Multiple origins of a frameshift insertion in a mitochondrial gene in birds and turtles

Sergio Andreu-Sánchez [1,2], Wanjun Chen[3], Josefin Stiller [1] and Guojie Zhang [1,3,4,5,*]

[1]Villum Centre for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark; [2]Present Address: University of Groningen, University Medical Center Groningen, Department of Pediatrics, 9700 RB Groningen, Netherlands; [3]China National Genebank, BGI-Shenzhen, Beishan Industrial Zone, 518083 Shenzhen, China; [4]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, 650223 Kunming, China and [5]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, 32 Jiaochang Donglu, 650223 Kunming, China

*__Correspondence address.__ Guojie Zhang, E-mail: guojie.zhang@bio.ku.dk ⃝ http://orcid.org/0000-0001-6860-1521Villum Centre for Biodiversity Genomics, Section for Ecologyy and Evolution, Department of Biology, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark

Abstract

__Background:__ During evolutionary history, molecular mechanisms have emerged to cope with deleterious mutations. Frameshift insertions in protein-coding sequences are extremely rare because they disrupt the reading frame. There are a few known examples of their correction through translational frameshifting, a process that enables ribosomes to skip nucleotides during translation to regain proper reading frame. Corrective frameshifting has been proposed to act on the single base pair insertion at position 174 of the mitochondrial NADH dehydrogenase subunit 3 gene (ND3) that has been observed in several turtles and birds. However, the relatively sparse taxonomic representation has hampered our understanding of the evolution of this insertion in vertebrates. __Results:__ Here, we analyzed 87,707 ND3 sequences from 10,309 vertebrate taxa to reveal the evolutionary history of this insertion and its common genomic characteristics. We confirmed that the insertion only appears in turtles and birds and reconstructed that it evolved independently in both groups with complex patterns of gains and losses. The insertion was observed in almost all bird orders but was absent in all members of the diverse Passeriformes. We found strong conservation in the nucleotides surrounding the insertion in both turtles and birds, which implies that the insertion enforces structural constraints that could be involved in its correction. __Conclusions:__ Our study demonstrates that frameshifts can be widespread and can be retained for millions of years if they are embedded in a conserved sequence theme.

Keywords: programmed frameshift; sequence analysis; ancestral state reconstruction; mitochondrion; ND3

Background

Comparative analysis of molecular sequences across the diversity of life lets us discover which molecular mechanisms have been conserved and which have been modified throughout evolution. Insertions or deletions in protein-coding genes are usually selected against because they result in frameshifts that disrupt the amino acid coding frame and result in dysfunctional proteins [1]. Albeit rare, examples of corrective frameshifting exist, in which ribosomes regain the proper reading frame [2]. Programmed translational frameshifts have been character-

ized in viruses, retrotransposons, bacteria, yeast, and in mammalian antizymes (reviewed in [3, 4]). Additionally, corrective frameshifting was proposed to be acting in the mitochondria of a range of animals (ants [5], glass sponges [6], oysters [7], birds [8, 9], and turtles [9, 10]), where +1 frameshift insertions have been reported in up to 6 different mitochondrially encoded genes [11] with no clear functional consequences. Alternatively, other mechanisms than corrective frameshifting could explain the absence of functional consequences of the frameshift insertion, such as non-canonical translation of tetra or penta codons, which might be an ancient translation mechanism [12], or RNA editing [13].

Three common sequence features that surround frameshift insertions were identified in the better-known examples of translational frameshifts (*Ty1* and *Ty3* genes in yeast, antizyme gene in mammals, *prfB* in *Escherichia coli*): a transfer RNA (tRNA) that enables the ribosome to "slip" on the ribosome P-site, a rarely used codon in the A-site promoting the stall, and a commonly used codon in the +1 frame [14]. In mitochondrial genes with frameshifts, the mechanism leading to a programmed corrective frameshift is not as well characterized but the DNA sequence surrounding the mitochondrial frameshift insertions has the same features, indicating that the mechanism that facilitates the correction could be the same [11]. Three models have been proposed that may enable translational frameshifting [13]. The "pause-and-slip" model proposes that a pause is induced at the A-site of the ribosome and that the P-site tRNA can pair with the +1 codon, allowing it to slip out of frame [15]. A second model proposes that abnormal tRNA structures enable the frameshift [16]. The "out-of-frame" model proposes that the recruited tRNA skips the additional nucleotide in the A-site [11]. This out-of-frame correction of the frameshift is thought to act in glass sponges because of the conserved genomic features in several mitochondrial genes containing frameshifts [13].

The out-of-frame model may also apply to the frameshift insertion found in the mitochondrial genomes of certain turtles and birds, where a +1 frameshift insertion occurs at position 174 of the NADH dehydrogenase 3 gene (hereafter ND3–174+1) [9]. The nucleotide sequence around the ND3–174+1 insertion is conserved in a way that would facilitate out-of-frame pairing [13]. The insertion is usually found in a codon CUN, with N being the insertion, which produces a wobble pairing with tRNA-leucine when entering the P-site of the ribosome [11]. The codon downstream of the insertion is usually AGU, a rarely used codon for serine, which enters the A-site [11]. The CUN in the P-site and the AGU in the A-site are thought to initiate the frameshift correction by causing a stall in decoding [11]. The AGU codon further forms the beginning of a 15-bp long stem-loop RNA secondary structure [9], which may enhance the stall [11]. After the stall has been initiated, there are 2 possibilities. One is to maintain the shifted reading frame, which would lead to an early termination of the protein (in birds ending after 207 bp instead of the usual 354 bp [9]), or to produce a +1 frameshift, leaving out A-175 and thus recovering the regular reading frame.

ND3–174+1 was the first mitochondrial +1 frameshift described in vertebrates, initially discovered in ostrich (*Struthio camelus*) [8]. An extended investigation found that the ND3–174+1 insertion was also present in a species of turtle and in many other bird species (46 of 61 examined bird species) [9]. Within turtle mitogenomes, the insertion was widespread (27 of 31 examined turtle species), likely present in the ancestor and lost 2–3 times within turtles, while being absent in crocodiles (2 species examined) and snakes (1 species examined) [11]. From this pattern of absence and presence, it was suggested that the insertion could have been present in a common ancestor of turtles and birds (i.e., Archelosauria: birds+crocodiles, and turtles [17]) and subsequently lost in crocodiles, and lost again in specific lineages of turtles and birds [11]. The insertion was not observed in any other vertebrate lineage [9]. The relatively small number of species investigated in each vertebrate group limited the resolution of the evolutionary history of the insertion. Denser sampling promises to provide additional insight into the distribution of the frameshift insertion in different lineages and to determine not only the common features of the sequence theme that allow the insertion to remain in the genome, but also to find potential deviations from a conserved sequence theme.

Here, we build on the abundance of mitochondrial sequence data publicly available for vertebrates to study the evolution of the frameshift insertion in the mitochondrial protein-coding gene ND3. We compiled ND3 sequences representing 10,309 vertebrate taxa and reconstructed ancestral states and transformations of ND3–174+1 in different groups of Diapsida. We further investigated sequence conservation and codon usage patterns around the insertion site to identify potential common sequence motifs associated with the absence or presence of the frameshift insertion. This large dataset provides improved resolution to understand the evolution of this frameshift insertion and highlights that common sequence patterns appear to be required for maintaining the programmed translational frameshift.

Data Description

In this work, we have compiled a total of 87,708 ND3 sequences from 10,397 unique vertebrate taxa (data available in GigaDB [18]). Even though our focus was to represent ND3 sequences for vertebrate species, the taxonomy used on NCBI is not always representing formally named species. The "Organism" field in NCBI mostly contains traditional Linnaean names for species or subspecies, but in some cases the field contains unnamed taxa such as undescribed species, cryptic species, or uncertain species identification (designated with sp., cf., or a placeholder name). Although some of the unique sequence identifiers may not present distinct species, we include these taxa here in addition to the formally named species to represent a diversity of patterns in ND3. This decision also allowed us to match each unique sequence identifier to a terminal in the Open Tree of Life phylogenetic tree (see Methods), which also integrates NCBI's taxonomy.

Sequences were extracted from fragments of ND3 sequences and partial or full mitochondrial genomes downloaded from NCBI's GenBank and RefSeq databases. We reduced the dataset to 1 representative sequence for each unique entry in the "Organism" field for the following analyses. If multiple sequences for the same taxon existed, we confirmed that all sequences agreed in the presence or absence of the insertion at position 174. Among intraspecific records, we gave sequences from RefSeq preference over GenBank sequences. Among multiple GenBank records for the same taxon, a random sequence was chosen. This resulted in a dataset of 10,397 vertebrate taxa, which was further filtered to exclude ambiguously aligned sequences to a total of 10,309 taxa for further analysis (dataset available at [18]).

Analysis

Patterns of presence and absence of ND3–174+1in vertebrates

Of the 10,309 sequences for ND3, a normal reading frame of ND3 without an insertion at position 174 was found in all included sequences of jawless fishes (Cyclostomata, N = 30 taxa included), cartilaginous fishes (Chondrichthyes, N = 208), lungfishes (Dipnoi, N = 5), bony fishes (Actinopterygii, N = 3,511), coelacanths (Actinistia, N = 2), amphibians (Amphibia, N = 627), mammals (Mammalia, N = 1,683), snakes, lizards, worm lizards and tuatara (Lepidosauria, N = 305) and crocodiles (Crocodylia, N = 22). A single-nucleotide insertion in position 174 was observed in some lineages of turtles (Testudines, N = 141; 98 with insertion) and some lineages of birds (Aves, N = 3,775; 827 with insertion).

Because the insertion was only found in birds and turtles, we focused further analyses on Diapsida (gene alignment in Supplementary Table S1 and ND3–174 status in Supplementary Table S2). We obtained a consensus phylogenetic tree from the Open Tree of Life [19] for 3,464 taxa (Supplementary File S1), including Archelosauria (birds [N = 3,063], crocodiles [N = 22], turtles [N = 126]) and their sister group Lepidosauria (N = 253) (Fig. 1A). The insertion was absent in 2,654 diapsid taxa, while ND3–174+1 was present in 811 bird and turtle taxa. The inserted nucleotide was cytosine (C) in 749 taxa, thymine (T) in 53, guanine (G) in 7, and adenine (A) in 2 taxa.

We used ancestral state reconstruction using maximum parsimony (MP) and maximum likelihood (ML) with an equal rates model to infer the likelihoods of the presence (any nucleotide) or absence (gap in the alignment) of the insertion at each node in the tree. Branch lengths were set to equal lengths because the Open Tree of Life tree does not include branch lengths. Absence of the ND3–174+1 insertion was inferred in both MP and ML frameworks as the likely ancestral state for the common ancestor of Diapsida, the common ancestor of lizards and snakes (Lepidosauria), the common ancestor of birds and crocodiles (Archosauria) and the common ancestor of Archosauria and turtles (Archelosauria) with high probability (likelihood of absence >0.99, Fig. 1A). The common ancestor of birds (Aves) was inferred to have contained the insertion (likelihood of presence 0.98). The common ancestor of turtles (Testudines) was reconstructed without the insertion (likelihood of absence 0.97).

To quantify the number of state changes between absence and presence of the insertion, we counted the number of transitions across nodes. We chose a conservative approach for inferring a transition, which required that both parent and child nodes had a minimal likelihood of at least 0.90 for a different state. Nodes with a likelihood <0.90 in the ML reconstruction also had multiple parsimonious solutions in the MP framework. The MP and ML models therefore inferred the same number of gains and losses in this 2-state model. Across turtles and birds, a total of 33 gains and 47 losses of the ND3–174+1 insertion were inferred.

To identify the specific nucleotides involved in the insertion and their transitions, we reconstructed ML ancestral states using the 5 possible states (gap, A, T, C, G) and counted state changes as above. The 5-state model inferred high likelihoods (≥0.90) for more nodes than the 2-state model, which resulted in more inferred gains and losses of the insertion under the model with 5 states (38 gains and 52 losses). The gain of the insertion was most frequently a C (34 gains), followed by T (4 gains). Loss of the insertion happened most frequently from C (50 losses), followed by T (2 losses). When the insertion was present, most transitions were from C to T (27 transitions), while other state changes were less frequent (2 transversions from C to G, 1 transversion from C to A, 1 transversion from T to A, 1 transition from T to C).

Complex patterns of gain and loss within turtles and birds

Within turtles, 87 of 126 examined sequences had the insertion and they were dispersed across the phylogeny (Fig. 1A). Using the 2-state models, we inferred 4 losses within turtles, but no gains could be unambiguously inferred because the nodes did not have likelihoods ≥0.90 in the ancestral state reconstruction. The 5-state model inferred a complex evolution with 3 gains (likelihood >0.90), of which 2 were insertions of C and 1 was an insertion of T (Fig. 1B). The 5-state model inferred 5 losses of the insertion, of which 3 losses were from an ancestral C state and 2 losses were from an ancestral T state (Fig. 1B). Transitions from C to T were most common (7 transitions), compared to other transitions (1 T to C) and transversions (1 T to A). An insertion of G was observed in turtles in a clade of 3 species (*Malaclemys terrapin*, *Trachemys scripta*, *Chrysemys picta* in Emysternia), but the ancestral nucleotide of the parent node could not be inferred unambiguously (Fig. 1B).

Within birds, 724 of 3,063 examined sequences had the insertion. The common ancestor of birds was inferred to have had an insertion of C (likelihood 0.99). The 2-state models predicted 32 gains and 49 losses, while the 5-state model inferred 34 gains and 47 losses (likelihood >0.90). Gains were mostly of C (31 gains) or of T (3 gains). As in turtles, transitions from C to T (20 transitions) were most common, while other state changes were less frequent (2 transversions from C to G, 1 transversion from C to A). Different orders of birds had different prevalence of the ND3–174+1 insertion or its absence (Fig. 1C). The only 2 orders consistently without the frameshift insertion were the speciose perching birds (Passeriformes, N = 2,096 included) and the tropicbirds (Phaethontiformes, N = 2), and their ancestors were reconstructed without the insertion (likelihood 0.99). The insertion was present in all examined sequences of 12 orders (Pterocliformes, Columbiformes, Mesitornithiformes, Musophagiformes, Otidiformes, Opisthocomiformes, Eurypygiformes, Suliformes, Cariamiformes, Falconiformes, Coliiformes, Trogoniformes). The remaining bird orders had both lineages with and without the insertion (Fig. 1C).

Strong sequence conservation surrounding the insertion

To identify a potential shared sequence pattern around the ND3–174+1 insertion, we compared the nucleotide diversity between diapsid sequences with and without the insertion. We calculated information content (R) based on Shannon entropy for each base pair, which measures conservation of a sequence position and has a maximum value of 2 bits if the position is fully conserved. In taxa without the insertion (Lepidosauria, crocodiles, certain turtles, and certain birds), the region around position 174 (from 163 to 180 bp) of ND3 had a similar conservation level as the remainder of the gene (Fig. 2A). In contrast, all turtle and bird species with the insertion had noticeably more conserved base pairs (higher R values) around position 174 than in other regions of ND3 (Fig. 2A). The distribution of values of R around the insertion from position 163 to 180 was 20% lower
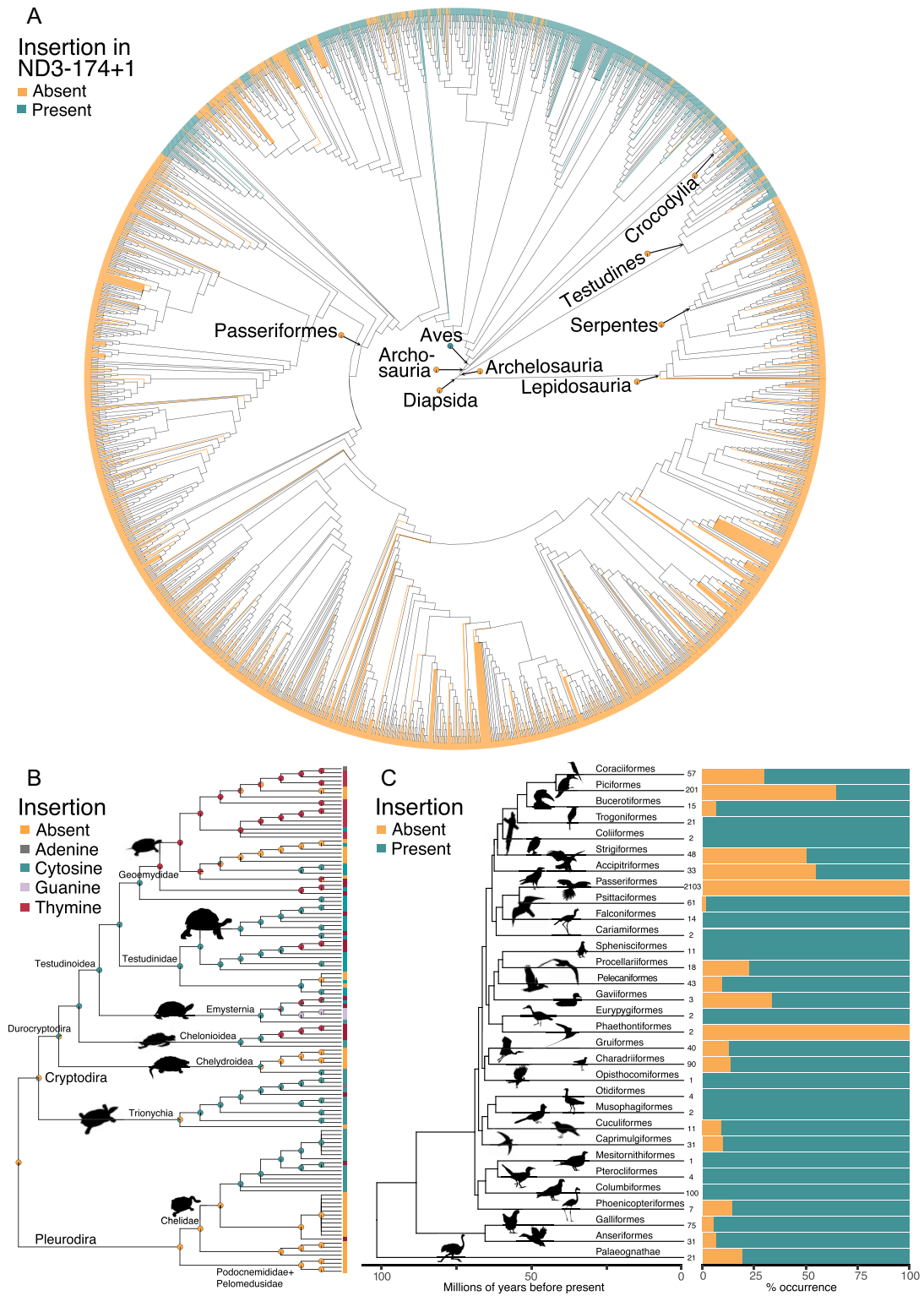
Figure 1: Phylogenetic distribution of the insertion in position 174 of the mitochondrial ND3 gene (ND3–174+1). (A) Synthetic phylogeny of 3,464 species of Diapsida with terminals colored according to the absence (orange) or presence (green) of the insertion. The corresponding figures with named terminals can be found in Supplementary Figure S1. (B) Ancestral state reconstruction for turtles with states of species indicated at the tips and pie charts showing the likelihoods of states on nodes of the phylogeny. (C) Frequency of absence (orange) or presence (green) of the insertion in each bird order. Numbers at the tips represent the number of species that were included for each order.
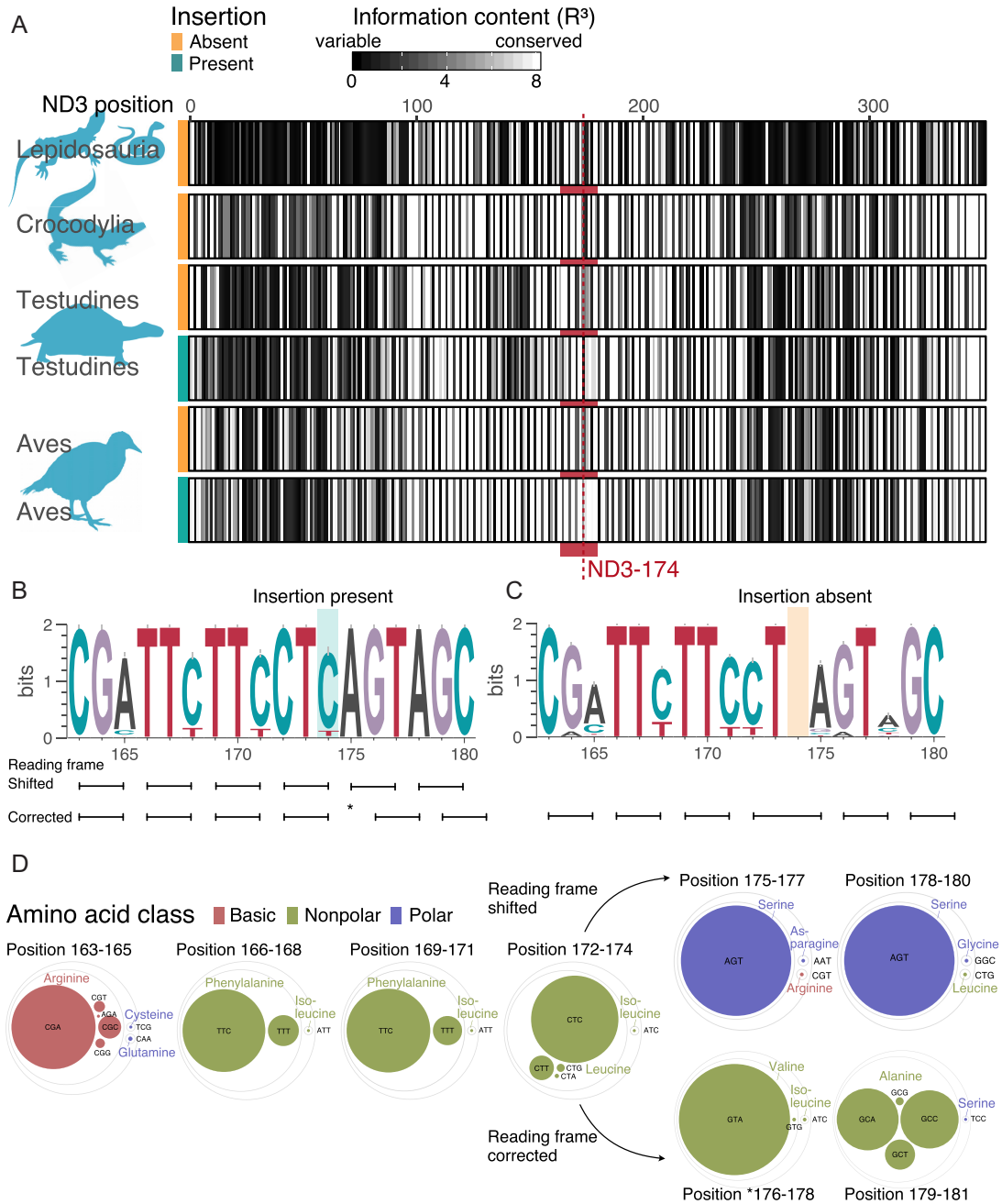
Figure 2: Nucleotide and codon usage variability in ND3 of Diapsida. (A) Information content R (cubed for visualization) across the ND3 sequence in different diapsid groups. The vertical red line marks the insertion at position 174. The red shading highlights an area of high conservation (high information content), which is only seen in birds and turtles that have the insertion. (B and C) Sequence conservation as a sequence logo from position 163 to 180 showing variability among species (B) with the insertion and (C) without the insertion. Note that the frameshift correction is thought to occur at the following nucleotide, by skipping the nucleotide A at position 175. (D) Circle packing showing the frequency of codon usage in species of birds and turtles that contain the insertion. The 2 options of the shifted and corrected reading frame following the insertion at position 174 are shown. Circle diameters indicate prevalence of a specific codon, which are grouped into larger circles if codons are synonymous. Circle color indicates amino acid class.

in species without the insertion than with the insertion (non-parametric Wilcoxon test, P-value $<10^{-3}$) (Fig. 2A). Specifically, when the insertion was present, nucleotides upstream of the insertion (position 163–174) were highly conserved with some variability on the third codon positions (Fig. 2B). The sequence downstream of the insertion (position 175–180) was completely conserved with a maximum information content (R = 2 bits, Fig. 2B). When the insertion was absent, sequence conservation

was lower (R < 2 bits), particularly on the third codon positions downstream of 174 (Fig. 2C).

We also analyzed the codon conservation in sequences containing the insertion, i.e. a combined set of turtles and birds with the insertion. The codon containing the insertion was a leucine codon (CTN) in all but one examined species, with CTC being most prevalent and the synonymous codons CTT, CTG, and CTA represented at lower frequencies (Fig. 2D). Notably, the

green wood hoopoe (*Phoeniculus purpureus*, Bucerotiformes) deviated from this pattern with an ATC encoding for isoleucine. Both leucine and isoleucine are nonpolar amino acids. For the codons following the insertion, we considered both the 0 reading frame, which is the shifted reading frame if the frameshift insertion was retained, and the +1 reading frame, which is the corrected reading frame if the A at position 175 was left out [11]. In the 0 reading frame, the first 2 codons downstream of the insertion showed almost complete conservation to AGT (encoding serine, a polar amino acid) in the first codon following the insertion (position 175–177) and AGC (encoding serine, a polar amino acid) in the second codon following the insertion (position 178–180) (Fig. 2D). The only exception to this high codon conservation following the insertion was Baillon's crake (*Porzana pusilla*, Gruiformes) with a CGT codon (encoding arginine, a basic amino acid) in the first following codon. In the +1 corrected reading frame leaving out A-175, the codon following the insertion (position 176–178) was a GTA codon (encoding valine, a nonpolar amino acid) in all sequences (Fig. 2D). The second codon following the insertion (position 179–181) was also conserved in coding for alanine (a nonpolar amino acid), albeit with all 4 synonymous codons present (Fig. 2D).

No major tRNA changes in taxa with the insertion

We investigated whether lineages containing the ND3 insertion showed differences in tRNA secondary structure, which could enhance programmed frameshifting. We aligned a set of tRNAs of leucine (CUN), serine (UCN), and valine (GUN). Leucine is the tRNAs decoding the codon where the insertion occurs serine and valine are the 2 tRNAs that compete for being decoded downstream of the insertion (Fig. 2D). We compared the predicted consensus secondary structure for each tRNA from 4 alignments, turtles and bird with and without ND3–174+1 but did not observe consistent differences in the secondary structure between bird and turtles with the frameshift insertion (Supplementary Figure S2).

Additional potential frameshifts in 5 turtle species

The translational machinery of certain birds and turtles seems to enable programmed frameshifting in order to correct single-nucleotide insertions in coding regions, and additional frameshift locations could exist. Using a subset of Diapsida sequences with full mitochondrial genomes, we checked for other frameshifts in the ND3 gene (Supplementary File S2). We did not find additional frameshift insertions in the ND3 sequences of birds (N = 703) nor in other Diapsida groups (N = 341), except for turtles. Of 106 investigated mitochondrial genomes of turtles, we identified 5 putative frameshifts upstream of position 174, namely, in *Cuora aurocapitata* (A inserted at alignment position 121; NC_009509.1)*, Cuora pani* (T deleted at alignment position 136 or 137; NC_014401.1)*, Cyclemys oldhami* (possible C or T deletion at alignment position 116; NC_023220.1), *Pelomedusa subrufa* (G inserted at alignment position 149; NC_001947.1), and *Pelusios castaneus* (G inserted at aligment position 149; NC_026049.1). The 2 *Cuora* turtles had frameshifts in different positions. *P. subrufa* and *P. castaneus* (both Pelomedusidae) had frameshifts at the same location.

For *P. castaneus*, we were able to verify that the frameshift was not due to sequencing errors by mapping high-throughput genomic sequencing reads (SRR9091361) and transcriptomic reads (SRR629649) to the mitochondrial genome (NC_026049.1). The insertion site was verified in both DNA and RNA short reads (Sup-

plementary Figure S3). We could not perform this check for the frameshift insertions in the other 4 turtle species because no additional genomic or transcriptomic data was available.

Discussion

In this work, we have inferred multiple origins of an insertion in position 174 of the mitochondrially encoded NADH dehydrogenase 3 complex gene (ND3–174+1) on the basis of a large collection of publicly available sequences for 10,309 vertebrates. This study significantly expands the sampling of previous studies (61 species in [9], 34 species in [11]) to provide a broader picture across vertebrates on one hand and more fine-scale resolution of sequence conservation on the other hand.

We confirm that the insertion is present only in turtles and birds [9, 11], but the improved sampling shows that the insertion was both more frequently gained and lost than previously thought. Different from previous interpretations, which predicted the presence of this insertion in the common ancestor of turtles and birds (Archelosauria), ML and MP ancestral state reconstruction suggested an independent evolution of this insertion in birds and turtles. In birds, the insertion was reconstructed as present in the most recent common ancestor of modern birds, which lived ~70–111 million years ago (depending on the phylogeny [20, 21]). The insertion was retained in many lineages but lost in the common ancestor of Passeriformes 39–49 million years ago [20–22] and not regained since. Additionally, within the other bird lineages, it appeared more likely to lose the insertion than to gain it (60% of changes were insertion losses, while 40% were insertion gains across all tested models). In turtles, the most recent common ancestor was inferred to not have had the insertion in ND3, as opposed to previous ideas [9, 11]. Our data included 126 turtles, 93 more than in the last study on turtle mitochondrial genomes [11], which produced an alternative interpretation of the gain and loss patterns. Within turtles, the insertion has been independently gained 1–3 times on the basis of our reconstructions. Within clades that have insertion, some lineages show mutations to other nucleotides. The dominance of C to T transitions in both turtles and birds (27 of 32 [84%] mutations) could be a consequence of cytosine methylation, which has also been described to be present in mitochondrial genomes [23].

The quality of the ND3 sequences and the observed absence or presence of the insertion at position 174 on these sequences is of crucial importance for our inferences. Most of the ND3 sequences used here originate from Sanger-sequenced ND3 genes and chromatograms may have been hand-curated for sequencing errors. While insertions at position 174 are likely to be genuine because they would have been flagged as problematic during submission to NCBI's Genbank and would require a special annotation to address the frameshift insertion (often to [9]), the absence of the insertion may be overrepresented. A frameshift insertion in ND3 may have been curated out of the sequence because such insertions in protein-coding sequences are extremely rare and could have been considered a sequencing error. It is therefore possible that the estimated number of loss events in turtles and birds is overestimated. Where losses were observed in multiple members of a clade, the most extreme case being the absence in all 2,096 included Passeriformes, the absence of the insertion is likely real. To estimate the prevalence of this potential problem, we compared ND3 annotations for 101 bird species that had ND3 sequences on NCBI's GenBank (mostly Sanger-sequenced) and also high-throughput sequenced mito-

chondrial genomes from the Bird 10,000 Genomes Project (B10K), for which we have ourselves created the annotations and can therefore exclude manual modification. Reassuringly, we found that there were no cases in which the B10K dataset contained the insertion while the GenBank sequence did not contain it. This assessment admittedly spans only a small fraction of the taxa investigated here but lends support that at least in birds, annotation errors may be limited. It is possible that submitters of bird and turtle sequences are more aware of the possibility of a frameshift insertion because the insertions have been reported from these taxa [9, 11] than submitters of taxa in which the insertion has not been observed, such as Lepidosauria or Mammalia.

It is intriguing that an insertion in exactly the same position of the ND3 gene has independently evolved 33 (MP model) to 37 times (ML model) across turtles and birds. This site specificity points to an underlying common feature that causes the frameshift to occur in this position. One possibility is that there is an increased probability to produce indels in this specific position. Alternatively, insertions may appear at a normal rate but can only be tolerated if they are embedded in a specific sequence motif that allows ribosomes to conduct the frameshift correction. This seems likely given the strong conservation at the nucleotide and at the codon level that evolved convergently in birds and turtles, despite their separate evolution for >240 million years. Intriguingly, this conservation is not found in birds and turtles without the insertion (Fig. 2A) nor in the other diapsids without the insertion, as expected with the degree of divergence of vertebrates. Furthermore, the similar sequence and codon conservation observed around ND3–174+1 have been described in other mitochondrial frameshifts across animals [11]. This indicates that the heavily conserved sequence is needed for correcting the insertion and tolerating it in the mitochondrial genome.

Most of the features observed in this study conformed with the "out-of-frame" frameshift model [11], which is characterized by a weak codon tRNA interaction in the P-site, a rare codon downstream, and an alternative codon if the reading frame is restored, which achieves a canonical Watson-Crick match with its tRNA [13]. Our extended sampling provides higher resolution of the sequence conservation features that may be involved in the programmed translational frameshift, adding examples of species that deviate from the most commonly used codons. This codon upstream from the translational frameshift position was proposed to produce a wobble pairing initiating a translation stall [11]. We found it to be highly conserved as a leucine codon (CTN) in all examined birds and turtles with the exception of 1 bird (green wood hoopoe *P. purpureus*) that uses an isoleucine codon (ATC). We further confirm that a serine codon (AGT), a polar amino acid, was always observed after the insertion [11]. However, we also show that this pattern can be more flexible at least in Baillon's crake (*P. pusilla*), which uses an arginine codon (CGT), encoding a basic amino acid (Fig. 2D). It is thought that rarely used codons, such as AGT being a rarely used codon for serine, promote the stall in the translation [11]. It is not known whether the observed variable codon CGT is also a rarely used codon for arginine in most vertebrates. In human mitochondria this seems to be the case because the CGT codon is only the third of 4 possible codons for arginine in codon preference [24]. If CGT was also rarely used in the other vertebrates, it could be an additional example of a rare codon enhancing the translational stall.

Regarding the insertion itself, we have observed all 4 nucleotides to be present in different frequencies. According to the out-of-frame frameshift model, wobble pairing between

a sequence and the tRNA anticodon promotes frameshifting [11]. Consequently, the insertion should rarely occur as an adenosine because a CTA codon produces a perfect match with the tRNA-leucine anticodon [11]. In addition to Reeve's turtle (*Chinemys reevesi*), which has previously been shown to contain an A-insertion [11], we found an independent occurrence of an A-insertion in a bird, Baillon's crake (*P. pusilla*), which also has a non-synonymous codon right after the insertion as described above. These 2 species may therefore be interesting candidates for further investigations on the programmed translational frameshift in the absence of wobble pairing.

In addition to the detailed investigation of position 174, we found 4 additional frameshifts in ND3 in 5 turtles. Of these, the African helmeted turtle (*P. subrufa*) was already described to contain such a frameshift [11] and the frameshift of the West African mud turtle (*P. castaneus*) in the same position was annotated as a frameshift insertion on NCBI. These 2 species are Pelomedusidae and both *Pelusios* and *Pelomedusa* contain a number of species [25] that could be sequenced for ND3 to investigate whether the insertion is shared across Pelomedusidae or independently obtained in the 2 species. The other 3 frameshifts are, to the best of our knowledge, potential new frameshifts in the ND3 gene. These findings and the ubiquitous presence of frameshifts in other mitochondrial genes of turtles [11] suggest a broad tolerance of turtles to frameshift insertions.

Our study demonstrates that incorporating a large number of sequences can improve resolution in inferred evolutionary patterns and give additional power to investigate sequence conservation. Our analyses suggest an independent origin of the frameshift insertion in both turtles and birds, and complex patterns of gains and losses within each group. The high sequence conservation surrounding the insertion suggests purifying selection retaining the sequence motifs needed for translational frameshifting. Nonetheless, a few species deviate from the conserved pattern. Additional losses and gains of the insertion and other deviations from the conserved motifs will likely be found once more sequences become available, within birds and turtles, and possibly also in other groups.

Potential Implications

The present work advances our understanding of the distribution of the frameshift insertion in the mitochondrial gene ND3 across the vertebrate tree of life and identifies highly conserved sequence features that seem to be associated with its occurrence. This will allow researchers to further study which sequence features allow for the corrective frameshift and to investigate the evolutionary constraints that keep the surrounding sequence heavily conserved. The fact that this insertion has remained in the mitochondrial genome for millions of years in certain birds and turtles opens the door to study the translational machinery in these lineages.

Methods

Dataset preparation

RefSeq mitochondrial genomes were downloaded from the NCBI FTP site [26] with the term "Vertebrata" contained in the taxonomy, resulting in 5,325 mitochondrial genomes of which we could retrieve the ND3 sequence in 5,320 records. We searched for nucleotide sequences containing the ND3 gene on

the GenBank nucleotide database (accessed 11 December 2019) with the query "NADH dehydrogenase subunit 3" AND Vertebrata[Organism] AND mitochondrion[filter]". A total of 92,352 sequence records were downloaded using a custom script. From those, we were able to find 88,880 records of ND3 sequences. We further included mitochondrial data from the second phase of the B10K, which produced 336 mitochondrial genomes as part of their whole-genome sequences [27]. Of these mitochondrial genomes, we included 207, which had ND3 sequences assembled and which were added if the species was not already present in the database.

Sequence alignment

One ND3 sequence for each unique taxon (10,397) was aligned using MAFFT FFT-NS-1 (v7.407) for initial alignment [28] [18]. Next, alignment positions not present in ≥5% of the sequences were removed using pxclsq (v0.1) [18] from the phyx suite [29] to exclude insertions only seen in a small proportion of taxa. Note that we also explicitly investigated other frameshifts than ND3–174+1 across ND3 in birds and turtles (section Other possible frameshift insertions), while this alignment was focused on the region around position 174. To remove potential low-quality sequences around the insertion, we removed sequences that contained a gap or an N in the 2 codons (6 nucleotides) upstream and downstream of the insertion and records that after visual examination were found not to be properly aligned around ND3–174. We performed a second alignment round on the remaining sequences with a slower and more accurate mode of MAFFT (L-INS-i). We repeated filtering positions not in ≥5% of sequences and repeated quality filters around position 174. The final filtered alignment of 10,312 sequences is given in GigaDB [18].

To confirm that all sequences of the same unique taxon had the same pattern at position 174, we used the filtered alignment as a framework for aligning all other intraspecific records, by using the –add function of MAFFT and applied the same filters as above [18].

Phylogenetic distribution of the insertion

The frameshift insertion ND3–174+1 was only observed in certain species of turtles and birds, and we therefore restricted analyses to Diapsida, i.e., birds, crocodiles, turtles, and Lepidosauria (tuatara, worm lizards, snakes, lizards). This left 4,233 of 10,397 vertebrate taxa. We recorded the state of position 174 in each sequence, either being a gap in the alignment (i.e., insertion absent) or being a nucleotide (i.e., insertion present as A, T, C, G).

We used the R package rotl (v3.0.10) [30] to obtain a phylogenetic tree for the included species. The package queries the Open Tree of Life database [31], which synthesizes phylogenetic hypotheses from published datasets and adds species that have not been included in phylogenetic analyses based on the taxonomic system [19, 32]. While a fully sampled tree for Diapsida would be preferable over a synthetic tree, it agrees in the relationships among the major Diapsida clades with phylogenetic analyses [33–36]. Of the 4,233 Diapsida species with ND3 records, 3,465 could be matched with a terminal of the Open Tree of Life tree (newick tree in Supplementary File S1). To summarize the distribution of absence or presence of the insertion on a dated bird phylogeny, we used the fossil-calibrated phylogenetic tree from [20].

Ancestral state reconstruction

ML ancestral states were reconstructed using the function hsp_mk_model in the R package castor (v1.5.5) [37]. The function first calculates the transition matrix between different states assuming equal rates for transitioning from 1 state to another and vice versa in a ML framework. We chose the equal-rates model because it makes the least assumptions about the probabilities of gain, loss, and transitions between different states. Given the known states of the tips and the phylogenetic tree, the likelihood of each node in the tree was calculated using the rerooting method [38]. Because the Open Tree of Life synthesis phylogeny did not include branch lengths, the function assumes equal branch lengths throughout the phylogeny. ML analyses were both done for 2 states (absence or presence of the insertion) and for the 5 possible states (absence, A, C, G, T).

MP reconstruction of ancestral states was performed with the function MPR in the R package ape v5.3 [39]. MP analysis was done for 2 states (absence or presence of the insertion). Prior to analysis, polytomies were arbitrarily resolved using the function multi2di in ape.

Inference of transitions between states

To count the number of nodes of the phylogeny where transitions from one state to another likely occurred, we related the likelihood from the ancestral state reconstruction of each descendant node to its parent node using the R package phangorn [40]. In the ML model, we only considered nodes with an ancestral state likelihood ≥0.90. If the likelihood was <0.90, the state was considered ambiguous. A transition was counted when a descendant node differed from its parent's state with high likelihood. This approach therefore only identifies transitions that are accompanied with strong changes in likelihoods. On the other hand, in the MP analysis, we counted transitions where both the parental and the descendant nodes had a single most parsimonious state. We counted the number of transitions between the 2 states (absence or presence of the insertion) in the ML and the MP model and the number of transitions between the 5 states (absence, A, C, G, T) in the ML model.

Sequence conservation and codon usage

Nucleotide frequencies per position across the entire ND3 sequence were obtained separately for diapsids without the insertion and with the insertion. We calculated Shannon entropy as a measure of nucleotide diversity [9]:

$$H_i = -\sum\nolimits_N \text{freq}_N * \log_2\left(\text{freq}_N\right),$$

where H is the Shannon entropy in position i of the DNA sequence and freq_N is the frequency of nucleotide N of state {A, T, G, C}. Shannon entropy was transformed into information content per nucleotide position:

$$R_i = \log_2(4) - H_i,$$

where R is the information content at position i of the DNA sequence [41]. The information content was compared between the 2 groups in the region surrounding the insertion (position 163–180) with a non-parametric Wilcoxon test and a significance threshold (α) of $P < 0.05$. Weblogo [42] was used to visualize the information content of this region (position 163–181) for both

groups. In the sequence logo, relative diversity or conservation of each nucleotide is reflected as the height of the nucleotide, measured in bits with a maximum value of 2 at complete conservation.

Codon frequencies were calculated for both the shifted reading frame, the 0 reading frame, and for the corrected reading frame, the +1 reading frame. Codons containing unknown nucleotides (N) were removed. Codon frequencies were calculated for the 6 codons surrounding the insertion (position 163–181). The calculation of codon frequencies in the +1 reading frame excluded the adenosine at position 175 (A-175) following the insertion [11].

tRNA secondary structure prediction

The sequences for 3 tRNAs involved in translation of codons surrounding the insertion were extracted from the compiled set of full mitochondrial genomes of birds and turtles. We extracted the tRNA for leucine, which translates the codon upstream of the insertion; serine, which translates the codon following the insertion in the shifted reading frame; and valine, which translates the codon following the insertion in the corrected reading frame. We split the sequences in a group of birds and a group of turtles, each with and without the insertion. Short sequences were removed if they were between 2 times the standard deviation of the mean nucleotide length of the group. An initial alignment was done using the mlocarna tRNA aligner (v2.0.0RC8) [43], which simultaneously aligns and folds RNA sequences. Two tRNAs exist for leucine and serine in mitochondrial genomes, of which CUN-leucine and UCN-serine are involved in translating the codons around the ND3–174+1. Because Genbank records do not always distinguish between the 2 tRNAs, we selected the tRNA copy with the higher pairwise identity to a known representative CUN-leucine and UCN-serine tRNA from the initial alignment. The filtered sequences were realigned with mlocarna. The predicted consensus secondary structure of the tRNA alignment was visualized using the ViennaRNA web services [44].

Other possible frameshift insertions in ND3

To extend the study to other potential insertions in the ND3 gene in Diapsida, we compiled a dataset of 1,050 complete mitochondrial genomes from the Refseq database. We excluded records without annotations or with undetermined nucleotides (N) in the ND3 gene, resulting in 1,044 ND3 gene sequences (Crocodilia N = 20, Lepidosauria N = 215, Testudines N = 106, Aves N = 703). We also investigated a second dataset focused on birds of the B10K including 328 mitochondrial genomes that were *de novo* assembled from shotgun genomic reads [27].

Both ND3 sequence sets were aligned using the vertebrate mitochondrial genetic codon table ("-gc_def 2") in the "alignSequence" module of MACSE (v2.01) [43], which respects reading frames and can tolerate frameshifts (Supplementary File S2). MACSE designates frameshifts with the symbol "!" and candidates were visually verified with the alignment viewer Seaview (v5.0.4) [45]. Individual ND3 gene sequences were also annotated using the "protein2genome" model of exonerate (v2.4.0) [46] by mapping the *Gallus gallus* ND3 amino acid sequence (NC_040902.1) to each of the diapsid ND3 nucleotide gene sequences. Exonerate designates potential frameshifts with the symbol "#."

Availability of Source Code and Requirements

Scripts used for data generation and analysis can be found at: https://github.com/sergioSEa/ND3_174_vertebrates2020
Operating system(s): Bash scripts should be run in Linux OS/Mac OS. Python and R scripts are platform independent.
Programming language: Bash, R, Python
Other requirements: Python 3 or higher, MAFFT v7.4, pxclsq v0.1.
Python packages: biopython. R packages: rotl, castor, ape, phytools, ggtree, ggimage, phangorn, ggstance, Biostrings, ggrepel, and tidyverse.
License: GNU

Data Availability

The datasets supporting the results of this article are available in the *GigaScience* database GigaDB [18].

Additional Files

Supplementary Table S1: Diapsida MAFFT alignment. Alignment of diapsid ND3 sequences after removal of positions not present in ≥5% of taxa.
Supplementary Table S2: Status at ND3–174+1 across Diapsida. Table of diapsid taxa and the corresponding status at ND3–174+1 used for ancestral state reconstruction extracted from Addtional_file_1.
Supplementary File S1: Diapsida phylogeny. Open Tree of Life synthetic phylogenetic tree matched to the Diapsida taxa included in the study.
Supplementary File S2: Diapsida MACSE alignment. Alignment of 1,044 Refseq Diapsida sequences used for identification of other frameshifts in the mitochondrial ND3 gene.
Supplementary Figure S1: Diapsida phylogeny with tip labels. Diapsida tree as presented in Fig. 1A but including taxon names.
Supplementary Figure S2: Consensus tRNA structure. Consensus predicted tRNA structure of birds and turtles with and without gap for the codons valine, leucine, and serine. Nucleotides presented are mock nucleotides that do not represent the actual consensus sequence.
Supplementary Figure S3: Confirmation of insertion in the mitochondrial genome of *Pelusios castaneus* by mapping of short read sequences. Genomic (SRR9091461) and transcriptomic (SRR629649) short reads were mapped to the *P. castaneus* mitochondrial genome (NC_026049.1) to check whether the predicted frameshift site was present in short reads as well.

Abbreviations

A: adenosine; B10K: Bird 10,000 Genomes Project; bp: base pairs; C: cytosine; G: guanine; MACSE: Multiple Alignment of Coding Sequences; MAFFT: Multiple Alignment using Fast Fourier Transform; ML: maximum likelihood; MP: maximum parsimony; N: undetermined nucleotide; NCBI: National Center for Biotechnology Information; ND3: NADH dehydrogenase 3 complex gene; T: thymine; tRNA: transfer RNA.

Competing Interests

The authors declare that they have no competing interests.

Funding

Authors' Contributions

S.A.-S.: Formal analysis, writing, visualization, conceptualization; W.C.: data curation, formal analysis, visualization; J.S.: conceptualization, supervision, visualization, writing; G.Z.: conceptualization, supervision, resources, writing.

References

1. Tse H, Cai JJ, Tsoi H-W, et al. Natural selection retains over-represented out-of-frame stop codons against frameshift peptides in prokaryotes. BMC Genomics 2010;**11**:491.

2. Atkins JF, Loughran G, Bhatt PR, et al. Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. Nucleic Acids Res 2016;**44**:7007–78.

3. Farabaugh PJ. Programmed translational frameshifting. Microbiol Rev 1996;**60**:103–34.

4. Dinman JD. Mechanisms and implications of programmed translational frameshifting. Wiley Interdiscip Rev RNA 2012;**3**:661–73.

5. Beckenbach AT, Robson SKA, Crozier RH. Single nucleotide +1 frameshifts in an apparently functional mitochondrial cytochrome b gene in ants of the genus *Polyrhachis*. J Mol Evol 2005;**60**:141–52.

6. Rosengarten RD, Sperling EA, Moreno MA, et al. The mitochondrial genome of the hexactinellid sponge *Aphrocallistes vastus*: Evidence for programmed translational frameshifting. BMC Genomics 2008;**9**:33.

7. Milbury CA, Gaffney PM. Complete mitochondrial DNA sequence of the eastern oyster *Crassostrea virginica*. Mar Biotechnol 2005;**7**:697–712.

8. Härlid A, Janke A, Arnason U. The mtDNA sequence of the ostrich and the divergence between paleognathous and neognathous birds. Mol Biol Evol 1997;**14**:754–61.

9. Mindell DP, Sorenson MD, Dimcheff DE. An extra nucleotide is not translated in mitochondrial ND3 of some birds and turtles. Mol Biol Evol 1998;**15**:1568–71.

10. Parham JF, Macey JR, Papenfuss TJ, et al. The phylogeny of Mediterranean tortoises and their close relatives based on complete mitochondrial genome sequences from museum specimens. Mol Phylogenet Evol 2006;**38**:50–64.

11. Russell RD, Beckenbach AT. Recoding of translation in turtle mitochondrial genomes: Programmed frameshift mutations and evidence of a modified genetic code. J Mol Evol 2008;**67**:682–95.

12. Seligmann H, Warthi G. Chimeric translation for mitochondrial peptides: Regular and expanded codons. Comput Struct Biotechnol J 2019;**17**:1195–202.

13. Haen KM, Pett W, Lavrov DV. Eight new mtDNA sequences of glass sponges reveal an extensive usage of +1 frameshifting in mitochondrial translation. Gene 2014;**535**:336–44.

14. Baranov PV, Gesteland RF, Atkins JF. Recoding: Translational bifurcations in gene expression. Gene 2002;**286**:187–201.

15. Huang Y, Koonin EV, Lipman DJ, et al. Selection for minimization of translational frameshifting errors as a factor in the evolution of codon usage. Nucleic Acids Res 2009;**37**:6799–810.

16. Sroga GE, Nemoto F, Kuchino Y, et al. Insertion (sufB) in the anticodon loop or base substitution (sufC) in the anticodon stem of tRNA Pro2 from *Salmonella typhimurium* induces suppression of frameshift mutations. Nucleic Acids Res 1992;**20**:3463–9.

17. Crawford NG, Parham JF, Sellas AB, et al. A phylogenomic analysis of turtles. Mol Phylogenet Evol 2015;**83**:250–7.

18. Sergio A-S, Josefin S, Wanjun C, et al. Supporting data for "Multiple origins of a frameshift insertion in a mitochondrial gene in birds and turtles." GigaScience Database 2020. http://dx.doi.org/10.5524/100839.

19. Redelings BD, Holder MT. A supertree pipeline for summarizing phylogenetic and taxonomic information for millions of species. PeerJ 2017;**5**:e3058.

20. Jarvis ED, Mirarab S, Aberer AJ, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 2014;**346**:1320–31.

21. Prum RO, Berv JS, Dornburg A, et al. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 2015;**526**:569–73.

22. Oliveros CH, Field DJ, Ksepka DT, et al. Earth history and the passerine superradiation. Proc Natl Acad Sci U S A 2019;**116**:7916–25.

23. Sirard M-A. Distribution and dynamics of mitochondrial DNA methylation in oocytes, embryos and granulosa cells. Sci Rep 2019;**9**:11937.

24. Jia W, Higgs PG. Codon usage in mitochondrial genomes: Distinguishing context-dependent mutation from translational selection. Mol Biol Evol 2008;**25**:339–51.

25. Fritz U, Branch WR, Hofmeyr MD, et al. Molecular phylogeny of African hinged and helmeted terrapins (Testudines: Pelomedusidae: *Pelusios* and *Pelomedusa*). Zool Scr 2011;**40**:115–25.

26. Refseq. ftp://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/. Accessed 3 December 2019.

27. Feng S, Stiller J, Deng Y, et al. Dense sampling of bird diversity increases power of comparative genomics. Nature 2020;**587**:252–7.

28. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 2013;**30**:772–80.

29. Brown JW, Walker JF, Smith SA. Phyx: Phylogenetic tools for unix. Bioinformatics 2017;**33**:1886–8.

30. Michonneau F, Brown JW, Winter DJ. rotl: An R package to interact with the Open Tree of Life data.Methods Ecol Evol 2016;**7**:1476–81.

31. Open Tree of Life. https://tree.opentreeoflife.org/opentree/argus/opentree12.3@ott93302. Accessed 3 December 2019.

32. Rees JA, Cranston K. Automated assembly of a reference taxonomy for phylogenetic data synthesis. Biodivers Data J 2017;**5**:e12581.

33. Green RE, Braun EL, Armstrong J, et al. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science 2014;**346**:1254449.

34. Irisarri I, Baurain D, Brinkmann H, et al. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat Ecol Evol 2017;**1**:1370–8.

35. Chiari Y, Cahais V, Galtier N, et al. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol 2012;**10**:65.

36. Crawford NG, Faircloth BC, McCormack JE, et al. More than 1000 ultraconserved elements provide evidence that tur-

tles are the sister group of archosaurs. Biol Lett 2012;**8**: 783–6.

37. Louca S, Doebeli M. Efficient comparative phylogenetics on large trees. Bioinformatics 2018;**34**:1053–5.

38. Yang Z, Kumar S, Nei M. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 1995;**141**:1641–50.

39. Paradis, E., Claude, J.and Strimmer, K.(2004) APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics **20**, 289–9014734327

40. Schliep, KP.(2011) phangorn: phylogenetic analysis in R. Bioinformatics **27**, 592–3 10.1093/bioinformatics/btq70621169378

41. Schneider TD, Stephens RM. Sequence logos: A new way to display consensus sequences. Nucl Acids Res 1990;**18**:6097–100.

42. Crooks GE, Hon G, Chandonia J-M, et al. WebLogo: A sequence logo generator. Genome Res 2004;**14**: 1188–90.

43. Smith C, Heyne S, Richter AS, et al. Freiburg RNA Tools: A web server integrating INTARNA, EXPARNA and LOCARNA. Nucleic Acids Res 2010;**38**:W373–7.

44. Kerpedjiev P, Hammer S, Hofacker IL. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics 2015;**31**:3377–9.

45. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 2010;**27**: 221–4.

46. Ranwez V, Harispe S, Delsuc F, et al. MACSE: Multiple Alignment of Coding Sequences accounting for frameshifts and stop codons. PLoS One 2011;**6**:e22594.