Long-term selection on mitocoding genes buried in mammalian and avian nuclear genomes

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

[1] TU Dortmund University, Computational Systems Biology, Faculty of Biochemical and Chemical Engineering, Emil-Figge-Str. 66, 44227 Dortmund, Germany

[2] Bielefeld University, Department of Animal Behaviour, Faculty of Biology, Konsequenz 45, 33615 Bielefeld, Germany

[3] Bielefeld University, Evolutionary Genetics, Faculty of Biology, Konsequenz 45, 33615 Bielefeld, Germany

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

[*]**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 an integration over longer 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 to the host's mitochondrial genome. Two major hypotheses have been put forward to explain the existence of such mitocoding nuclear genes: (A) recent introgression from another species and (B) long-term selection. To address whether these intriguing possibilities do play a role we scanned the genomes of more than 1,000 avian and mammalian species for NUMTs. We indeed identified that the subclass of divergent NUMTs harbouring mitogenes with intact reading frames are widespread across mammals and birds. We can show that for these NUMTs signatures of cross-species introgression are widespread in birds, but not mammals with the exception of ungulates. We can also show that a substantial fraction of deeply divergent NUMTs are maintained by selection. For a small number of NUMT genes we identify an evolutionary signature that is consistent with adaptive evolution including one human NUMT that is shared among seven ape species. This highlights the intriguing possibility that NUMT insertions occasionally may be functional.

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

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 rise and manifestation 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). 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. 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) they were then found in many different eukaryotic species from yeast (Ricchetti, Fairhead, and Dujon 1999) to humans (Lopez et al. 1997) and 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 human 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 it 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 investigated.

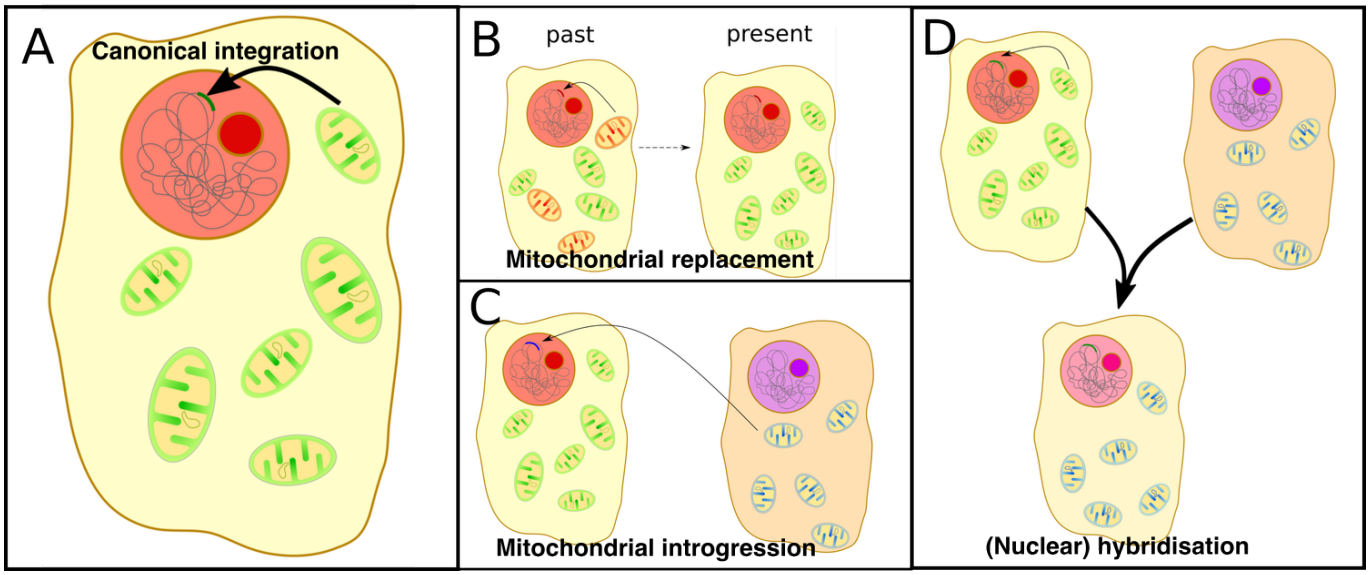


Figure 1: Potential sources of mitochondrial DNA of NUMTs. (A) Canonical integration through host species' mitochondrial genome, (B) Integration of the host species' mitochondrial genome stemming from heteroplasmic DNA, (C) Integration of foreign species mitochondrial DNA, e.g. through retroviruses, (D) Integration of nuclear DNA from another species, e.g through hybrid formation, hat happens to carry a NUMT

To unravel the role of organelle DNA in creating novel and potentially adaptive variation in nuclear genomes, we applied a comparative genomics approach, which we term 'phylonumtomics'. We use this approach to identify NUMTs in more than 1,000 publicly available avian and mammalian genomes and to investigate the evolutionary origins of dcNUMTs. We show that dcNUMTs are widespread across mammalian and avian genomes, with distinct phylogenetic differences between mammals and birds. We conduct comprehensive analyses to uncover traces of selection by incorporating consequences of

low local mutation rates and introgression. We show that in birds, the vast majority of dcNUMTs are reflects of ancient mitochondrial genomes that are preserved as 'genomic fossils'. However, we also highlight numerous instances representing likely signatures of recent introgression, as well as a limited, but highly significant, role of positive selection. Taken together, our results suggest that mitochondria may have a much greater contribution to the evolution of the nuclear genome than previously envisaged.

Results

Class-wide NUMT identification

To investigate the evolutionary consequences of NUMT integration in a comparative context, we identified mitochondrial insertions into the nuclear genomes of more than one thousand mammalian and avian species. Specifically, we analysed 680 mammalian genomes and 458 avian genomes for which both a genomic and mitogenomic assembly was available from public databases. We used a previously published bioinformatic pipeline that uses the host's mitochondrial genome DNA as a BLAST query to identify NUMT insertions in the nuclear genome (Vendrami et al. 2022). Altogether, we identify 85,099 NUMTs in 435 mammalian species (66.62%) and 28,947 NUMTS from 458 avian species (100%). This observation is in line with previous findings on 70 vertebrate species that NUMT integration is a frequent and ongoing process across mammals and birds (Hazkani-Covo 2022). We observed fewer NUMTs in birds relative to mammals (average ≈ 125 versus ≈ 63 NUMTs per species, respectively), and also find that for mammals the number of NUMTs is positively correlated with genome size, which is not the case for birds ($P < 2.2 \times 10^{-16}$ and $P = 0.79$, for mammals and birds, respectively, phylogenetic generalized least-squares regression).

Phylogenetic patterns of NUMTs across mammals and birds

As we observe contrasting pattern in terms of NUMT numbers between the two taxonomic groups, we next considered the NUMT occurrences within a phylogenetic context. For this we used established species phylogenies for mammals and birds (Kumar et al. (2017), Timetree.org) and reconstructed the number of NUMTs as a phylogenetic trait using a Baysian approach with BayesTraits V4.0.1 (Pagel and Meade 2022). We observe several clear phylogenetic pattern in the data. First, the number of NUMTs varies substantially across clades (Figure 2). For example, among mammals, most marsupials show a complete lack or a very low number of NUMTs with the exception of the Tasmanian devil (*Sarcophilus harrisii*) (Hazkani-Covo 2022). In contrast, a few clades, such as old world monkeys and Cetaceans carry comparatively large numbers of NUMTs. For avian species, three clades in Passeriformes show an conspicuous excess of NUMTs. The first of these clades includes warblers, cowbirds and blackbirds. The second clade includes sparrows (e.g. *Zonotrichia albicollis*), junco, seedeaters, Tanager and finches (e.g. *Geospiza fortis*) and the third clade includes cardinals (*Cardinalis cardinalis*; Sin, Lu, and Edwards (2020)), buntings, grosbeak, longspur and grackle. Second, the number of NUMTs varies substantially between closely related species. For example for birds, *Calidris pygmaea*, a highly endangered species (Chowdhury, Foysal, and Green 2022), shows

a massive number of NUMT integrations suggesting a highly degenerated genome. Mammalian species with exceptionally high numbers of NUMTs include the Tasmanian devil (*Sarcophilus harrisii*, 1,546 NUMTs), gray short-tailed opossum (*Monodelphis domestspizella passerinaica*, 811 NUMTs) and yellow-footed antechinus (*Antechinus flavipes*, 792 NUMTs) in mammals. In birds, beside the already mentioned spoon-billed sandpiper (*Calidris pygmaea*, 13,170 NUMTs) are also an island finch (*Nesospiza acunhae*, 403 NUMTs) and the northern cardinal (*Cardinalis cardinalis*, 373 NUMTs) species with a substantial NUMT enrichment in their genomes.

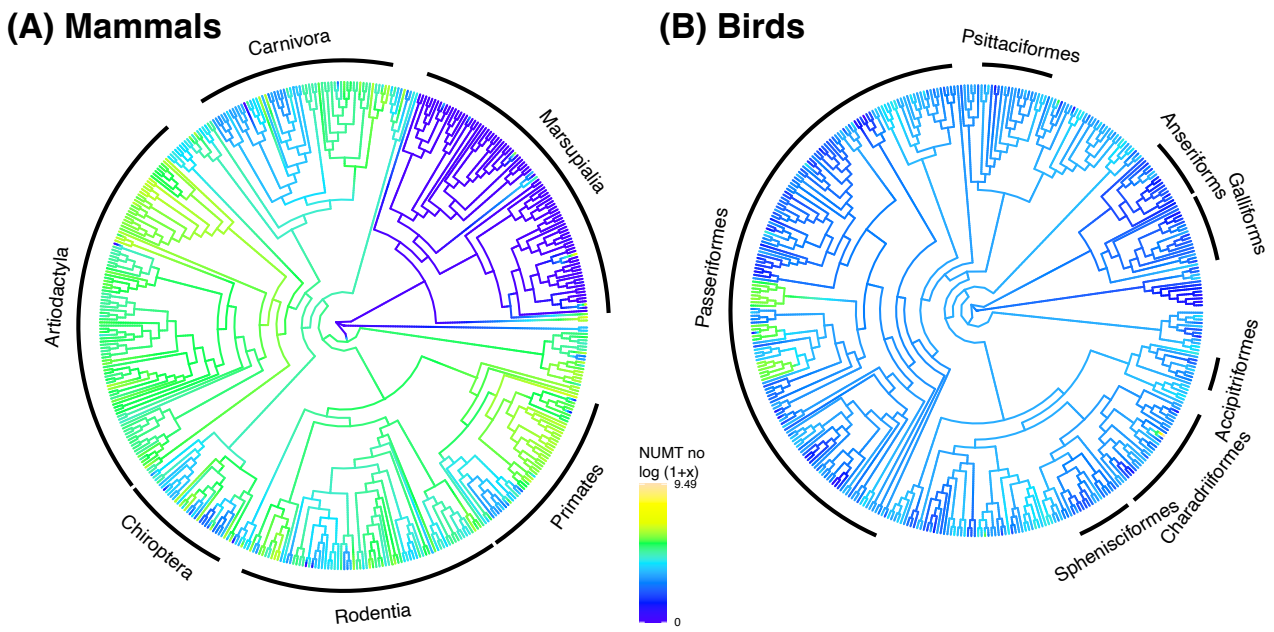


Figure 2: Phylogenetic distribution of NUMT occurences in mammals and birds. Shown are the species cladograms from timetree and colours represent reconstructed number of NUMTs for **(A)** mammals and **(B)** birds, respectively. Note that the colorised representation refers to the log-transformed (i.e.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

Distinct distributions of coding NUMTs, a subclass of NUMTs

While our data show that mitochondrial genome integration into the nuclear genome is a frequent and ongoing process in many species, it remains unclear what the origins and evolutionary trajectories of such NUMTs are, in particular those that are retained over long evolutionary time harbouring genes with intact mitocoding reading frames. To address this, we focus on a subgroup of NUMTs retaining at least one coding gene with an intact reading frame (coding NUMTs, cNUMTs). Among the 114,046 NUMTs identified in this study, 6,429 NUMTs from 698 species (5,229 NUMTs for 379 mammals and 1,200 NUMTs for 319 for birds) possess at least one mitochondrial gene with an intact reading frame. In mammals the number of cNUMTs is correlated with the number of NUMTs that do not have at least one gene with intact reading frame, which is not the case for birds ($P < 2.2 \times 10^{-16}$ and $P = 0.9$ for mammals and birds, respectively, phylogenetic generalized least-squares regression). This suggests that cNUMTs may be considered as a by-product of NUMT integration in mammals,

but not in birds. Using pairwise coding divergence estimates (Wang et al. 2010), we obtain a proxy for the evolutionary age of the cNUMTs using the synonymous divergence between the cNUMT genes and the respective genes on the host mitochondrial DNA. The distribution of these dS values may uncover common integration times across species (Figure 3). We find that while the number of cNUMT becomes lower for increasing dS values in mammals, for birds there is a clear peak at $dS \approx 1$, which might reflect a historic episode of high transposable element activity.

Non-host origin of dcNUMTS reveal introgressed 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 to the respective gene in the host mitochondrial genome (dcNUMTs, deep 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 non-canonicial 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 NUMTs contain multiple divergent mitocoding genes.

We identify 67 nuclear mitocoding genes (1.62 %) in mammals and 132 nuclear mitocoding genes (14.39 %) in birds that we can assign to a non-host mitochondrial genome with high confidence (Table 1, Figure 4). Surprisingly we find that 77.01% and 41.88% 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 group but the exact origins remain obscure. This suggests that these integrations are either not very recent, or that the current databases do not fully reflect the true biological mitochondrial diversity of mammals and birds.

Table 1: Numbers of BLAST hits of dcNUMTs against available mitochondrial genomes for birds and mammals. *High confidence is defined as a BLAST hit with at least 95% identity to the query.For a hit to be assigned 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 (dcNUMT as query)	Mammals	Birds
host species mitochondrial genome	557	134
with high confidence*	**81**	**116**
non-host species mitochondrial genome	392	399
with high confidence*	**67**	**132**
no species in the database	3179	384

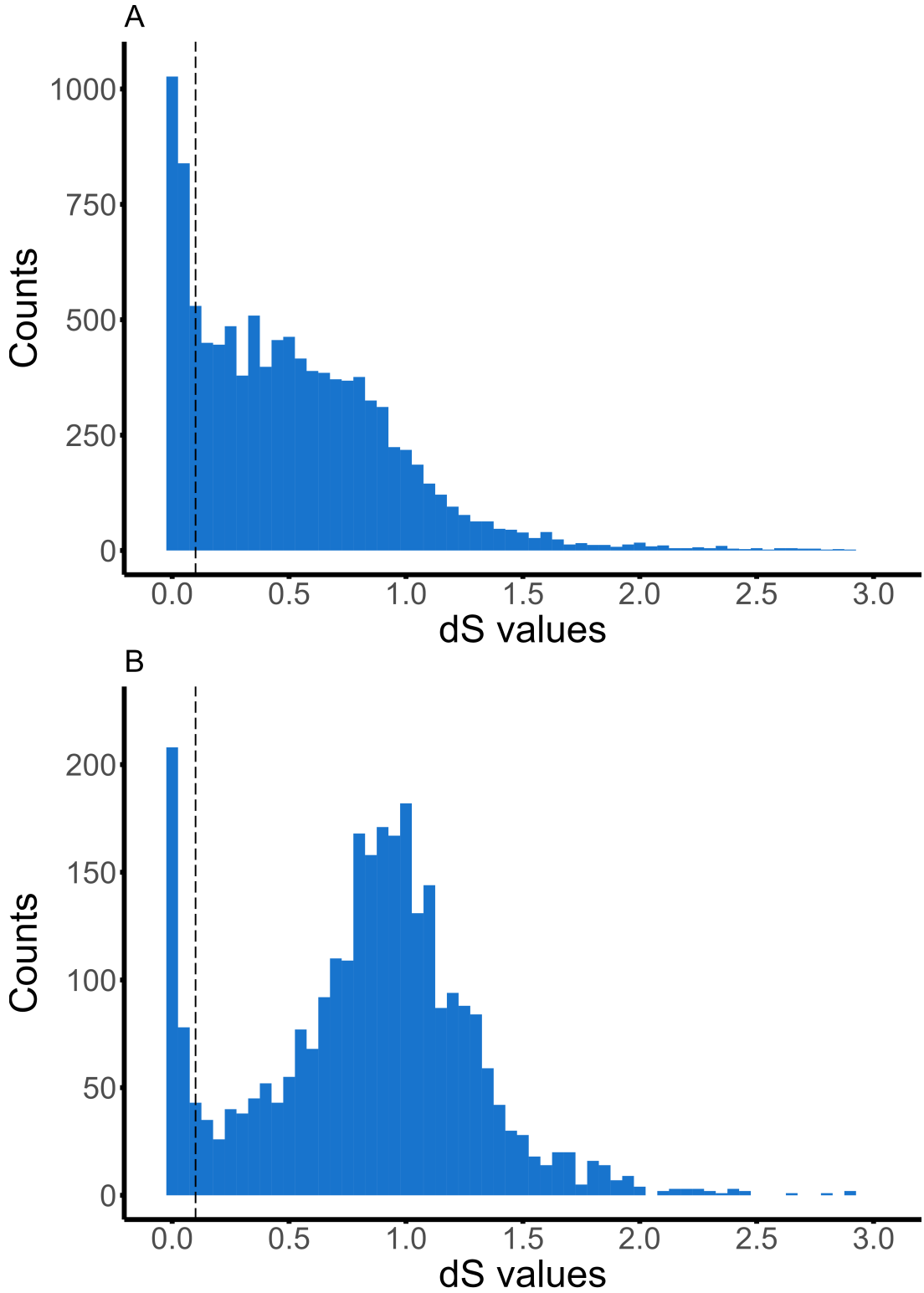


Figure 3: Distribution of dS of genes in cNUMTs. Shown are the distribution of dS between genes on the NUMTs and the host species mitochondrial DNA. NUMTs that possess at least one genes with $dS > 0.1$ are considered dcNUMTs, $dS = 0.1$ are indicated by a dashed line. Distributions are shown for **(A)**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

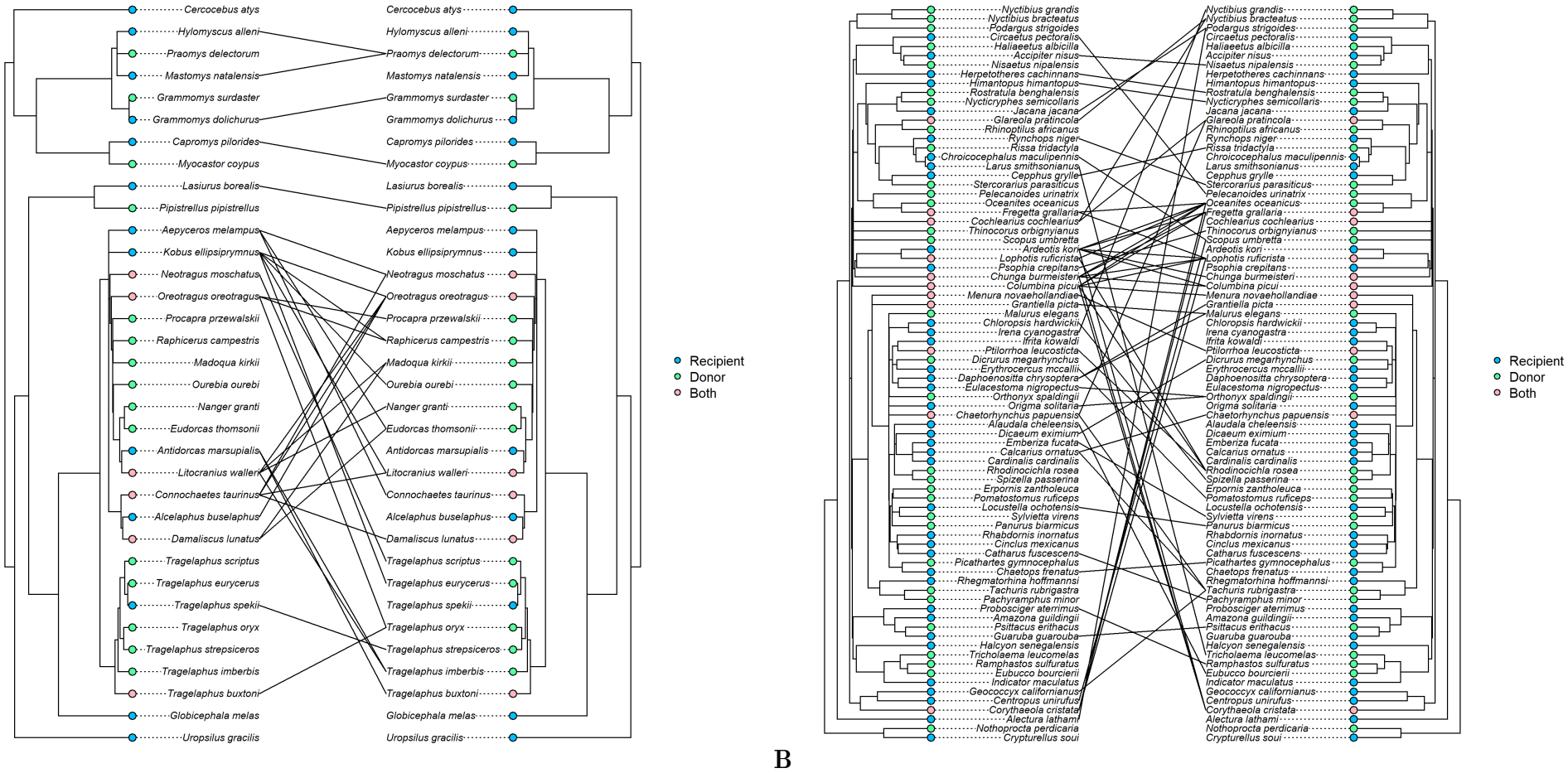
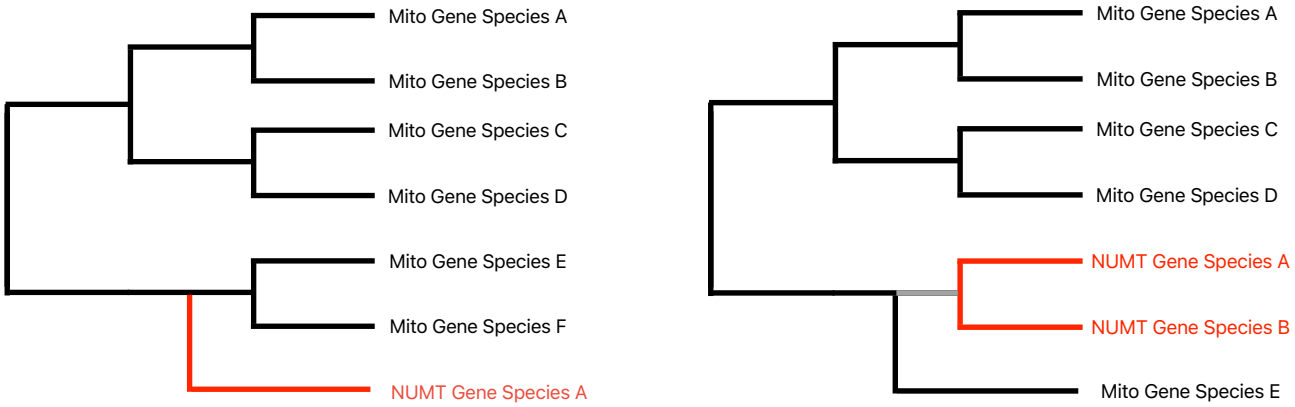


Figure 4: Phylogenetic pattern of NUMT integrations of non-host mitochondrial DNA. For **(A)** mammals and **(B)** birds. Shown are species that either show evidence of NUMTs of non-host origin or that appear be the donor species of the NUMT. Note that some species can be donor and recipient and that 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 S1 and S2.

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 (dN/dS) to be approximately one. To test this, we divided our dataset into: (A) Single NUMT genes (Figure 5A), which highly divergent to 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 cluster together with available mitochondrial genes of related species to calculate dcNUMT gene (Figure 5A) and gene cluster specific dN/dS ratios (Figure 5B).



A **B**

Figure 5: Illustration of foreground and background branches for codeml analysis. (A) Foreground branch in red for a single NUMT gene, background branches (black) are for the respective mitochondrial genes of related species **(B)** Foreground branch in red for NUMT gene clusters (here illustrated with cluster size two), background branches (black and grey) are for the respective mitochondrial genes of related species. Note that the grey branch for (B) is a background, not a foreground branch consitent with a node based last common ancestor clade assignment.

Selection analysis on isolated dcNUMT genes suggest non-canonical 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 the proportion 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.

However, it is possible that the last mitochondrial lineages from which these NUMT were derived are not included in our dataset, either because they became extinct or because they are lacking from our database. Therefore signatures of purifying selection for single NUMT genes could reflect the fact that our database lacks substantial parts of ancient or recent mitochondrial species diversity. However, our results are also consistent with the possibility of true purifying selection on the NUMTs.

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 (dN/dS<1)	1,337	377
Neutral evolution (dN/dS=1)	2,211	1,865
Positive selection (dN/dS>1)	1	4

Selection analysis on NUMT gene clusters suggest 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). In particular we tested the selection pressure 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). By that it is 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 would the estimated selective pressure analysis reflect selection on the NUMT, but not mitochondrial DNA.

We identified over ten times more of these NUMT gene clusters for mammalian species than for birds (1,169 versus 110). 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 to the host's mitogenome than avian dcNUMTs (Figure 3), they are as frequently shared between species. Based on the evolutionary rate analysis we identify almost 400 gene clusters that show evidence of purifying selection, most of them specific to mammals. Moreover, we identify an additional 25 gene clusters, 23 for mammals and 2 for birds, that show evolutionary signatures consistent with positive selection. This suggests that the identified NUMTs are potentially functionally conserved and therefore contribute towards adaptive variation.

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 (dN/dS<1)	394	33
Neutral evolution (dN/dS=1)	752	75
Positive selection (dN/dS>1)	23	2

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 S1) and find very little 3D structural variation between the human ND4L protein and the 3D protein prediction of the NUMT gene. Therefore, this hypothetical prediction does not suggest that the accumulated mutations in the NUMT strongly alters 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). Instead we find that the genomic region where the NUMT is located is known to have evolved under positive selection in the ancestral lineage of modern humans (Green et al. 2010; Prüfer et al. 2013; Racimo 2015).

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 dN/dS=ω value for the entire tree). Positive selection was inferred from a two ratio model and a fixed two-ratio model for which the dN/dS 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)

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.

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. We use a previously established pipeline that uses a species' mitochondrial genome as a query for a nuclear genomic BLAST identification of the same species (Vendrami et al. 2022). Altogether we 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 necessarily surprising 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, as they mainly located on small scaffolds that are often not much bigger than a mitochondrial genome (e.g. <20kb) itself (Figure S2). 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.

Some of the identified NUMTs are very distantly related to the host's mitochondrial DNA. Because 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 NUMTs per species are potentially underestimates. For example, we do not find any NUMTs that were 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. Hence our NUMT numbers are an overestimate of the number of NUMT integration events from non-nuclear 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 (Caswell et al. 2019; Sangster and Luksenburg 2021), mitochondrial DNA and the NUMT DNA could appear more divergent. To limit the impact of this to our dcNUMT analysis, we focused on coding NUMTs that have 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 high 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 genome in the database (Table S3), 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.

Table 5: Number of NUMTS, coding NUMTs (cNUMTs) and deep coding NUMTs (dcNUMTS) in birds and mammals.

NUMT type	No in mammals	No in birds
NUMTS	85,099	28,947
cNUMTS	5,229	1,200
dcNUMTS	4,182	917
dcNUMTS/non-dcNUMTS in groups	2,159/144	131/6
single dcNUMTS	1,512	760

dcNUMTs To identify non-canonical origins of NUMTs and potential signatures of selection, we focused on dcNUMTs, which still contain genes with an intact mitochondrial reading frame. Because the nuclear and mitochondrial genome use different genetic codes, it is unlikely that mitochondrial genes integrated into the nuclear genome can be translated into intact proteins. Moreover, after integration of the mitochondrial DNA into the nuclear genome, NUMTs are expected to be accumulate mutations over time that likely disrupt the mitochondrial reading frame. However, we find more than 5,000 NUMTs in mammals and birds (Table 5) that contain genes with an intact mitochondrial reading frame and that show relatively high divergence to the respective hosts mitochondrial encoding genes (dcNUMTs with genes with $dS > 0.1$, i.e. a synonymous per site divergence of more than 10%). For mammals we find that the number of dcNUMT becomes lower for increasing dS 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, 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 and may have a function on the RNA level, or that the DNA itself has some regulatory function as observed for mitochondrial DNA (Acker et al. 2023) or creates novel exons (Noutsos et al. 2007). For example, the transcription of non-coding genomic DNA is very common (Gao et al. 2020) and active retroviruses (Bolisetty et al. 2012) may could lead to expression of NUMTs as well. However, this remains speculative and would need to be supported by a signature signature consistent with functionality. To address this, we investigated the evolutionary signatures of mitocoding NUMT genes. Specifically, two separate analyses that may explain the existence of dcNUMTs were conducted. First, we focused on the hypothesis that the NUMTs result from introgression or hybridisation events. Second, we analysed the selection pressures on coding DNA to identify traces of directional selection.

Introgression/Hybridisation To identify potentially introgressed/hybridised NUMTs we BLASTed the 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).

We identify 67 dcNUMTs in mammals and 132 in birds that can be assigned to a different other species than the host's species mitogenome with high confidence (Tables 1, S1 and S2 as well as Figures 4A and 4B). 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 non-canonical NUMT origins. In mammals, almost all of the introgressed NUMTs occur within the ungulates (Figure 4A and Table S1) - a taxonomic group in which hybridization plays an extensive role (Iacolina et al. 2018).

In birds, we find introgression events between distantly related species (Figure 4B, Table S2). For example, we identified a NUMT that shows a signature of introgression from the tawny frogmouth (*Podargus strigoides*), a species of frogmouth native to the Australian mainland and Tasmania, to the gweela (*Alectura lathami*), a mound-building bird from the family Megapodiidae found in eastern Australia. A possible explanation of cross-species NUMT transfer 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 cross species transmission of retroviruses 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. Because our dataset heavily relies on genome assemblies retrieved from short-read sequencing, we cannot pinpoint the exact genomic locations of many of the identified NUMTs, because they tend to be located on very short scaffold (Figure S2) - and this appears to particularly true for avian dcNUMTs. Unable to locate the precise genomic locations, however, hampers the possibility to pinpoint integration events that happened as a consequence of 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 could help to further unravel 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. Surprisingly, the number is almost ten-fold higher in mammals compared to birds. A deeper mammalian sampling or the retrieval of ancient mammalian mitogenomes might be enable us to reveal the evolutionary origins and fates of further mammalian NUMTs in the future.

Tests of selection To test for potential signatures of selection specific to NUMTs we conducted substitution rate analysis of genes located in NUMTs. We restricted our analysis to those genes with intact reading frames showing substantial synonymous divergence to the host's mitochondrial genome. Specifically, we conducted our analysis using the mitochondrial genetic code as there is no known

mechanism that would allow nuclear genes to be translated into proteins using the mitochondrial code. Instead, it is likely that the reading frames of NUMT genes should deteriorate over evolutionary time due to the acquisition of new mutations. The long-term maintenance of NUMT genes might therefore be a signal of selection. For this, we classified the NUMT genes into two groups, single NUMT genes (i.e. isolated NUMT genes with no other NUMT genes in close proximity) and grouped NUMT genes (i.e. NUMT genes with high similarity to a NUMT gene of another species or another NUMT gene of the same species). The latter group would allow us to evaluate the evolutionary pressures that have acted on the NUMT gene when it integrated into the nuclear genome (Figure 5).

For the majority of single NUMT genes, we observe a signature of selection consistent with drift (e.g. dN/dS=1). However, we also identify more than 2,000 genes showing signatures of purifying selection (Table 2), comprising 54% and 83% of the tested NUMT genes for mammals and birds, respectively. Surprisingly, we also identify five NUMT genes showing signatures of positive selection. While it is intriguing that many 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 divergence before the NUMT integration took place. To address this limitation, we conducted clade specific tests of selection on groups of NUMT genes. By that it is likely that the considered branches in the clade with the NUMTs reflect an evolutionary time when the NUMT was truly integrated into the nuclear genome. However, it is also possible that by that approach we group sequences that have evolved the same sequence independently (Hazkani-Covo, Zeller, and Martin 2010; Song et al. 2013). Again, for the majority of grouped NUMT genes we find an evolutionary signature consistent with genetic drift (Table 3), with 64% and 68% of the groups showing selection pressures consistent with drift for mammals and birds, respectively.

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 the 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). Hence, 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.

Conclusion

Here we have shown that NUMTs may originate from non-canonical introgression 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 based 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. 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 introgression maybe understood better when placing NUMTs into phylogenetic context. Even if the NUMTs themselves are not functional, they provide excellent markers of (positive) selection and introgression/hybridisation.

Materials and methods

Data collection and selection

The genome assemblies of all mammalian species (available before November, 11, 2021) species and all avian species (available before November, 02, 2021) species were obtained from GenBank (Sayers et al. 2021) using ncbi-genome-download v0.3.1 (https://github.com/kblin/ncbi-genome-download). If one species had multiple assemblies, the representative genome or the genome with higher coverage (if no assemblies were stated as representative) was included in our database.

The mitogenomes with at least one partial coding gene of all mammalian species (available before November, 22, 2021) or all avian species (available before December, 16, 2021) were obtained from GenBank for building BLAST databases. The mitogenomes from the same species and subspecies (or closely related subspecies if the same subspecies were not found) with genome assemblies were selected for NUMT identification. If one species had multiple mitogenomes, the reference assembly was included. Unverified mitochondrial genomes were ignored. The protein-coding gene annotations were also obtained from GenBank. The mitogenomes without protein-coding gene annotations were annotated using the MITOS2 webserver (Donath et al. 2019). See Table S4 for the complete list of genomes and mitochondrial genomes and their accession numbers.

NUMT identification

For NUMT identification, we performed local BLASTN v2.6.0+ (Camacho et al. 2009) with default settings and used mitochondrial genomes as queries to search the corresponding genome assemblies of the same species. The hits with subject lengths shorter than 200 bp were discarded. The corresponding NUMT sequences were extracted with BEDTools v2.26.0 (Quinlan and Hall 2010). NUMTs and their reverse sequences (generated by EMBOSS v6.6.0.0 revseq algorithm (Rice, Longden, and Bleasby 2000) were pairwise aligned with the protein coding genes of the corresponding mitochondrial genome using MAFFT v7.310 (Katoh and Standley 2013).

The NUMTs numbers were mapped to the species trees from Timetree (Kumar et al. 2017) to examined their evolutionary trends across the phylogenies. The rate of evolution were estimated with independent contrast model using reversible jump Markov chain Monte Carlo (RJMCMC) in BayesTraits v.4.0.1 (Pagel and Meade 2022). For each dataset we ran 12,000,000 iterations, sampling once every 2,000 iterations after a 2,000,000 burn-in iterations. The branch length transformation based on the rate change were applied to ancestral reconstructions of NUMTs numbers with ace function in the phytools packages (Revell 2012) in R (v4.2.3).

Substitution rate analyses of NUMTs with coding genes

We first checked whether the aligned sequences were coding in frame (e.g. lacked stop codons). If this was the case, we calculated the synonymous (dS) substitution rates of the NUMTs with KaKs calculator (Wang et al. 2010) under the NG substitution model (Nei and Gojobori 1986). The NUMTs with at least one coding gene with dS $\geq$ 0.1 were considered as potentially coding

divergent NUMTs. These potentially coding divergent NUMTs were used as queries to BLAST against (Camacho et al. 2009) the mitogenomes of all mammalian species and all avian species, respectively (Table S4). The NUMTs with best hits to mitogenomes from another species with percentages of identical matches $\geq$ 98% but less than 95% to its own miochondrial genome were examined on a case-by-case basis to test for potential lateral transfer between species pairs.

Clustering NUMTs and selection analysis

To investigate the functional constraints of coding sequences on cNUMTs, we first identified NUMTs derived from the same source and then analyzed the ratio of non-synonymous to synonymous (dN/dS) of each NUMT clusters. We used mafft v7.310 (Katoh and Standley 2013) to align protein coding sequences identified in cNUMTs and reconstructed the coding sequence trees with iqtree v2.0.3 (Minh et al. 2020). Then we used CODEML implemented in PAML (Yang 2007) to calculate pairwise dN/dS comparisons and substitutions per codon. The coding sequence pairs with substitutions per codon less than 0.1 were examined for their best hits against the mitogenomes. The pairs with both coding sequences' 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 in pairs were grouped into one cluster.

After clustering, we first translated the coding sequences into protein sequences, and then aligned each cluster to the respective mitochondrial genes from the species used for NUMT identification 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 3 branch models: a one-ratio model, a two-ratio model and a two-ratio model with the omega of NUMT clades fixed to 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.

Data deposition

All scripts are available at github https://github.com/chnyuch/numt_vertebrate

Acknowledgment

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).

Author contribution

TIG designed the study with input from JIH and DV. YC conducted the majority of the analyses. MH and LH conducted analyses on species subsets. TIG and YC drafted the manuscript with input from all of the authors.

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.

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.

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/10.1038/s41598-017-16750-2.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (421): 1–9. https://doi.org/10.1186/1471-2105-10-421.

Caswell, Jacob, Jason D. Gans, Nicholas Generous, Corey M. Hudson, Eric Merkley, Curtis Johnson, Christopher Oehmen, et al. 2019. "Defending Our Public Biological Databases as a Global

Critical Infrastructure." *Frontiers in Bioengineering and Biotechnology* 7 (April). https://doi.org/10.3389/fbioe.2019.00058.

Chowdhury, Sayam U., Mohammod Foysal, and Rhys E. Green. 2022. "Accelerating Decline of an Important Wintering Population of the Critically Endangered Spoon-billed Sandpiper *Calidris Pygmaea* at Sonadia Island, Bangladesh." *Journal of Ornithology* 163 (4): 891–901. https://doi.org/10.1007/s10336-022-01995-0.

Clark, Alice, Gizem Koc, Ying Eyre-Walker, and Adam Eyre-Walker. 2023. "What Determines Levels of Mitochondrial Genetic Diversity in Birds?" Edited by Aida Andres. *Genome Biology and Evolution* 15 (5). https://doi.org/10.1093/gbe/evad064.

Danchin, Etienne G. J. 2016. "Lateral Gene Transfer in Eukaryotes: Tip of the Iceberg or of the Ice Cube?" *BMC Biology* 14 (1). https://doi.org/10.1186/s12915-016-0330-x.

Dayama, Gargi, Sarah B. Emery, Jeffrey M. Kidd, and Ryan E. Mills. 2014. "The Genomic Landscape of Polymorphic Human Nuclear Mitochondrial Insertions." *Nucleic Acids Research* 42 (20): 12640–49. https://doi.org/10.1093/nar/gku1038.

Diehl, William E, Nirali Patel, Kate Halm, and Welkin E Johnson. 2016. "Tracking Interspecies Transmission and Long-Term Evolution of an Ancient Retrovirus Using the Genomes of Modern Mammals." *eLife* 5 (March). https://doi.org/10.7554/elife.12704.

Donath, Alexander, Frank Jühling, Marwa Al-Arab, Stephan H. Bernhart, Franziska Reinhardt, Peter F. Stadler, Martin Middendorf, and Matthias Bernt. 2019. "Improved Annotation of Protein-Coding Genes Boundaries in Metazoan Mitochondrial Genomes." *Nucleic Acids Research* 47 (20): 10543–52. https://doi.org/10.1093/nar/gkz833.

Dunning, Luke T., Jill K. Olofsson, Christian Parisod, Rimjhim Roy Choudhury, Jose J. Moreno-Villena, Yang Yang, Jacqueline Dionora, et al. 2019. "Lateral Transfers of Large DNA Fragments Spread Functional Genes Among Grasses." *Proceedings of the National Academy of Sciences* 116 (10): 4416–25. https://doi.org/10.1073/pnas.1810031116.

Flavell, Andrew J. 1999. "Long Terminal Repeat Retrotransposons Jump Between Species." *Proceedings of the National Academy of Sciences* 96 (22): 12211–12. https://doi.org/10.1073/pnas.96.22.12211.

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., 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.

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.

Jin, Jian-Jun, Wen-Bin Yu, Jun-Bo Yang, Yu Song, Claude W. dePamphilis, Ting-Shuang Yi, and De-Zhu Li. 2020. "GetOrganelle: A Fast and Versatile Toolkit for Accurate de Novo Assembly of Organelle Genomes." *Genome Biology* 21 (1). https://doi.org/10.1186/s13059-020-02154-5.

Junier, Thomas, and Evgeny M. Zdobnov. 2010. "The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell." *Bioinformatics* 26 (13): 1669–70. https://doi.org/10.1093/bioinformatics/btq243.

Kapusta, Aurélie, Alexander Suh, and Cédric Feschotte. 2017. "Dynamics of Genome Size Evolution in Birds and Mammals." *Proceedings of the National Academy of Sciences* 114 (8). https://doi.org/10.1073/pnas.1616702114.

Katoh, Kazutaka, and Daron M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30 (4): 772–80. https://doi.org/10.1093/molbev/mst010.

Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges. 2017. "TimeTree: a resource for timelines, timetrees, and divergence times." *Molecular Biology and Evolution* 34 (7): 1812–19. https://doi.org/10.1093/molbev/msx116.

Lee, Michael S. Y. 1998. "Ancestors and Taxonomy." *Trends in Ecology & Evolution* 13 (1): 26. https://doi.org/10.1016/s0169-5347(97)01272-x.

Liang, Bin, Ning Wang, Nan Li, Rebecca T. Kimball, and Edward L. Braun. 2018. "Comparative Genomics Reveals a Burst of Homoplasy-Free Numt Insertions." Edited by Anne Yoder. *Molecular Biology and Evolution* 35 (8): 2060–64. https://doi.org/10.1093/molbev/msy112.

Lopez, Jose, Melanie Culver, J. Claiborne Stephens, Warren E. Johnson, and Warren Stephen J. O'brien. 1997. "Rates of Nuclear and Cytoplasmic Mitochondrial DNA Sequence Divergence in Mammals." *Molecular Biology and Evolution* 14 (3): 277–86. https://doi.org/10.1093/oxfordjournals.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.

Pozzi, Andrea, and Damian K. Dowling. 2019. "The Genomic Origins of Small Mitochondrial RNAs: Are They Transcribed by the Mitochondrial DNA or by Mitochondrial Pseudogenes Within the Nucleus (NUMTs)?" Edited by Daniel Sloan. *Genome Biology and Evolution* 11 (7): 1883–96.

https://doi.org/10.1093/gbe/evz132.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments." *PloS One* 5 (3): e9490. https://doi.org/10.1371/journal.pone.0009490.

Prüfer, Kay, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, et al. 2013. "The Complete Genome Sequence of a Neanderthal from the Altai Mountains." *Nature* 505 (7481): 43–49. https://doi.org/10.1038/nature12886.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Racimo, Fernando. 2015. "Testing for Ancient Selection Using Cross-Population Allele Frequency Differentiation." *Genetics* 202 (2): 733–50. https://doi.org/10.1534/genetics.115.178095.

Revell, Liam J. 2012. "Phytools: An R Package for Phylogenetic Comparative Biology (and Other Things)." *Methods in Ecology and Evolution* 3 (2): 217–23. https://doi.org/https://doi.org/10.1111/j.2041-210X.2011.00169.x.

Rhoads, Anthony, and Kin Fai Au. 2015. "PacBio Sequencing and Its Applications." *Genomics, Proteomics & Bioinformatics* 13 (5): 278–89. https://doi.org/10.1016/j.gpb.2015.08.002.

Ricchetti, Miria, Cécile Fairhead, and Bernard Dujon. 1999. "Mitochondrial DNA Repairs Doublestrand Breaks in Yeast Chromosomes." *Nature* 402: 96–100. https://doi.org/10.1038/47076.

Ricchetti, Miria, Fredj Tekaia, and Bernard Dujon. 2004. "Continued Colonization of the Human Genome by Mitochondrial DNA." Edited by Jonathan A. Eisen. *PLoS Biology* 2 (9): e273. https://doi.org/10.1371/journal.pbio.0020273.

Rice, Peter, Ian Longden, and Alan Bleasby. 2000. "EMBOSS: The European Molecular Biology Open Software Suite." *Trends in Genetics* 16 (6): 276–77. https://doi.org/10.1016/s0168-9525(00)02024-2.

Richly, Erik, and Dario Leister. 2022. "NUMTs in Sequenced Eukaryotic Genomes." *Molecular Biology and Evolution* 21 (6): 1081–84. https://doi.org/10.1093/molbev/msh110.

Sangster, George, and Jolanda A Luksenburg. 2021. "Sharp Increase of Problematic Mitogenomes of Birds: Causes, Consequences, and Remedies." Edited by Liliana Milani. *Genome Biology and Evolution* 13 (9). https://doi.org/10.1093/gbe/evab210.

Sayers, Eric W., Mark Cavanaugh, Karen Clark, Kim D. Pruitt, Conrad L. Schoch, Stephen T. Sherry, and Ilene Karsch-Mizrachi. 2021. "GenBank." *Nucleic Acids Research* 49 (D1): D92–96. 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.

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.

Appendix

Supplementary Figures

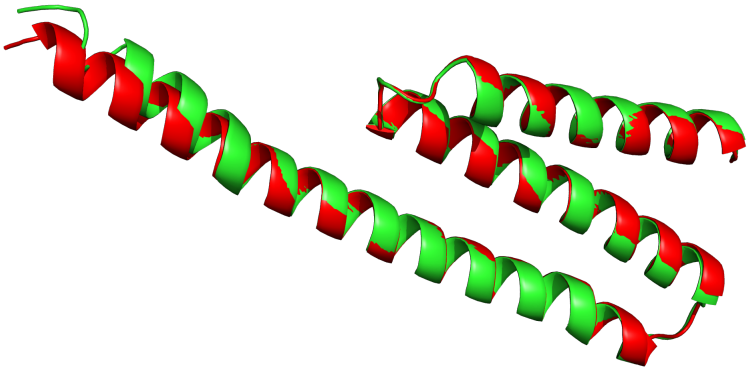


Figure S1: Alphafold 3D prediction of a human NUMT gene along with the corresponding human mitochondrial gene. The NUMT gene is part of a dcNUMT cluster that is shared among seven ape species (Figure 6). The NUMT gene with an intact reading frame is ND4L. Shown are the 3D structures of human ND4L as well as a structure predcition conducted with alphafold based on the hypothetical protein seqeunce assuming mitocondrial code

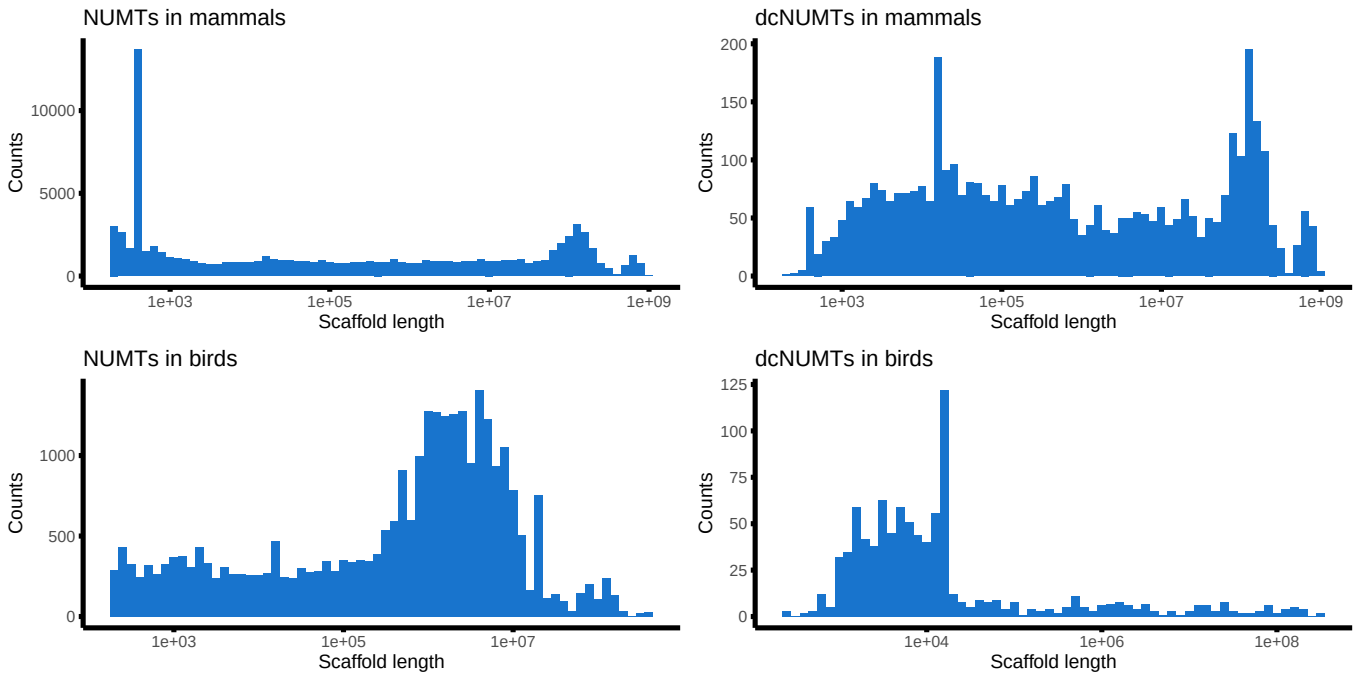


Figure S2: Distribution of genomic scaffold length of scaffolds that contain NUMTs shown for mammalian and avian species. Many identified NUMTs are on short genomic fragments that are not assigned to chromosomes.

Supplementary Tables

Table S1: All mammalian host - donor pairs of NUMTs that likely originated from a different species. For full table see Supplementary_tables.xlsx

Recipient species	NUMT ID	Donor scaffold ID	Donor species
Aepyceros melampus	SJXQ01006944.1:63574-66381	NC_020699.1	Connochaetes taurinus
Aepyceros melampus	SJXQ01006944.1:78349-91308	NC_039669.1	Eudorcas thomsonii
Aepyceros melampus	SJXQ01012544.1:217760-229605	NC_020728.1	Neotragus moschatus
Alcelaphus buselaphus	SJXN01419338.1:1-924	NC_020731.1	Oreotragus oreotragus
Alcelaphus buselaphus	SJXN01502202.1:1-1258	NC_020731.1	Oreotragus oreotragus
Alcelaphus buselaphus	SJXN01519847.1:1-1458	NC_020731.1	Oreotragus oreotragus

Table S2: All avian host - donor pairs of NUMTs that likely originated from a different species. For full table see Supplementary_tables.xlsx

Recipient species	NUMT ID	Donor scaffold ID	Donor species
Accipiter gentilis	BJER01014652.1:2587-13711	NC_040858.1	Haliaeetus albicilla
Accipiter nisus	BJBX01032538.1:1946-9915	NC_007598.1	Nisaetus nipalensis
Alaudala cheleensis	VWYE01023216.1:285-2530	MN356176.1	Eubucco bourcierii
Alaudala cheleensis	VWYE01025130.1:1181-2744	MN356176.1	Eubucco bourcierii
Alectura lathami	VXAV01019081.1:1-243	MN356205.1	Podargus strigoides
Amazona guildingii	VXAR01004813.1:3-2347	MN356178.1	Ciccaba nigrolineata

Table S3: Comparison of the genetic and mitogenetic diversity of 22 avian species for which mitogenomic diversity was retrieved form raw sequencing reads For full table see Supplementary_tables.xlsx

Species	Genome wide heterozygosity	Pairwise mitochondrial difference	Assembly ID
Ardeotis kori	0.00203	0.003	NC_52775.1
Calcarius ornatus	0.01058	0.000	NC_53083.1
Centropus bengalensis	0.00150	0.037	NC_50049.1
Ceuthmochares aereus	0.00357	0.001	NC_52776.1
Chaetorhynchus papuensis	0.00547	0.000	NC_51007.1
Chunga burmeisteri	0.00447	0.000	MN356168.1

Table S4: All identifiers of the genomes and mitogenomes and their respective species used in this study For full table see Supplementary_tables.xlsx

Genome	Assembly_accession	Genome	Accession_number	Length	cds_number	Download_link
Mammal	NA	NA	NA	NA	NA	NA
Acinonyx jubatus	GCA_003709585	Acinonyx jubatus	NC_005212.1	17047	13	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/003/709/585/GCA_003709585.1_Aci_jub_2
Acomys cahirinus	GCA_004027535	Acomys cahirinus	JN571155.1	16217	13	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/004/027/535/GCA_004027535.1_AcoCah_v1_BIUU
Acrobates pygmaeus	GCA_904813335	Acrobates pygmaeus	KJ868094.1	15322	11	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/904/813/335/GCA_904813335.1_ERS5079794_MarsupialExonCaptKit
Addax nasomaculatus	GCA_019593525	Addax nasomaculatus	NC_020674.1	16751	13	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/019/593/525/GCA_019593525.1_ASM1959352v1
Aepyceros melampus	GCA_006408695	Aepyceros melampus	NC_020675.1	16422	13	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/006/408/695/GCA_006408695.1_IMP