar to mitochondria of distantly related species (Figure 4B, Table S3). For example, we identified a NUMT in the gweela (*Alectura lathami*), a mound-building bird from the family Megapodiidae found in eastern Australiathat, that shows high similarity to the tawny frogmouth mitogenome (*Podargus strigoides*), a species native to the Australian mainland and Tasmania. A possible explanation might be that some NUMTs co-localise with retrotransposon elements (Tsuji et al. 2012; Bolisetty et al.

2012; Dayama et al. 2014; Schiavo et al. 2017), and the transmission of retroviruses between species has been documented in mammals and non-vertebrates (Flavell 1999; Diehl et al. 2016). We also note that the W chromosome is a refugium for retroviruses in birds (Peona et al. 2021), which might imply that we may have missed NUMTs that are located on the W if the respective genome of a species was obtained from a male individual, which has the homogametic genome in birds. We cannot pinpoint the exact genomic locations of many of the identified NUMTs, because they tend to be located on very short scaffolds (Figure S4) - and this appears to particularly true for avian dcNUMTs. It has been noted previously that some avian chromosomes have not been assembled in many bird genomes, such as Chromosome 16 (Laine et al. 2016), smaller Microchromosomes (Kapusta, Suh, and Feschotte 2017) and the germline restricted chromosome (Kinsella et al. 2019). The reduced ability to locate the precise genomic locations, however, hampers the possibility to pinpoint integration events that happened as a consequence of past genomic hybridization. In the future this can be overcome by the application of long-range sequencing approaches, which will subsequently allow us to use dcNUMTs as markers for genomic hybridisation, and which could facilitate the unraveling of their potential role as adaptive elements.

We also identify 3,179 and 384 dcNUMTs in mammals and birds, respectively for which we cannot identify a homologous sequences with a sequence identity of more than 95% (Table 1). While this could mean that these are simply older NUMTs that are highly diverged but which have remained in the genome for a long time, it could also mean that some mitogenomes are lacking from our database. This is either because the respective species has not been yet sampled or because the respective species has already gone extinct. The number of "orphan" NUMTs is almost ten-fold higher in mammals compared to birds. Sequencing more mammalian species' genomes and the retrieval of ancient mammalian (mito)genomes might enable us to reveal the evolutionary origins and fates of further dcNUMTs in the future.

Choice of genetic code Here we use of the mitochondrial genetic code for d_N/d_S calculations. In our analysis, we opted for the mitochondrial genetic code because our primary focus is to investigate the (ancient) integration of mitochondrial sequences into the nuclear genome, and how these sequences have evolved over time relative to mitochondrial genome of the host species. The maintenance of mitochondrial reading frames and the accumulation of mutations that disrupt these frames in the nuclear context can provide important insights into the evolutionary history of these elements. However, it is crucial to recognize that if some of these NUMTs are functional in a nuclear environment (e.g., through transcriptional or regulatory roles), then a re-assessment using the standard nuclear code might be warranted for a more nuanced understanding of their evolutionary trajectory.

Tests of selection For the majority of single NUMT genes, we observe a signature consistent with drift (e.g. d_N/d_S=1), but also identify more than 2,000 genes showing signatures of purifying selection (Table 2). We also identify five NUMT genes showing signatures of positive selection. While it is intriguing that many of these genes have been evolving under selection pressure, this analysis is hampered by the fact that for the tested branches, the single NUMT genes may have existed as mitochondrial genome sequences. Unfortunately we cannot incorporate the precise date of the nuclear

integration, as we potentially lack a substantial amount of mitodiversity (Table 1) and therefore cannot be sure that we include a sequence that did not diverge before the NUMT integration took place. To address this limitation, we conducted clade-specific tests of selection on groups of NUMT genes. Therefore the considered branches in the clade with the NUMTs may reflect an evolutionary time when the NUMT was truly integrated into the nuclear genome. However, it is also possible that this approach groups sequences that have evolved the same sequence independently (Hazkani-Covo, Zeller, and Martin 2010; Song et al. 2013).

Positive selection in a human NUMT gene Finally, we identified one gene within a human NUMT that is part of a cluster of seven ape NUMTs. For humans, the genes located in the genomic region where the NUMT is located are known to have evolved under positive selection in the ancestral lineage of modern humans (Green et al. 2010; Prüfer et al. 2013; Racimo 2015). Moreover, structural mutations, such as deletions and duplications, in the respective region are known to cause Mowat-Wilson syndrome in humans (Baxter et al. 2017; Goyal et al. 2022), a rare but severe genetic disorder. Because the NUMT is linked to functionally important genes, it is a possible explanation of the signal of positive selection - the NUMT might be hitchhiking with a gene under positive selection (Barton 2000). The fact that one mitochondrial gene on the NUMT shows a signature of positive selection at the (mitochondrial) protein coding level that is shared between many ape species, fuels speculation to a potential functional role of this NUMT or one of its genes. Our hypothetical 3D prediction (Figure S3) does not suggest that the accumulated mutations in the NUMT strongly alter the three-dimensional structure. It therefore appears unlikely that the signal of positive selection of the mitocoding nuclear DNA is the consequence of a false positive, e.g. through misalignment or rapid accumulation of random amino-acid altering mutations (Mallick et al. 2009).

Haplotype resolved sequencing Here we have shown that NUMTs may originate from noncanonical integration and that signatures of purifying and positive selection are common and widespread across two main vertebrate clades. Currently, in many genome assemblies NUMTs cannot be placed onto chromosomes, because they are usually flanked by highly repetitive sequences due to retro-virus integration. Haplotype-resolved sequencing will likely be able to place more NUMTs on chromosomal locations, and thereby reveal the evolutionary signatures of NUMTs as a consequence of linkage (Wilcox et al. 2022). Our example of a human NUMT has shown that it is crucial to know the genomic context to deduce the likely functional roles of these intriguing genetic elements. Moreover, recent and ancient integration may be understood better when placing NUMTs into phylogenetic context. Even if NUMTs are not inherently functional, they serve as valuable markers of selection and may arise as by-products of genomic introgression or hybridization events. Additionally, their integration near functional elements suggests a potential role for background selection in maintaining coding-like NUMTs within nuclear regions, particularly in areas of low recombination (Gossmann et al. 2014). This selective pressure could further suppress the rate of mutation accumulation in NUMTs.

Methods

Data Collection and Selection

Genome assemblies for all mammalian species (available before November 11, 2021) and all avian species (available before November 2, 2021) were downloaded from GenBank (Sayers et al. 2021) using ncbi-genome-download v0.3.1 (https://github.com/kblin/ncbi-genome-download). For species with multiple assemblies, the representative genome or the assembly with higher coverage (if no representative was specified) was selected.

Mitogenomes containing at least one partial coding gene for all mammalian species (available before November 22, 2021) and all avian species (available before December 16, 2021) were obtained from GenBank to build BLAST databases. Mitogenomes from the same species or closely related subspecies with genome assemblies were prioritized for NUMT identification. When multiple mitogenomes existed for a species, the reference assembly was used. Unverified mitochondrial genomes were excluded. Protein-coding gene annotations were obtained from GenBank, and mitogenomes lacking annotations were annotated via the MITOS2 webserver (Donath et al. 2019). Table S5 lists the genomes, mitochondrial genomes, and their accession numbers.

NUMT Identification

NUMTs were identified using local BLASTN v2.6.0+ (Camacho et al. 2009), with a word size of 20. Mitochondrial genomes were used as queries against genome assemblies from the same species, following a modified approach from previous studies (Lammers et al. 2017; Vendrami et al. 2022). Hits shorter than 200 bp were discarded, and NUMT sequences were extracted using BEDTools v2.26.0 (Quinlan and Hall 2010).

To identify coding NUMTs (cNUMTs), NUMTs and their reverse complements (generated with the EMBOSS v6.6.0.0 *revseq* algorithm (Rice, Longden, and Bleasby 2000)) were aligned to the protein-coding genes of the corresponding mitochondrial genome using MAFFT v7.310 (Katoh and Standley 2013).

NUMTs were mapped to species phylogenies from Timetree (Kumar et al. 2017) to assess evolutionary trends. Evolutionary rates were estimated using an independent contrast model with reversible jump Markov chain Monte Carlo (RJMCMC) in BayesTraits v4.0.1 (Pagel and Meade 2022). Each dataset was run for 12,000,000 iterations, with sampling every 2,000 iterations following a 2,000,000-iteration burn-in. Ancestral reconstructions of NUMT counts, adjusted for branch length rate changes, were performed using the *ace* function in the *phytools* package (Revell 2012) in R v4.2.3.

Substitution Rate Analysis of NUMTs with Coding Genes

We first assessed whether cNUMTs were in-frame by checking pairwise alignments with mitochondrial protein-coding genes for the absence of internal stop codons. In-frame cNUMTs and their corresponding mitochondrial genes were used to calculate synonymous substitution rates (d_S) using the KaKs Calculator (Wang et al. 2010) under the NG substitution model (Nei and Gojobori 1986).

NUMTs with at least one coding gene and $d_S \geq 0.1$ were classified as potentially divergent cNUMTs (dcNUMTs).

These dcNUMTs were used as queries in BLAST searches (Camacho et al. 2009) against mitogenomes of all mammalian and avian species (Table S5). The best BLAST hits were defined as those with the highest bit score, while high-confidence hits were defined by at least 95% sequence identity to the query. dcNUMTs with high-confidence hits to mitogenomes from other species, where the percentage of identical matches differed by at least 5% from hits to their host species, were considered potential cases of lateral transfer between species pairs.

Clustering NUMTs and selection analysis

To investigate the functional constraints on coding sequences on cNUMTs, we first identified NUMTs derived from the same source and then analyzed the ratio of nonsynonymous mutations to synonymous mutations (d_N/d_S) for each NUMT cluster. We used MAFFT v7.310 (Katoh and Standley 2013) to align protein coding sequences identified in cNUMTs and reconstructed the coding sequence trees with IQ-TREE v2.0.3 (Minh et al. 2020). Then we used CODEML implemented in PAML (Yang 2007) to calculate pairwise d_N/d_S comparisons and substitutions per codon on all the protein coding genes from cNUMTs which corresponding to the same mitochondrial protein coding genes. The coding sequence pairs with substitutions per codon less than 0.1 were examined for their best hits against the mitogenomes. The pairs where both coding sequences had a percentage of identical matches less than 98% were examined if any of the coding sequences matched with coding sequences in other pairs. The coding sequences matched into groups were grouped into clusters.

We then translated the mitochondrial protein-coding genes from the species we used to identify NUMTs with Biopython (Cock et al. 2009) and aligned each mitochondrial protein-coding gene with MAFFT v7.310 (Katoh and Standley 2013). The translated sequences with lengths 20% shorter than the longest sequences in the alignment were removed to exclude incomplete mitochondrial protein sequences. The alignments of corresponding mitochondrial protein-coding genes were used to align with each dcNUMT clusters generated in the previous step with prank v.170427 (Löytynoja 2014). The alignments were used to reconstruct coding sequence trees with FastTree v2.1.11 (Price, Dehal, and Arkin 2010). In order to increase computational efficiency, we extracted the subtrees with the clades containing cNUMT genes and 3 hierarchical levels above the focal clades using Newick Utilities v1.7.0 (Junier and Zdobnov 2010). The codon alignments of the corresponding subtrees were converted from mitochondrial gene sequences and their translated protein sequences using pal2nal v14 (Suyama, Torrents, and Bork 2006). Then we used CODEML implemented in PAML (Yang 2007) for the selection analysis. We implemented two different branch models: a three-ratio model (model = 2, NSsites = 0) and a three-ratio model with the omega of NUMT clades fixed to 1 (model = 2, NSsites = 0, fix_omega = 1, omega=1), for each cNUMT gene cluster. The likelihood differences among the models were used to derive statistical significance. Benjamin-Hochberg adjustments (Benjamini and Hochberg 1995) were applied for multiple testing corrections.

Hetereogenous substitution models

We also conducted a set of analysis to test for the effect of heterogeneous versus homogeneous substitution models. Potential differences in mitochondrial and nuclear DNA might be reflected in the transition to transversion ratio (ts/tv, κ) which is much higher in mitochondrial DNA relative to nuclear DNA in humans, although there is lots of variation across animals (Belle et al. 2005). While codeml does not incorporate heterogeneity in κ or codon composition, it is possible to conduct sequence simulation using branch specific models of sequence evolution (Fletcher and Yang 2009). For this we performed simulations with INDELible and used estimated κ and codon frequencies from our data (i.e. cluster depicted in Figure 6, with NUMT and mitochondrial specific codon composition and κ estimated separately) as input parameters. We also simulated a homogenous model where we assumed codon composition and κ from the mitochondrial sequences only. We then tested these simulated sequences using codeml assuming homogeneous codon composition and a single κ across branches similar to our test setup (Figure S2 and Table S1). The INDELIBLE input file control.txt is provided as Supplemental code and available on github.

Software availability

Scripts are available at github https://github.com/chnyuch/numt_vertebrate (NUMT detection) and https://github.com/tgossmann/numt_vertebrate (INDELIBLE simulations) and as Supplemental zip file.

Competing Interest Statement

The authors declare that they have no competing interests related to this study.

Acknowledgments

We sincerely thank Justin Wilcox for his valuable input and insightful contributions to this project. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement No. 947636. This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A) and CeBiTec computer cluster at Bielefeld University.

Author contributions

TIG designed the study with input from JIH and DV. YC conducted the majority of the analyses. MH, LH, CRC, DV and TIG contributed to analyses. TIG and YC drafted the manuscript with input from all of the authors. All authors contributed to finalizing the manuscript.

References

Acker, Zoë P. Van, Anika Perdok, Ruben Hellemans, Katherine North, Inge Vorsters, Cedric Cappel, Jonas Dehairs, et al. 2023. "Phospholipase D3 Degrades Mitochondrial DNA to Regulate Nucleotide Signaling and APP Metabolism." *Nature Communications* 14 (1). https://doi.org/10.1038/s41467-023-38501-w.

Baltazar-Soares, Miguel, Patrik Karell, Dominic Wright, Jan-Åke Nilsson, and Jon E. Brommer. 2023. "Bringing to Light Nuclear-Mitochondrial Insertions in the Genomes of Nocturnal Predatory Birds." *Molecular Phylogenetics and Evolution* 181 (April): 107722. https://doi.org/10.1016/j.ympev.2023.107722.

Bank, Claudia, Gregory B. Ewing, Anna Ferrer-Admettla, Matthieu Foll, and Jeffrey D. Jensen. 2014. "Thinking Too Positive? Revisiting Current Methods of Population Genetic Selection Inference." *Trends in Genetics* 30 (12): 540–46. https://doi.org/10.1016/j.tig.2014.09.010.

Barton, N. H. 2000. "Genetic Hitchhiking." Edited by B. Charlesworth and P. H. Harvey. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 355 (1403): 1553–62. https://doi.org/10.1098/rstb.2000.0716.

Baxter, Adrianne L., Jay L. Vivian, R. Tanner Hagelstrom, Waheeda Hossain, Wendy L. Golden, E. Robert Wassman, Rena J. Vanzo, and Merlin G. Butler. 2017. "A Novel Partial Duplication of ZEB2 and Review of ZEB2 Involvement in Mowat-Wilson Syndrome." *Molecular Syndromology* 8 (4): 211–18. https://doi.org/10.1159/000473693.

Belle, Elise M. S., Gwenael Piganeau, Mike Gardner, and Adam Eyre-Walker. 2005. "An Investigation of the Variation in the Transition Bias Among Various Animal Mitochondrial DNA." *Gene* 355 (August): 58–66. https://doi.org/10.1016/j.gene.2005.05.019.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300. https://doi.org/https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Biró, Bálint, Zoltán Gál, Giuseppina Schiavo, Anisa Ribari, Valerio Joe Utzeri, Michael Brookman, Luca Fontanesi, and Orsolya Ivett Hoffmann. 2022. "Nuclear Mitochondrial DNA Sequences in the Rabbit Genome." *Mitochondrion* 66 (September): 1–6. https://doi.org/10.1016/j.mito.2022.07.003.

Bolisetty, Mohan, Jonas Blomberg, Farid Benachenhou, Göran Sperber, and Karen Beemon. 2012. "Unexpected Diversity and Expression of Avian Endogenous Retroviruses." Edited by Stephen P. Goff. *mBio* 3 (5). https://doi.org/10.1128/mbio.00344-12.

Bomblies, Kirsten, and Catherine L. Peichel. 2022. "Genetics of Adaptation." *Proceedings of the National Academy of Sciences* 119 (30). https://doi.org/10.1073/pnas.2122152119.

Butenko, Anzhelika, Julius Lukeš, Dave Speijer, and Jeremy G. Wideman. 2024. "Mitochondrial Genomes Revisited: Why Do Different Lineages Retain Different Genes?" *BMC Biology* 22 (1). https://doi.org/10.1186/s12915-024-01824-1.

Calabrese, F. M., D. L. Balacco, R. Preste, M. A. Diroma, R. Forino, M. Ventura, and M. Attimonelli. 2017. "NumtS Colonization in Mammalian Genomes." *Scientific Reports* 7 (1). https://doi.org/

503 10.1038/s41598-017-16750-2.
504 Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin
505 Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC*
506 *Bioinformatics* 10 (421): 1–9. https://doi.org/10.1186/1471-2105-10-421.
507 Caswell, Jacob, Jason D. Gans, Nicholas Generous, Corey M. Hudson, Eric Merkley, Curtis Johnson,
508 Christopher Oehmen, et al. 2019. "Defending Our Public Biological Databases as a Global
509 Critical Infrastructure." *Frontiers in Bioengineering and Biotechnology* 7 (April). https://doi.or
510 g/10.3389/fbioe.2019.00058.
511 Chowdhury, Sayam U., Mohammod Foysal, and Rhys E. Green. 2022. "Accelerating Decline of an
512 Important Wintering Population of the Critically Endangered Spoon-billed Sandpiper *Calidris*
513 *Pygmaea* at Sonadia Island, Bangladesh." *Journal of Ornithology* 163 (4): 891–901. https:
514 //doi.org/10.1007/s10336-022-01995-0.
515 Clark, Alice, Gizem Koc, Ying Eyre-Walker, and Adam Eyre-Walker. 2023. "What Determines
516 Levels of Mitochondrial Genetic Diversity in Birds?" Edited by Aida Andres. *Genome Biology*
517 *and Evolution* 15 (5). https://doi.org/10.1093/gbe/evad064.
518 Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew
519 Dalke, Iddo Friedberg, et al. 2009. "Biopython: Freely Available Python Tools for Computational
520 Molecular Biology and Bioinformatics." *Bioinformatics* 25 (11): 1422–23. https://doi.org/10.1
521 093/bioinformatics/btp163.
522 Cohen, Tal, Liron Levin, and Dan Mishmar. 2016. "Ancient Out-of-Africa Mitochondrial DNA
523 Variants Associate with Distinct Mitochondrial Gene Expression Patterns." Edited by Vivian G.
524 Cheung. *PLOS Genetics* 12 (11): e1006407. https://doi.org/10.1371/journal.pgen.1006407.
525 Danchin, Etienne G. J. 2016. "Lateral Gene Transfer in Eukaryotes: Tip of the Iceberg or of the Ice
526 Cube?" *BMC Biology* 14 (1). https://doi.org/10.1186/s12915-016-0330-x.
527 Dayama, Gargi, Sarah B. Emery, Jeffrey M. Kidd, and Ryan E. Mills. 2014. "The Genomic
528 Landscape of Polymorphic Human Nuclear Mitochondrial Insertions." *Nucleic Acids Research*
529 42 (20): 12640–49. https://doi.org/10.1093/nar/gku1038.
530 Diehl, William E, Nirali Patel, Kate Halm, and Welkin E Johnson. 2016. "Tracking Interspecies
531 Transmission and Long-Term Evolution of an Ancient Retrovirus Using the Genomes of Modern
532 Mammals." *eLife* 5 (March). https://doi.org/10.7554/elife.12704.
533 Donath, Alexander, Frank Jühling, Marwa Al-Arab, Stephan H. Bernhart, Franziska Reinhardt,
534 Peter F. Stadler, Martin Middendorf, and Matthias Bernt. 2019. "Improved Annotation of
535 Protein-Coding Genes Boundaries in Metazoan Mitochondrial Genomes." *Nucleic Acids Research*
536 47 (20): 10543–52. https://doi.org/10.1093/nar/gkz833.
537 Dunning, Luke T., Jill K. Olofsson, Christian Parisod, Rimjhim Roy Choudhury, Jose J.
538 Moreno-Villena, Yang Yang, Jacqueline Dionora, et al. 2019. "Lateral Transfers of Large DNA
539 Fragments Spread Functional Genes Among Grasses." *Proceedings of the National Academy of*
540 *Sciences* 116 (10): 4416–25. https://doi.org/10.1073/pnas.1810031116.
541 Flavell, Andrew J. 1999. "Long Terminal Repeat Retrotransposons Jump Between Species."
542 *Proceedings of the National Academy of Sciences* 96 (22): 12211–12. https://doi.org/10.107
543 3/pnas.96.22.12211.

Fletcher, W., and Z. Yang. 2009. "INDELible: A Flexible Simulator of Biological Sequence Evolution." *Molecular Biology and Evolution* 26 (8): 1879–88. https://doi.org/10.1093/molbev/msp098.

Gabaldón, Toni. 2020. "Patterns and Impacts of Nonvertical Evolution in Eukaryotes: A Paradigm Shift." *Annals of the New York Academy of Sciences* 1476 (1): 78–92. https://doi.org/10.1111/nyas.14471.

Gao, Fan, Ye Cai, Philipp Kapranov, and Dongyang Xu. 2020. "Reverse-Genetics Studies of lncRNAs—what We Have Learnt and Paths Forward." *Genome Biology* 21 (1). https://doi.org/10.1186/s13059-020-01994-5.

Gossmann, Toni I., Anna W. Santure, Ben C. Sheldon, Jon Slate, and Kai Zeng. 2014. "Highly Variable Recombinational Landscape Modulates Efficacy of Natural Selection in Birds." *Genome Biology and Evolution* 6 (8): 2061–75. https://doi.org/10.1093/gbe/evu157.

Gossmann, Toni I., Achchuthan Shanmugasundram, Stefan Börno, Ludovic Duvaux, Christophe Lemaire, Heiner Kuhl, Sven Klages, et al. 2019. "Ice-Age Climate Adaptations Trap the Alpine Marmot in a State of Low Genetic Diversity." *Current Biology* 29 (10): 1712–1720.e7. https://doi.org/10.1016/j.cub.2019.04.020.

Goyal, Manisha, Mohammed Faruq, Ashok Gupta, Divya Shrivastava, and Uzma Shamim. 2022. "Deletion of 2q22.2q22.3 in Mowat–wilson Syndrome: A Case Report and Review of the Literature." *Journal of Pediatric Neurology* 20 (06): 440–44. https://doi.org/10.1055/s-0042-1749670.

Green, Richard E., Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, et al. 2010. "A Draft Sequence of the Neandertal Genome." *Science* 328 (5979): 710–22. https://doi.org/10.1126/science.1188021.

Hazkani-Covo, Einat. 2022. "A Burst of Numt Insertion in the Dasyuridae Family During Marsupial Evolution." *Frontiers in Ecology and Evolution* 10 (844443). https://doi.org/10.3389/fevo.2022.844443.

Hazkani-Covo, Einat, Raymond M. Zeller, and William Martin. 2010. "Molecular Poltergeists: Mitochondrial DNA Copies (Numts) in Sequenced Nuclear Genomes." Edited by Harmit S. Malik. *PLoS Genetics* 6 (2): e1000834. https://doi.org/10.1371/journal.pgen.1000834.

Hu, Gengxi, and William G. Thilly. 1994. "Evolutionary Trail of the Mitochondrial Genome as Based on Human 16S rDNA Pseudogenes." *Gene* 147 (2): 197–204. https://doi.org/10.1016/0378-1119(94)90065-5.

Huang, Wen, Richard F Lyman, Rachel A Lyman, Mary Anna Carbone, Susan T Harbison, Michael M Magwire, and Trudy FC Mackay. 2016. "Spontaneous Mutations and the Origin and Maintenance of Quantitative Genetic Variation." *eLife* 5 (May). https://doi.org/10.7554/elife.14625.

Iacolina, Laura, Luca Corlatti, Elena Buzan, Toni Safner, and Nikica Šprem. 2018. "Hybridisation in European Ungulates: An Overview of the Current Status, Causes, and Consequences." *Mammal Review* 49 (1): 45–59. https://doi.org/10.1111/mam.12140.

Javaheri Tehrani, Sahar, Laura Kvist, Omid Mirshamsi, Seyed Mahmoud Ghasempouri, and Mansour Aliabadian. 2021. "Genetic Divergence, Admixture and Subspecific Boundaries in a Peripheral

585 Population of the Great Tit, Parus Major (Aves: Paridae)." *Biological Journal of the Linnean*
586 *Society* 133 (4): 1084–98. https://doi.org/10.1093/biolinnean/blab064.
587 Jin, Jian-Jun, Wen-Bin Yu, Jun-Bo Yang, Yu Song, Claude W. dePamphilis, Ting-Shuang Yi, and
588 De-Zhu Li. 2020. "GetOrganelle: A Fast and Versatile Toolkit for Accurate de Novo Assembly
589 of Organelle Genomes." *Genome Biology* 21 (1). https://doi.org/10.1186/s13059-020-02154-5.
590 Junier, Thomas, and Evgeny M. Zdobnov. 2010. "The Newick utilities: high-throughput
591 phylogenetic tree processing in the Unix shell." *Bioinformatics* 26 (13): 1669–70. https:
592 //doi.org/10.1093/bioinformatics/btq243.
593 Kapusta, Aurélie, Alexander Suh, and Cédric Feschotte. 2017. "Dynamics of Genome Size Evolution
594 in Birds and Mammals." *Proceedings of the National Academy of Sciences* 114 (8). https:
595 //doi.org/10.1073/pnas.1616702114.
596 Katoh, Kazutaka, and Daron M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software
597 Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30 (4):
598 772–80. https://doi.org/10.1093/molbev/mst010.
599 Kinsella, Cormac M., Francisco J. Ruiz-Ruano, Anne-Marie Dion-Côté, Alexander J. Charles, Toni
600 I. Gossmann, Josefa Cabrero, Dennis Kappei, et al. 2019. "Programmed DNA Elimination
601 of Germline Development Genes in Songbirds." *Nature Communications* 10 (1): 1–10. https:
602 //doi.org/10.1038/s41467-019-13427-4.
603 Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges. 2017. "TimeTree: a resource for
604 timelines, timetrees, and divergence times." *Molecular Biology and Evolution* 34 (7): 1812–19.
605 https://doi.org/10.1093/molbev/msx116.
606 Laine, Veronika N., Toni I. Gossmann, Kees van Oers, Marcel E. Visser, and Martien A. M. Groenen.
607 2019. "Exploring the Unmapped DNA and RNA Reads in a Songbird Genome." *BMC Genomics*
608 20 (1). https://doi.org/10.1186/s12864-018-5378-2.
609 Laine, Veronika N., Toni I. Gossmann, Kyle M. Schachtschneider, Colin J. Garroway, Ole Madsen,
610 Koen J. F. Verhoeven, Victor de Jager, et al. 2016. "Evolutionary Signals of Selection on
611 Cognition from the Great Tit Genome and Methylome." *Nature Communications* 7 (1). https:
612 //doi.org/10.1038/ncomms10474.
613 Lammers, Fritjof, Axel Janke, Cornelia Rücklé, Vera Zizka, and Maria A. Nilsson. 2017. "Screening
614 for the Ancient Polar Bear Mitochondrial Genome Reveals Low Integration of Mitochondrial
615 Pseudogenes (Numts) in Bears." *Mitochondrial DNA Part B* 2 (1): 251–54. https://doi.org/10
616 .1080/23802359.2017.1318673.
617 Lee, Michael S. Y. 1998. "Ancestors and Taxonomy." *Trends in Ecology & Evolution* 13 (1): 26.
618 https://doi.org/10.1016/s0169-5347(97)01272-x.
619 Liang, Bin, Ning Wang, Nan Li, Rebecca T. Kimball, and Edward L. Braun. 2018. "Comparative
620 Genomics Reveals a Burst of Homoplasy-Free Numt Insertions." Edited by Anne Yoder.
621 *Molecular Biology and Evolution* 35 (8): 2060–64. https://doi.org/10.1093/molbev/msy112.
622 Lopez, Jose, Melanie Culver, J. Claiborne Stephens, Warren E. Johnson, and Warren Stephen J.
623 O'brien. 1997. "Rates of Nuclear and Cytoplasmic Mitochondrial DNA Sequence Divergence in
624 Mammals." *Molecular Biology and Evolution* 14 (3): 277–86. https://doi.org/10.1093/oxfordjo
625 urnals.molbev.a025763.

Lopez, Jose, Naoya Yuhki, Ryuichi Masuda, William Modi, and Stephen O'brien. 1994. "Numt, a Recent Transfer and Tandem Amplification of Mitochondrial DNA to the Nuclear Genome of the Domestic Cat." *Journal of Molecular Evolution* 39: 174–90. https://doi.org/10.1007/BF00163806.

Löytynoja, Ari. 2014. "Phylogeny-Aware Alignment with PRANK." In *Multiple Sequence Alignment Methods*, edited by David J. Russell, 155–70. Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-62703-646-7_10.

Lu, Hengyun, Francesca Giordano, and Zemin Ning. 2016. "Oxford Nanopore MinION Sequencing and Genome Assembly." *Genomics, Proteomics & Bioinformatics* 14 (5): 265–79. https://doi.org/10.1016/j.gpb.2016.05.004.

Lucas, Torres, Bretagnolle Vincent, and Pante Eric. 2022. "Translocation of Mitochondrial DNA into the Nuclear Genome Blurs Phylogeographic and Conservation Genetic Studies in Seabirds." *Royal Society Open Science* 9 (6). https://doi.org/10.1098/rsos.211888.

Mallick, Swapan, Sante Gnerre, Paul Muller, and David Reich. 2009. "The Difficulty of Avoiding False Positives in Genome Scans for Natural Selection." *Genome Research* 19 (5): 922–33. https://doi.org/10.1101/gr.086512.108.

Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. "IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era." *Molecular Biology and Evolution* 37 (5): 1530–34. https://doi.org/10.1093/molbev/msaa015.

Mirdita, Milot, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. 2022. "ColabFold: Making Protein Folding Accessible to All." *Nature Methods* 19 (6): 679–82. https://doi.org/10.1038/s41592-022-01488-1.

Nacer, Deborah F., and Fabio Raposo do Amaral. 2017. "Striking Pseudogenization in Avian Phylogenetics: Numts Are Large and Common in Falcons." *Molecular Phylogenetics and Evolution* 115 (October): 1–6. https://doi.org/10.1016/j.ympev.2017.07.002.

Nei, Masatoshi, and Takashi Gojobori. 1986. "Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions." *Molecular Biology and Evolution* 3 (5): 418–26. https://doi.org/10.1093/oxfordjournals.molbev.a040410.

Noutsos, Christos, Tatjana Kleine, Ute Armbruster, Giovanni Dalcorso, and Dario Leister. 2007. "Nuclear Insertions of Organellar DNA Can Create Novel Patches of Functional Exon Sequences." *Trends in Genetics* 23 (12): 597–601. https://doi.org/10.1016/j.tig.2007.08.016.

Pagel, M., and A. Meade. 2022. "BayesTraits, version 4." University of Reading, Berkshire, UK.

Peona, Valentina, Octavio M. Palacios-Gimenez, Julie Blommaert, Jing Liu, Tri Haryoko, Knud A. Jønsson, Martin Irestedt, Qi Zhou, Patric Jern, and Alexander Suh. 2021. "The Avian w Chromosome Is a Refugium for Endogenous Retroviruses with Likely Effects on Female-Biased Mutational Load and Genetic Incompatibilities." *Philosophical Transactions of the Royal Society B: Biological Sciences* 376 (1833): 20200186. https://doi.org/10.1098/rstb.2020.0186.

Plazzi, Federico, Youn Le Cras, Alessandro Formaggioni, and Marco Passamonti. 2023. "Mitochondrially Mediated RNA Interference, a Retrograde Signaling System Affecting Nuclear Gene Expression." *Heredity* 132 (3): 156–61. https://doi.org/10.1038/s41437-023-00650-5.

667 Popadin, Konstantin, Konstantin Gunbin, Leonid Peshkin, Sofia Annis, Zoe Fleischmann, Melissa
668 Franco, Yevgenya Kraytsberg, Natalya Markuzon, Rebecca R. Ackermann, and Konstantin
669 Khrapko. 2022. "Mitochondrial Pseudogenes Suggest Repeated Inter-Species Hybridization
670 Among Direct Human Ancestors." *Genes* 13 (5). https://doi.org/10.3390/genes13050810.
671 Pozzi, Andrea, and Damian K. Dowling. 2019. "The Genomic Origins of Small Mitochondrial RNAs:
672 Are They Transcribed by the Mitochondrial DNA or by Mitochondrial Pseudogenes Within the
673 Nucleus (NUMTs)?" *Genome Biology and Evolution* 11 (7): 1883–96. https://doi.org/10.1093/
674 gbe/evz132.
675 Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. "FastTree 2 – Approximately
676 Maximum-Likelihood Trees for Large Alignments." *PloS One* 5 (3): e9490. https://doi.org/10.1
677 371/journal.pone.0009490.
678 Prüfer, Kay, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer,
679 Anja Heinze, et al. 2013. "The Complete Genome Sequence of a Neanderthal from the Altai
680 Mountains." *Nature* 505 (7481): 43–49. https://doi.org/10.1038/nature12886.
681 Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing
682 Genomic Features." *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq
683 033.
684 Racimo, Fernando. 2015. "Testing for Ancient Selection Using Cross-Population Allele Frequency
685 Differentiation." *Genetics* 202 (2): 733–50. https://doi.org/10.1534/genetics.115.178095.
686 Revell, Liam J. 2012. "Phytools: An R Package for Phylogenetic Comparative Biology (and Other
687 Things)." *Methods in Ecology and Evolution* 3 (2): 217–23. https://doi.org/https://doi.org/10.1
688 111/j.2041-210X.2011.00169.x.
689 Rhoads, Anthony, and Kin Fai Au. 2015. "PacBio Sequencing and Its Applications." *Genomics,*
690 *Proteomics & Bioinformatics* 13 (5): 278–89. https://doi.org/10.1016/j.gpb.2015.08.002.
691 Ricchetti, Miria, Cécile Fairhead, and Bernard Dujon. 1999. "Mitochondrial DNA Repairs
692 Doublestrand Breaks in Yeast Chromosomes." *Nature* 402: 96–100. https://doi.org/10.103
693 8/47076.
694 Ricchetti, Miria, Fredj Tekaia, and Bernard Dujon. 2004. "Continued Colonization of the Human
695 Genome by Mitochondrial DNA." Edited by Jonathan A. Eisen. *PLoS Biology* 2 (9): e273.
696 https://doi.org/10.1371/journal.pbio.0020273.
697 Rice, Peter, Ian Longden, and Alan Bleasby. 2000. "EMBOSS: The European Molecular Biology
698 Open Software Suite." *Trends in Genetics* 16 (6): 276–77. https://doi.org/10.1016/s0168-
699 9525(00)02024-2.
700 Richly, Erik, and Dario Leister. 2022. "NUMTs in Sequenced Eukaryotic Genomes." *Molecular*
701 *Biology and Evolution* 21 (6): 1081–84. https://doi.org/10.1093/molbev/msh110.
702 Sangster, George, and Jolanda A Luksenburg. 2021. "Sharp Increase of Problematic Mitogenomes
703 of Birds: Causes, Consequences, and Remedies." Edited by Liliana Milani. *Genome Biology and*
704 *Evolution* 13 (9). https://doi.org/10.1093/gbe/evab210.
705 Sayers, Eric W., Mark Cavanaugh, Karen Clark, Kim D. Pruitt, Conrad L. Schoch, Stephen T.
706 Sherry, and Ilene Karsch-Mizrachi. 2021. "GenBank." *Nucleic Acids Research* 49 (D1): D92–96.
707 https://doi.org/10.1093/nar/gkaa1023.

Schiavo, Giuseppina, Orsolya Ivett Hoffmann, Anisa Ribani, Valerio Joe Utzeri, Marco Ciro Ghionda, Francesca Bertolini, Claudia Geraci, Samuele Bovo, and Luca Fontanesi. 2017. "A Genomic Landscape of Mitochondrial DNA Insertions in the Pig Nuclear Genome Provides Evolutionary Signatures of Interspecies Admixture." *DNA Research* 24 (5): 487–98. https://doi.org/10.1093/dnares/dsx019.

Setter, Derek, Sylvain Mousset, Xiaoheng Cheng, Rasmus Nielsen, Michael DeGiorgio, and Joachim Hermisson. 2020. "VolcanoFinder: Genomic Scans for Adaptive Introgression." Edited by Alex Buerkle. *PLOS Genetics* 16 (6): e1008867. https://doi.org/10.1371/journal.pgen.1008867.

Sin, Simon Yung Wa, Lily Lu, and Scott V. Edwards. 2020. "De Novo Assembly of the Northern Cardinal (Cardinalis cardinalis) Genome Reveals Candidate Regulatory Regions for Sexually Dichromatic Red Plumage Coloration." *G3 Genes|Genomes|Genetics* 10 (10): 3541–48. https://doi.org/10.1534/g3.120.401373.

Song, Hojun, Matthew J. Moulton, Kevin D. Hiatt, and Michael F. Whiting. 2013. "Uncovering Historical Signature of Mitochondrial DNA Hidden in the Nuclear Genome: The Biogeography of Schistocerca Revisited." *Cladistics* 29 (6): 643–62. https://doi.org/10.1111/cla.12013.

Soto-Calderón, Iván Darío, Nicholas Jonathan Clark, Julia Vera Halo Wildschutte, Kelly DiMattio, Michael Ignatius Jensen-Seaman, and Nicola Mary Anthony. 2014. "Identification of Species-Specific Nuclear Insertions of Mitochondrial DNA (Numts) in Gorillas and Their Potential as Population Genetic Markers." *Molecular Phylogenetics and Evolution* 81 (September): 61–70. https://doi.org/10.1016/j.ympev.2014.08.018.

Suyama, Mikita, David Torrents, and Peer Bork. 2006. "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments." *Nucleic Acids Research* 34 (suppl_2): W609–12. https://doi.org/10.1093/nar/gkl315.

Tsuji, Junko, Martin C. Frith, Kentaro Tomii, and Paul Horton. 2012. "Mammalian NUMT Insertion Is Non-Random." *Nucleic Acids Research* 40 (18): 9073–88. https://doi.org/10.1093/nar/gks424.

Uvizl, Marek, Sebastien J. Puechmaille, Sarahjane Power, Martin Pippel, Samuel Carthy, Wilfried Haerty, Eugene W. Myers, Emma C. Teeling, and Zixia Huang. 2023. "Comparative Genome Microsynteny Illuminates the Fast Evolution of Nuclear Mitochondrial Segments (NUMTs) in Mammals," April. https://doi.org/10.1101/2023.04.20.537758.

Vendrami, David L. J., Toni I. Gossmann, Nayden Chakarov, Anneke J. Paijmans, Vivienne Litzke, Adam Eyre-Walker, Jaume Forcada, and Joseph I. Hoffman. 2022. "Signatures of Selection on Mitonuclear Integrated Genes Uncover Hidden Mitogenomic Variation in Fur Seals." Edited by Andrea Betancourt. *Genome Biology and Evolution* 14 (7). https://doi.org/10.1093/gbe/evac104.

Wang, Dapeng, Yubin Zhang, Zhang Zhang, Jiang Zhu, and JunYu. 2010. "KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies." *Genomics, Proteomics Bioinformatics* 8 (1): 77–80. https://doi.org/10.1016/S1672-0229(10)60008-3.

Wei, W., and P. F. Chinnery. 2020. "Inheritance of Mitochondrial DNA in Humans: Implications for Rare and Common Diseases." *Journal of Internal Medicine* 287 (6): 634–44. https://doi.org/10.1111/joim.13047.

Wei, Wei, Katherine R. Schon, Greg Elgar, Andrea Orioli, Melanie Tanguy, Adam Giess,

Marc Tischkowitz, Mark J. Caulfield, and Patrick F. Chinnery. 2022. "Nuclear-Embedded Mitochondrial DNA Sequences in 66,083 Human Genomes." *Nature* 611 (7934): 105–14. https://doi.org/10.1038/s41586-022-05288-7.

Wilcox, Justin J S, Barbara Arca-Ruibal, Jaime Samour, Victor Mateuta, Youssef Idaghdour, and Stéphane Boissinot. 2022. "Linked-Read Sequencing of Eight Falcons Reveals a Unique Genomic Architecture in Flux." Edited by Adam Eyre-Walker. *Genome Biology and Evolution* 14 (6). https://doi.org/10.1093/gbe/evac090.

Yang, Ziheng. 2007. "PAML 4: Phylogenetic analysis by maximum likelihood." *Molecular Biology and Evolution* 24 (8): 1586–91. https://doi.org/10.1093/molbev/msm088.

Figures

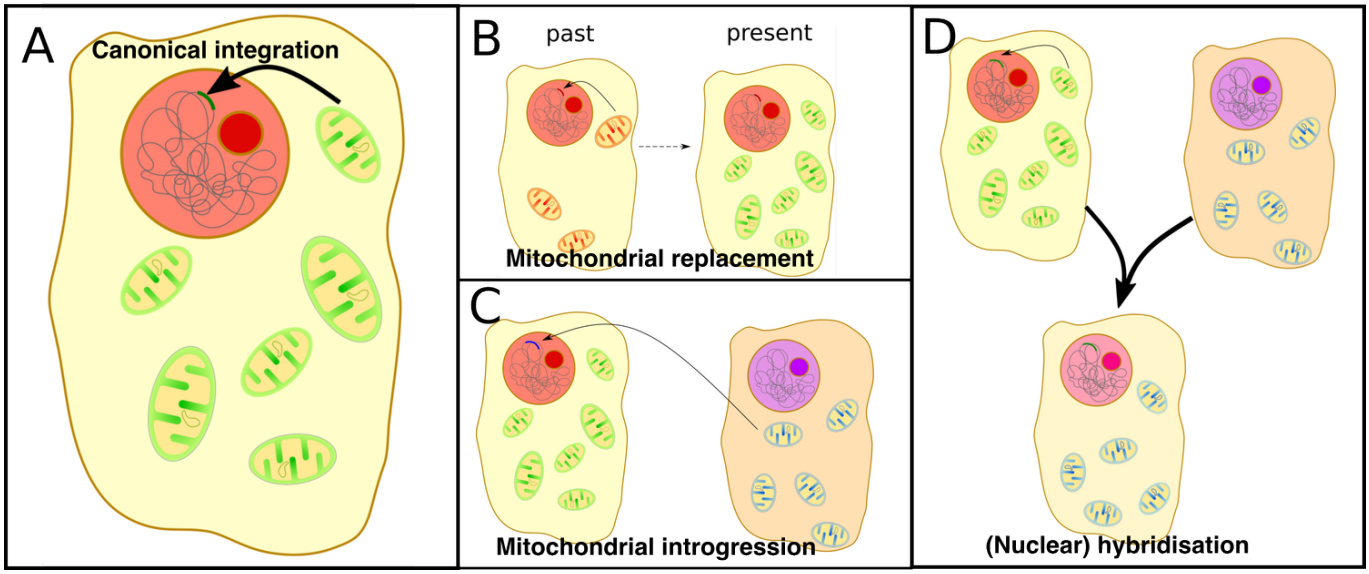

Figure 1: Potential sources of NUMTs. **(A)** Canonical integration through host species' mitochondrial genome, **(B)** Integration of the host species' mitochondrial genome stemming from mitochondrial replacement such as a consequence of hybridisation (Javaheri Tehrani et al. 2021), **(C)** Integration of the mitochondrial DNA of another species, e.g. through a retrovirus (Tsuji et al. 2012) **(D)** Integration of nuclear DNA from another species, e.g through nuclear hybrid genome formation (Popadin et al. 2022), that happens to carry a NUMT.

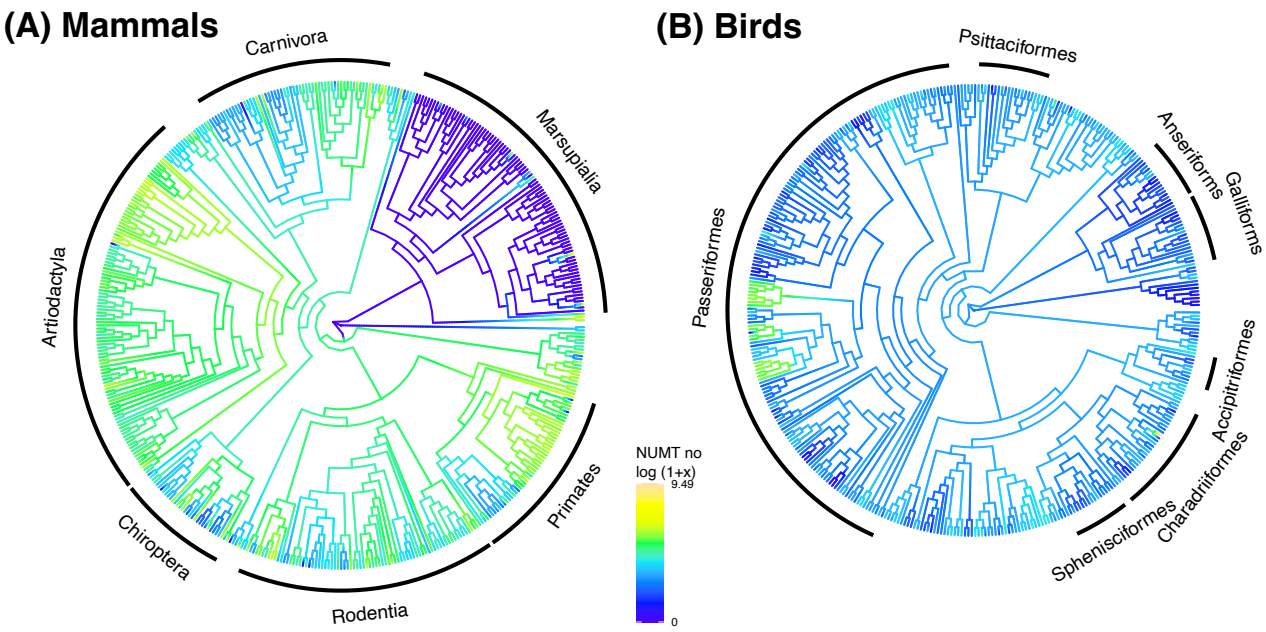

Figure 2: Phylogenetic distribution of NUMT occurences in mammals and birds. Shown are the species cladograms from timetree, with the colours representing reconstructed number of NUMTs for **(A)** mammals and **(B)** birds, respectively. Note that the colorised representation refers to the log-transformed (i.e. *log(1+x)*) NUMTs number.

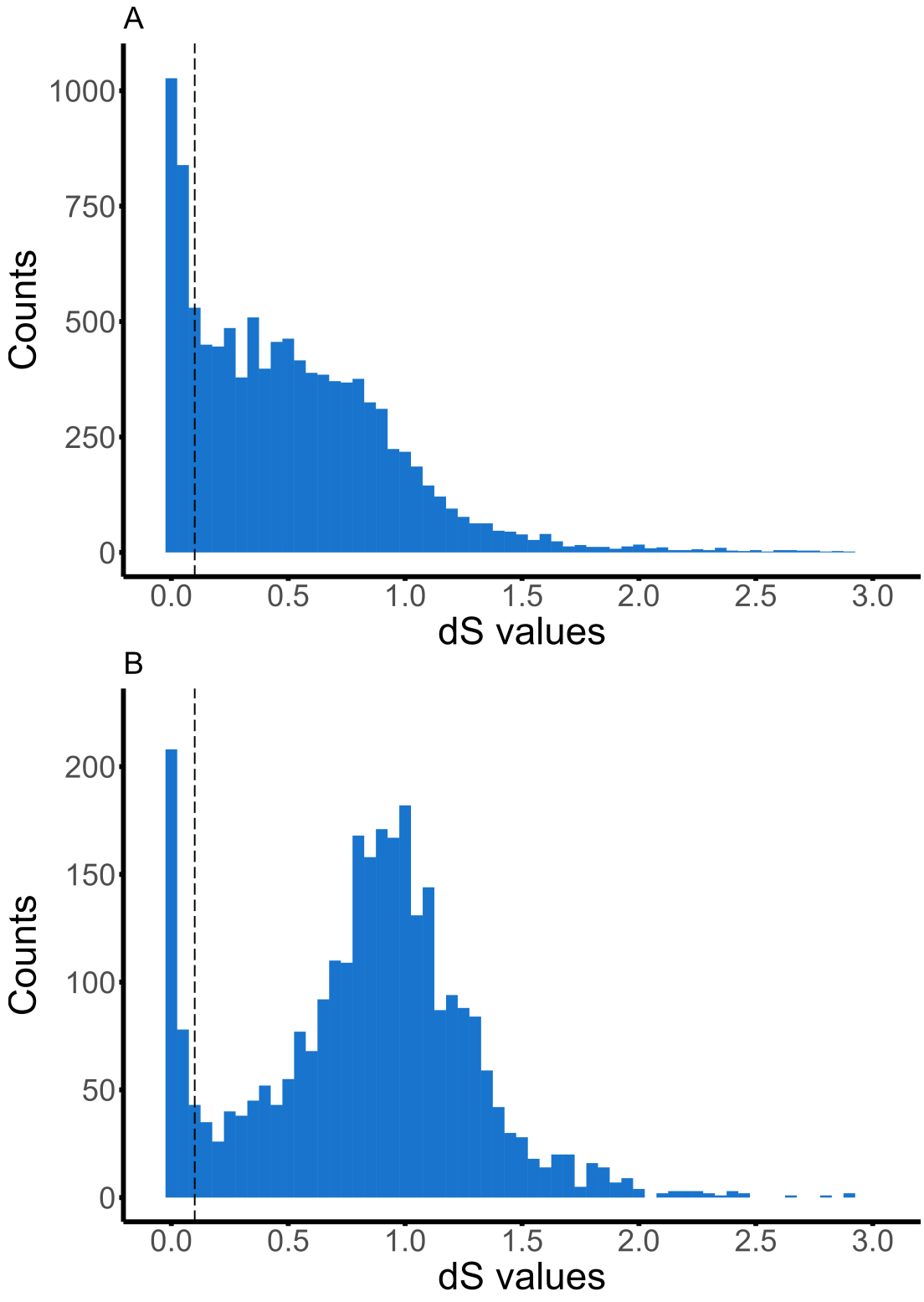

Figure 3: Distribution of dS values of genes in cNUMTs. Shown are the distribution of dS values between genes in the NUMTs and the host species mitochondrial DNA. NUMTs that possess at least one genes with $dS > 0.1$ are considered dcNUMTs, $dS = 0.1$ is indicated by a dashed line. Distributions are shown for **(A)** mammals and **(B)** birds. Note that count data for dS values >3 are not shown.

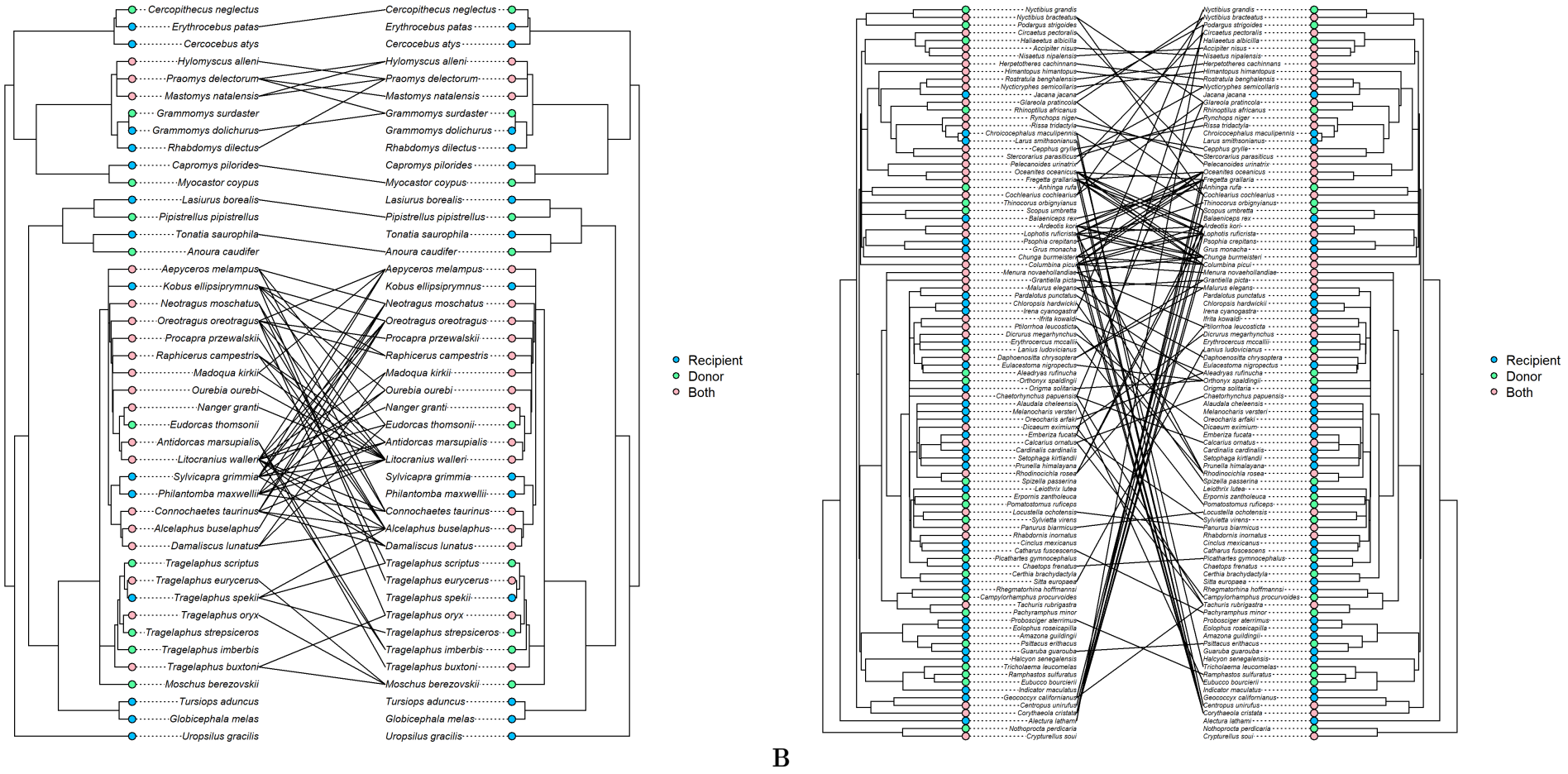

Figure 4: Phylogenetic pattern of NUMT integrations of non-host mitochondrial DNA. For **(A)** mammals and **(B)** birds. Shown are species that either show evidence of NUMTs of non-host origin or that appear be the donor species of the NUMT. Note that some species can be either a donor and/or a recipient and that not all species included in our analysis are resolved phylogenetically in Timetree, so that not all pairs can be assigned. A complete list can be found in Tables S2 and S3.

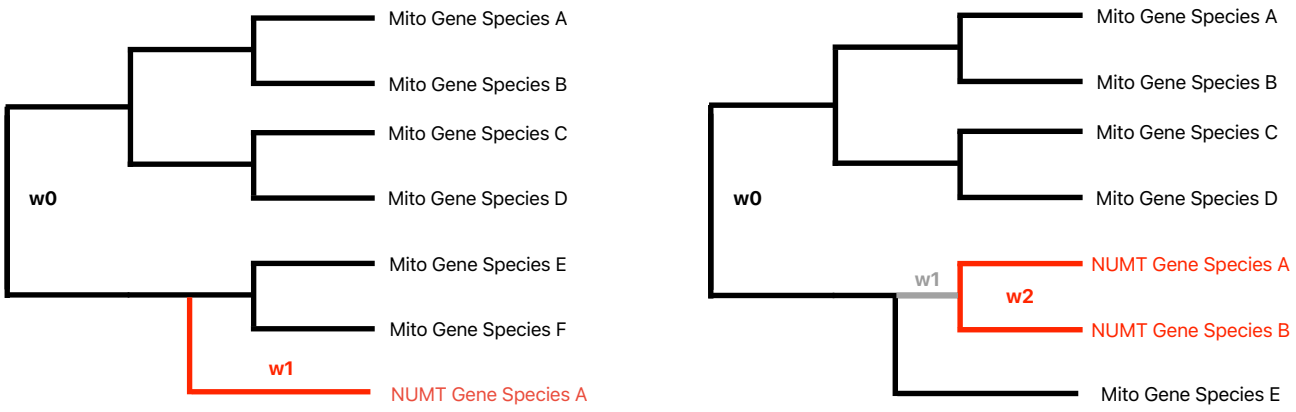

Figure 5: Schematic illustration of the single gene and clustered genes codeml analysis. **(A)** Foreground branch in red (d_N/d_S=ω_1) for a single NUMT gene, background branches (black, d_N/d_S=ω_0) are for the respective mitochondrial genes of related species **(B)** Foreground branch in red for NUMT gene clusters (d_N/d_S=ω_2, here illustrated with cluster size two), background branches (black, d_N/d_S=ω_0 and grey, d_N/d_S=ω_1) are for the respective mitochondrial genes of related species. Note that the grey branch is a seperate background branch, because it denotes the branch when the potential nuclear genomic integration took place.

Figure 6: Phylogenetic tree and sequences used for codeml analysis of a NUMT cluster that contains dcNUMTs of seven ape species along with mitochondrial sequences of 59 (sub)species. The NUMT gene with an intact reading frame is ND4L. Evolutionary pressures were estimated with codeml using the mitochondrial genetic code. The coding sequence is shown in color using the 4-letter-DNA notation illustrating no pinpointed divergence of the NUMT sequences.

Tables

Table 1: Number of BLAST hits of dcNUMTs against available mitochondrial genomes for birds and mammals. The best BLAST hits were required to show at least 95% identity to the query. * For a hit to be assigned as high confidence to a species other than the host, there needed to be a 5% or greater similarity to the non-host, e.g., a highly distinct hit to another species' mitochondrial genome.

Best BLAST hit to (>95% identity, dcNUMT as query)	Mammals	Birds
host species' mitochondrial genome	441	18
non-host species' mitochondrial genome	513	528
with high confidence*	**148**	**249**
no species in the database	3,228	371

Table 2: Evolutionary pressure analysis on single dcNUMT genes for mammals and birds. Shown are the numbers of genes that show branch-specific signatures of purifying selection, neutral evolution or positive selection.

Evolutionary pressure on dcNUMT gene	# of genes in mammals	# of genes in birds
Purifying selection (d_N/d_S<1)	1,337	377
Neutral evolution (d_N/d_S=1)	2,211	1,865
Positive selection (d_N/d_S>1)	1	4

Table 3: Evolutionary pressure analysis on dcNUMT gene clusters for mammals and birds. Shown are the numbers of gene clusters that show clade-specific signatures of purifying selection, neutral evolution and positive selection, respectively.

Evol. pressure on dcNUMT gene clusters	# of clusters mammals	# of clusters in birds
Purifying selection (d_N/d_S<1)	247	24
Neutral evolution (d_N/d_S=1)	726	73
Positive selection (d_N/d_S>1)	23	1

Table 4: Codeml analysis results of a mitocoding gene located on a NUMT that is shared among seven ape species. Rate heterogeneity between the NUMT clade and the mitochondrial sequences was inferred by comparing a two-ratio model with a one-ratio model (single d_N/d_S=ω value for the entire tree). Positive selection was inferred from a two ratio model and a fixed two-ratio model for which the d_N/d_S of the NUMT subtree was fixed to ω_2=1. The tree and sequences used in the analysis are shown in Figure 6.

Model	ln Likelihood	ω values	Comparison
One-ratio	$\ln L_1$=6,946.87	$\omega = 0.06$	
Two-ratio (fixed)	$\ln L_{2a}$=6,905.41	$\omega_1 = 0.05$, $\omega_2 = 1$	
Two-ratio	$\ln L_{2b}$=-6,902.98	$\omega_1 = 0.05$, $\omega_2 = 2.9$	$2\Delta(\ln L_{2a}, \ln L_{2b})$ P=0.027 (χ^2,d.f.=1)

Table 5: Number of NUMTS, coding NUMTs (cNUMTs) and deep coding NUMTs (dcNUMTS) in birds and mammals. As dcNUMTs can have multiple genes, they might group with other dcNUMTs multiple times.

NUMT type	Number in mammals	Number in birds
NUMTS	85,100	28,947
cNUMTS	5,229	1,200
dcNUMTS	4,182	917
dcNUMTS/non-dcNUMTS in groups	2,670/144	157/6
single dcNUMTS	2,024	786
dcNUMTS single+grouped	511	26