Molecular Phylogenetics and Evolution 201 (2024) 108221

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Genomic and transcriptomic perspectives on the origin and evolution of NUMTs in Orthoptera

Xuanzeng Liu [a], Nian Liu [a], Xuan Jing [a], Hashim Khan [a], Kaiyan Yang [a], Yanna Zheng [a], Yimeng Nie [a], Hojun Song [b,*], Yuan Huang [a,*]

[a] *College of Life Sciences, Shaanxi Normal University, Xi'an, China*
[b] *Department of Entomology, Texas A&M University, College Station, TX, USA*

ARTICLE INFO

Keywords:
NUMT
Molecular fossil
Mitochondrial evolution
NUMTs origin
Orthoptera

ABSTRACT

Nuclear mitochondrial pseudogenes (NUMTs) result from the transfer of mitochondrial DNA (mtDNA) to the nuclear genome. NUMTs, as "frozen" snapshots of mitochondria, can provide insights into diversification patterns. In this study, we analyzed the origins and insertion frequency of NUMTs using genome assembly data from ten species in Orthoptera. We found divergences between NUMTs and contemporary mtDNA in Orthoptera ranging from 0 % to 23.78 %. The results showed that the number of NUMT insertions was significantly positively correlated with the content of transposable elements in the genome. We found that 39.09 %-68.65 % of the NUMTs flanking regions (2,000 bp) contained retrotransposons, and more NUMTs originated from mitochondrial rDNA regions. Based on the analysis of the mitochondrial transcriptome, we found a potential mechanism of NUMT integration: mitochondrial transcripts are reverse transcribed into double-stranded DNA and then integrated into the genome. The probability of this mechanism occurring accounts for 0.30 %-1.02 % of total mitochondrial nuclear transfer events. Finally, based on the phylogenetic tree constructed using NUMTs and contemporary mtDNA, we provide insights into ancient evolutionary events such as species-specific "autaponumts" and "synaponumts" shared among different species, as well as post-integration duplication events.

1. Introduction

Nuclear mitochondrial pseudogenes (NUMTs) are non-functional mitochondrial DNA (mtDNA-like) fragments found in the nuclear genomes (Hazkani-Covo et al., 2010; Kleine et al., 2009; Ricchetti et al., 1999; Richly and Leister, 2004; Timmis et al., 2004), found in almost all eukaryotic genomes (Hazkani-Covo et al., 2010; Wei et al., 2022). The presence of NUMTs has long been considered both an intriguing molecular evolutionary problem and a major source of error in molecular systematics (Bensasson et al., 2001b; Song et al., 2008). How and why DNA fragments from mitochondria, which have a distinctly different mode of inheritance across generations, are continuously integrated into the nuclear genome is still not fully understood, although some hypotheses have been put forth (Bensasson et al., 2001b; Hazkani-Covo et al., 2010). From the early days of molecular systematics, mtDNA has been recognized as a versatile maker for inferring geographic genealogy and inference of phylogenetic relationships of species (Avise et al., 1987). The use of mtDNA for systematics became widely popular as scientists started designing conserved primers (often referred to as universal primers) that could be used to amplify target regions from broad taxon sampling via polymerase chain reaction (PCR) (Simon et al., 1994; Vrijenhoek, 1994). However, it was soon realized that the conserved primers would amplify not only the target mtDNA sequences but also any NUMTs that have enough sequence similarities to the target (Bensasson et al., 2001b; Collura and Stewart, 1995; Sorenson and Quinn, 1998; Thalmann et al., 2004; Zhang and Hewitt, 1996a), which could potentially compromise orthology assumptions in many areas of mitochondrial systematics, including phylogeography (Hlaing et al., 2009; Podnar et al., 2007; Thalmann et al., 2005), phylogenetics (Arctander, 1995; Lopez et al., 1997; Sunnucks and Hales, 1996), molecular ecology (Triant and Hayes, 2011), and DNA barcoding (Buhay, 2009; Song et al., 2008). The inadvertent inclusion of NUMTs in the analyses leads to incorrect evolutionary inferences.

Although the accidental co-amplification of NUMTs poses a significant challenge in mitochondrial systematics, NUMTs can also provide valuable insights into past evolutionary events, because they represent

* Corresponding authors.
E-mail addresses: Hojun.Song@ag.tamu.edu (H. Song), yuanh@snnu.edu.cn (Y. Huang).

https://doi.org/10.1016/j.ympev.2024.108221
Received 21 March 2024; Received in revised form 8 October 2024; Accepted 21 October 2024
Available online 23 October 2024

fossilized mtDNA lodged in the nuclear genome (Song et al., 2013). The mutation rate of the animal mitochondrial genome is about 10-fold faster than that of the nuclear genome (Brown et al., 1979; Brown et al., 1982; Haag-Liautard et al., 2008; Perna and Kocher, 1996), which means that the mutation rate of NUMTs slows down as soon as the NUMT integration happens (Hazkani-Covo et al., 2010; Perna and Kocher, 1996; Zhang et al., 2020b; Zhang et al., 2022). Because of the differences in genetic code between mitochondrial and nuclear genomes, NUMTs are non-functional and should follow the neutral mutation rate of the nuclear genome. Indeed, the fact that NUMTs can be amplified with mtDNA-specific primers and identified as mtDNA-like strongly supports that their mutation rate is slow. NUMT integration is ongoing throughout species diversification and can happen before and after species divergence (Bensasson et al., 2000; Perna and Kocher, 1996; Zischler et al., 1995). It is possible that the nuclear genome of distantly related contemporary species may harbor NUMTs that were integrated when the two species shared a common ancestor (Hazkani-Covo, 2009). Song et al. (2013) demonstrated this using the grasshopper genus *Schistocerca* (Orthoptera: Acrididae) as a test case by generating NUMT sequences of DNA barcoding region (cytochrome *c* oxidase subunit 1, or COI) from 21 species using PCR and cloning. By carefully comparing the sequence characteristics between the orthologous mtDNA and NUMTs, they were able to categorize NUMTs into two types. The first type was termed “synaponumt,” meaning shared derived NUMTs, similar to the concept of synapomorphy in phylogenetics. Synaponumts are NUMTs that were integrated in the common ancestor, which were subsequently passed to its descendant species. A contemporary species may harbor synaponumts from its most recent common ancestor as well as distant common ancestors. The second type was termed “autaponumt,” meaning uniquely derived NUMTs, similar to the concept of autapomorphy. Autaponumts are NUMTs that were integrated after the species divergence and uniquely found in one species, and therefore most similar to the contemporary mtDNA sequences. Song et al. (2013) suggested that NUMTs represent molecular fossils of mtDNA, which can provide useful phylogenetic insights, although they are affected by random mutations in the nuclear genome that may hinder identification.

How mtDNA fragments integrate into the nuclear genome still remains unclear. One of the prevailing hypotheses is that integration occurs via non-homologous end-joining (NHEJ) of double-stranded breaks (DSBs) (Blanchard and Schmidt, 1996; Ricchetti et al., 1999; Ricchetti et al., 2004; Wang and Timmis, 2013). This DNA-DNA process does not require short microhomology (Blanchard and Schmidt, 1996; Hazkani-Covo and Covo, 2008), and it is hypothesized that the mtDNA fragments generated by the breakage of the mitochondrial genome escape from the mitochondria and then integrate into the nuclear genome through the DNA repair mechanism. There are differences in the size, abundance, age, and chromosomal positioning of NUMTs in different species (Schiavo et al., 2017; Schiavo et al., 2018). Research on plant genomes revealed that the size of NUMTs ranged from 327 bp to 11.42 Mb, and the proportion of nuclear genomes varied from 0.0002 % to 2.08 % in different species (Zhang et al., 2020b). The number of NUMTs in bird genomes can vary 100-fold (Liang et al., 2018). In addition, studies in Hymenoptera have shown that NUMT inserts generally occur in AT-rich regions and near transposable elements (Ding et al., 2021; Wang et al., 2020).

Previous studies have shown that there appears to be a positive correlation between the abundance of NUMTs and genome size (Bensasson et al., 2001a; Bensasson et al., 2001b; Hazkani-Covo et al., 2010). Orthoptera is the only known insect group with a significantly enlarged genome (0.95–21.96 pg) (Alfsnes et al., 2017; Hanrahan and Johnston, 2011; Hawlitschek et al., 2023; Yuan et al., 2021). Several studies have demonstrated that NUMTs were rampant in all lineages of Orthoptera (Bensasson et al., 2000; Gellissen et al., 1983; Moulton et al., 2010; Song et al., 2008; Song et al., 2013; Zhang and Hewitt, 1996b). Orthoptera insects have a high proportion of transposable elements in the nuclear genome (Jing et al., 2024; Liu et al., 2022; Liu et al., 2024; Majid and Yuan, 2021), and whether these transposable elements caused the prevalence of NUMTs in Orthoptera is still unresolved.

To explore the evolutionary history and insertion characteristics of NUMTs and to explore the complex relationship between NUMT abundance, genome size, and TE content in Orthoptera, we used genome assembly data of ten species, representing both orthopteran suborders, Caelifera and Ensifera, to identify and characterize NUMTs (species list is in supplementary Table S1). We conducted additional analyses of the NUMT flanking regions to gain insight into which regions of the genome are prone to NUMT insertion. Full-length transcriptome data from three species were utilized to reveal a new potential mechanism of NUMT insertion. Taking advantage of the taxon sampling that included six closely related species, we reassessed the concept of synaponumts and autaponumts. This study provides new insights into NUMTs as molecular genetic markers and reveals ancient molecular evolutionary events that can be identified in contemporary species.

2. Materials and methods

2.1. Materials and sequencing

We included a total of ten orthopteran species for this study. They were *Xya riparia* (Saussure, 1877) (Tridactylidae), *Gryllus bimaculatus* (De Geer, 1773) (Gryllidae), *Teleogryllus occipitalis* (Serville, 1838) (Gryllidae), *Locusta migratoria* (Linnaeus, 1758) (Acrididae), and six species of *Schistocerca* (Acrididae): *S. gregaria* (Forskål, 1775), *S. cancellata* (Serville, 1838), *S. americana* (Drury, 1773), *S. piceifrons* (Walker, 1870), *S. serialis cubense* (Saussure, 1861), and *S. nitens* (Thunberg, 1815). Of these, an assembled genome was newly generated for *X. riparia*. To do this, living individuals of *X. riparia* were collected from Leshan, Sichuan, China (29°35′30.5″N, 103°20′22.6″E) in August 2019. Genomic DNA was then extracted from the hind leg of a single female using an SDS-based lysis method, and the DNA was purified with chloroform. The extracted DNA was then sonicated to a fragment size of 350 bp, before being fixed onto a microarray by bridge PCR and sequenced using the Illumina NovaSeq 6000 platform (PE150bp).

Full-length transcriptome data of *X. riparia* male and female were obtained using PacBio Sequel SMRT sequencer. The SMRTbell library was prepared and sequenced following the process of the PacBio Sequel platform (Pacific Biosciences, CA, USA). The cDNA fragments were obtained via PCR with oligo dT as a primer and screened according to length. Then, after a series of modifications, the SMRT adaptor was connected to complete library construction. The polished Circular Consensus Sequencing (CCS) data have been uploaded to the NCBI under the SRA database (accession number SRR27686112 and SRR27686113).

Raw sequencing data and genome assembly data for other species are available in the NCBI database (accession number in supplementary Tables S1 and S11).

2.2. Mitochondrial genome assembly and annotation

To ensure the generation of orthologous mitochondrial genomes to use as a reference to identify NUMTs and to avoid potential errors caused by specimen mismatches between the published mitochondrial genomes and nuclear genome sequencing samples, we reassembled the mitochondrial genomes of *Locusta migratoria* and *Xya riparia*, using the same specimen from which the nuclear genomes were generated. The mitochondrial genomes of the remaining species were assembled at the time of nuclear genome sequencing and thus did not require reassembly. We used 2 GB sequencing data as input to complete mitochondrial assembly and annotation through MitoZ (Meng et al., 2019) (mitoz all −-outprefix outputname −-clade Arthropoda −-fq1 input.fastq −-requiring_taxa Arthropoda −-kmers_megahit 31 41 51 61 71). Circos plots of two newly assembled mitochondrial genomes are in supplementary Fig. S3.

2.3. NUMT identification pipelines

We identified NUMTs by BLASTn aligning the reference mtDNA of each species to the genome (https://blast.ncbi.nlm.nih.gov/Blast.cgi, USA) (makeblastdb −dbtype nucl −in mt.fa −out mito) (blastn −query genome.fa −out Blast.out −db mito −outfmt 6 −evalue 1e-6). To avoid the influence of assembly quality on the identification of NUMTs quantity, we removed sequences of less than 20 kb from the genome file and kept the chromosome sequence for the genome assembly file at the chromosome level. Due to the Blastn linear alignment causing both the head and tail of the mtDNA, which has a circular genome, to align to the same position on the nuclear genome, we merged these two alignment results into a single identified NUMT (python NUMT_blast_allmaker.py) (https://github.com/Liuxuanzeng/NUMTs_evolution).

2.4. NUMT characteristics and insertion statistics

We used a Python script to extract NUMT fragments, and NUMT flanking regions of 25 bp, 100 bp, and 2,000 bp sequences (python NUMT_blast_allmaker.py). The GC content of NUMT flanking regions was counted using the Python script, and the insertion frequency of 13 mitochondrial protein-coding genes was calculated according to the mitochondrial genome annotation file. We visualized which regions NUMTs are from on the mitochondrial genome by using MitoZ (Meng et al., 2019)(mitoz visualize −-gb mt.gb −-depth_file numtdepth.bed −-outdir output).

2.5. De novo transposable element (TE) family identification

The TE content of the genomes was analyzed using RepeatModeler2 and RepeatMasker. RepeatModeler2 was used for the automated genomic discovery of transposable element families (https://github.com/Dfam-consortium/TETools) (BuildDatabase −name Speciesname genome.fa) (RepeatModeler −database Speciesname −LTRStruct). RepeatMasker was used to analyze genomic repeats using genome assembly data (https://repeatmasker.org) (RepeatMasker −html −gff −poly −lib merged_OTElib genome −dir output_dir).

Orthoptera insects are known to have a high proportion of TE in the genome (Liu et al., 2022; Majid and Yuan, 2021). To explore whether the DNA breaks caused by TE insertion are associated with NUMT insertions, and whether replication events after NUMT insertion into the genome are related to TEs near the insertion sites, we aligned the 2,000 bp NUMT flanking sequence to the TE library by BLASTn to analyze the TE distribution in the flanking sequence (blastn −query mt_flank_2000bp.fa −out blast.out −db TElib −outfmt 6 −evalue 1e-6 −max_hsps 1 −num_alignments 1).

2.6. NUMT duplication event identification

To explore whether the TE distribution around NUMTs could lead to duplication after NUMT insertion, we analyzed whether there were NUMT duplication events in the genome. Within each species, each NUMT and its 25 bp flanking fragments was compared, and fragments with the same NUMT flanking region were identified as having been duplicated after initial insertion into the genome. First, an internal mutual alignment was performed using the extracted NUMTs and 25 bp flanking sequences (makeblastdb −dbtype nucl −in numt_with25bp.fa −out with25) (blastn −query numt_with25bp.fa −out self25_blast.out −db with25 −outfmt 6 −evalue 1e-6). Then BLAST result files were screened using a Python script to identify post-insertion replication events (python same_NUMT.py) (https://github.com/Liuxuanzeng/NUMTs_evolution). We set the screening conditions to require that the NUMTs be the same and that there were identical sequences of more than 15 bp on both flanks (Hazkani-Covo, 2022).

2.7. Mitochondrial transcriptome analysis

Circular consensus sequences (CCS) of three species (*X. riparia, L. migratoria*, and *S. gregaria*) were obtained from subread BAM files. We used minimap2 to map CCS data to the mitochondrial genome (minimap2 −c −x map-pb mt.fa isoseq.fa > minimap.out) (Li, 2018, 2021). Mitochondrial transcripts were then obtained by removing incomplete alignments through a Python script. We searched for mitochondrial transcripts consistent with NUMTs through a Python script based on the minimap2 result and the BLASTn result of NUMTs (python3 NUMT_map_to_MT_isoseq_trans.py minimap.out spname 0.8 20 isoseq.fa) (https://github.com/Liuxuanzeng/NUMTs_evolution), sequence length differences were allowed within 20 bp. The NUMTs and their consensus mitochondrial transcripts were visualized in Geneious (https://www.geneious.com). In addition, to verify the universality of the presence of NUMTs consistent with mitochondrial transcripts observed in Orthoptera, we also performed the same analysis using the full-length transcriptome data of *Homo sapiens* and *Mus musculus* (Download from SRA database ERR6221531 and SRR23696580).

2.8. Phylogenetic analyses and divergence time estimates

To establish a phylogenetic framework for this study, we reconstructed a phylogeny using the mitochondrial genome. The taxon sampling included two outgroup species (*Periplaneta americana* and *Mantis religiosa*) and eleven ingroup species. Because our original ingroup sampling included only one member of Tridactylidae, we included an additional member of the family (*Xya japonica*) to mitigate potential long-branch attraction during the maximum likelihood tree construction. Phylogenetic analyses were performed on 13 protein-coding genes (PCGs), and multiple alignments were performed on each gene with MAFFT (v7.505) (Katoh and Standley, 2013), after which the 13 multiple sequence alignments were concatenated. Based on the best-fit model of nucleotide substitution (GTR + F + I + G4), a maximum likelihood tree was constructed using IQ-TREE (v2.2.0) (Nguyen et al., 2015). The entire analyses were completed using PhyloSuite (v1.2.3) (Xiang et al., 2023; Zhang et al., 2020a).

We estimated divergence time using mcmctree implemented in PAML (v4.9j) (mcmctree mcmctree.ctl) (Yang, 2007). We used the mitochondrial genome phylogeny without branch lengths as an input tree (ultrametric tree) and the multiple sequence alignment file as input data. We set multiple sequence alignment result file and rooted tree file with fossil calibration points as input in the configuration file (seed = -1). We selected three fossil-calibration points. The maximum range of the Tridactyloidea based only on fossils was in the Hauterivian (the minimum age was 131.15 Mya) (Sharov, 1971). The maximum range of the Acridoidea and Acrididae was in the Priabonian (the minimum age was 35.95 Mya) (Schimper and von Zittel, 1885). The base of Ensifera was rescaled to the minimum age of the earliest definitive fossil of Ensifera (the minimum age was 255.7 Mya) (Bethoux et al., 2002; Song et al., 2015). The final generated hypermetric tree file containing divergence-time estimates was viewed in FigTree (v1.4.4) (https://tree.bio.ed.ac.uk/software/figtree/).

To identify synaponumts and autaponumts, we performed a phylogenetic analysis including all identified NUMTs of COI genes and all orthologous COI genes. Multiple sequence alignment was done by MAFFT (Katoh and Standley, 2013) (mafft −-auto NUMT_mt.fa > NUMT_mt.fa.aln). based on the best-fitting model of nucleotide substitution (GTR + F + G4) predicted by ModelFinder (Kalyaanamoorthy et al., 2017), a maximum likelihood tree was constructed using (v2.2.5) (Minh et al., 2020) (iqtree2 −s NUMT_mt.fa.aln −m MFP −bb 2000 −nt AUTO). Phylogenetic trees were checked and visualized with ITOL (Letunic and Bork, 2021).

2.9. Numts and mtDNA divergence analysis

The sequence divergence between NUMTs and mtDNA was calculated according to the alignment length and the number of mismatches in the BLASTn results (p = nd / n). The divergence degree between the NUMTs of COI and the orthologous COI was calculated, and the p-distance was obtained in MEGA X according to the multiple sequence alignment results (Kumar et al., 2018). The ANOVA followed by Tukey's HSD test was performed for the multiple comparisons of p-distances (anova_result <- aov(GC_dis ~ factor(Prange, levels), data), tukey_result <- TukeyHSD(anova_result)).

2.10. COI ancestral sequence reconstruction and 3D structure prediction

First, the amino acid sequences of the COI gene of all ten ingroup species were aligned through ClustalW (Thompson et al., 2003), and then a maximum likelihood tree was constructed through MEGA X (Kumar et al., 2018). Finally, the amino acid sequences of node1 and node2 (Fig. 5) were predicted by EasyCodeML (v1.31) (Gao et al., 2019). The 3D structure of COI was obtained through online prediction by AlphaFold2 (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb) (Jumper et al., 2021) ("msa_mode"="mmseqs2_uniref_env", "model_type"="alphafold2_ptm", "num_models"=5, "num_recycles"=3). Visualization of protein structures and alignment of 3D structures were done in PyMOL (v2.5) (https://pymol.org/2/).

3. Results

3.1. Overview of NUMT insertions in Orthoptera genomes

The number of NUMT insertions in the ten orthopteran nuclear genomes ranged from 489 (*Xya riparia*) to 3,392 (*Schistocerca serialis cubense*) (Fig. 1). The total length of NUMTs in the genomes ranged from 243,595 bp (*Gryllus bimaculatus*) to 10,336,822 bp (*Schistocerca gregaria*). Based on the number and cumulative length of NUMT insertions in the genome, Acrididae had a generally higher NUMT count than Gryllidae and Tridactylidae. The genome sizes of Acrididae insects were larger than those of other families, and the number of NUMT fragments was also larger. The correlation analysis showed that the number of NUMT insertions was significantly positively correlated with genome size (Fig. 2E). Due to the small number of Orthoptera species investigated, we did not find a correlation between the evolutionary relationships among these species and the number of NUMT insertions (Fig. 1). The average length of NUMTs in the six species of *Schistocerca* with similar genome sizes varied greatly (771 bp-3,313 bp) (supplementary Table S2), as did the proportion of NUMT sequence in the genomes of these species. The NUMTs of *S. gregaria* accounted for 0.118 % of the genome, while that of *S. nitens* only accounted for 0.025 % (Fig. 1). The TE content in the genomes of the ten species of Orthoptera varied greatly, from 31.48 % of the genome in *G. bimaculatus* to 72.33 % in *L. migratoria* (Fig. 1 and supplementary Table S3). The TE content in the genomes of Acrididae was generally high, ranging from 68.51 % to 72.33 %, and there was a significantly positive correlation between the TE content and the number of NUMT insertions (Fig. 2F).

3.2. NUMT insertion frequency of mitochondrial genes and association with TE

The analysis of NUMT length distribution showed that 30–200 bp short fragment insertions accounted for a relatively high proportion in ten species (Fig. 2A). There were many long NUMT insertions (longer than 2,000 bp) in the genomes of *X. riparia* and *S. gregaria*. The analysis of the insertion frequency of NUMTs of 13 mitochondrial protein-coding regions showed that there was no particular mitochondrial gene that were most frequently inserted into the nuclear genome than the others (Fig. 2B). When we mapped the coverage depth of NUMT on the mitochondrial genome of each species (Fig. 2C), we found that the mitochondrial rRNA region had a coverage depth greater than the upper quartile, except for *Teleogryllus occipitalis* and *S. gregaria* (Fig. 2C).

Noting the high proportion of TEs in Orthoptera genomes, we investigated the distribution of TEs flanking NUMTs insertion sites. The results showed that more than 93 % of the NUMT flanks contained TE in the seven Acrididae species; the lowest was 65.67 % of the NUMT flanks by TE in *G. bimaculatus* (Fig. 2D). In addition, our analysis revealed that *retro*-transposable elements were present in 39.09 %-68.65 % of the 2,000 bp flanking NUMT. (Fig. 2D). Retrotransposons are known to cause sequence duplication and proliferation.

Among the ten species, *G. bimaculatus* had the fewest NUMT duplication events at 17 times, and *S. serialis cubense* had the most NUMT duplication events at 234 times (Fig. 1 and supplementary Table S4) (if two identical NUMT and flanking fragments are found in the genome, it

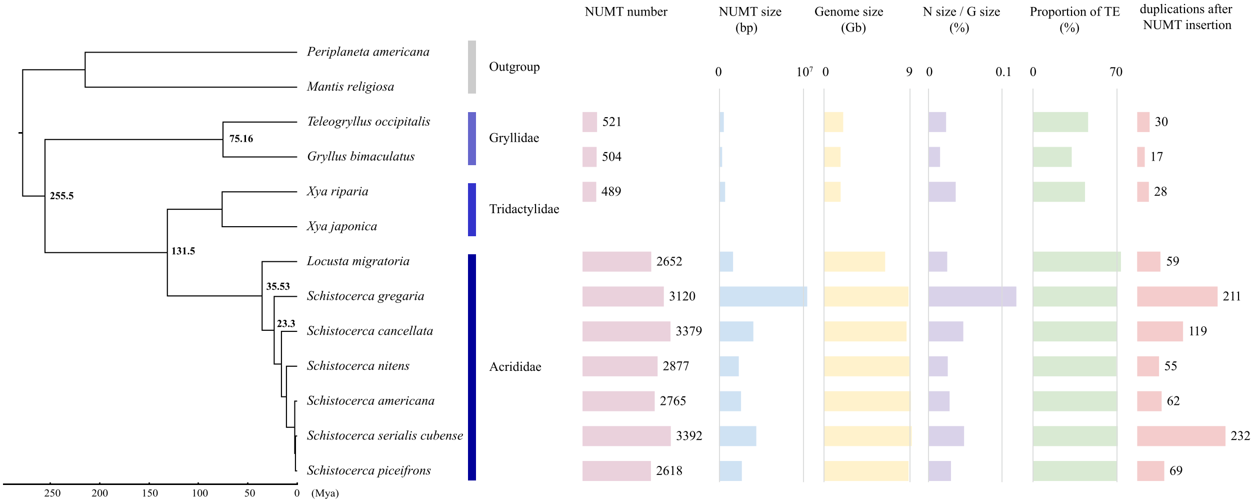

Fig. 1. Overview of NUMT insertions in Orthoptera genomes. The maximum likelihood tree of ten species on the left is constructed from 13 mitochondrial protein-coding genes. *Periplaneta americana* (Blattodea) and *Mantis religiosa* (Mantodea) were used as outgroups to construct a phylogenetic tree of Orthoptera species. Divergence time estimates were calculated using 3 fossil calibrations (see Method) and the phylogenetic tree file with 95% credibility intervals of node ages was shown in supplementary Table S10. The maximum likelihood tree with bootstrap values is shown in Fig. S4. On the right are genome features and NUMTs features. The "N size/ G size" means "NUMT size / genome size". The detailed NUMT size, genome size, and TE proportion values are in supplementary Table S2.

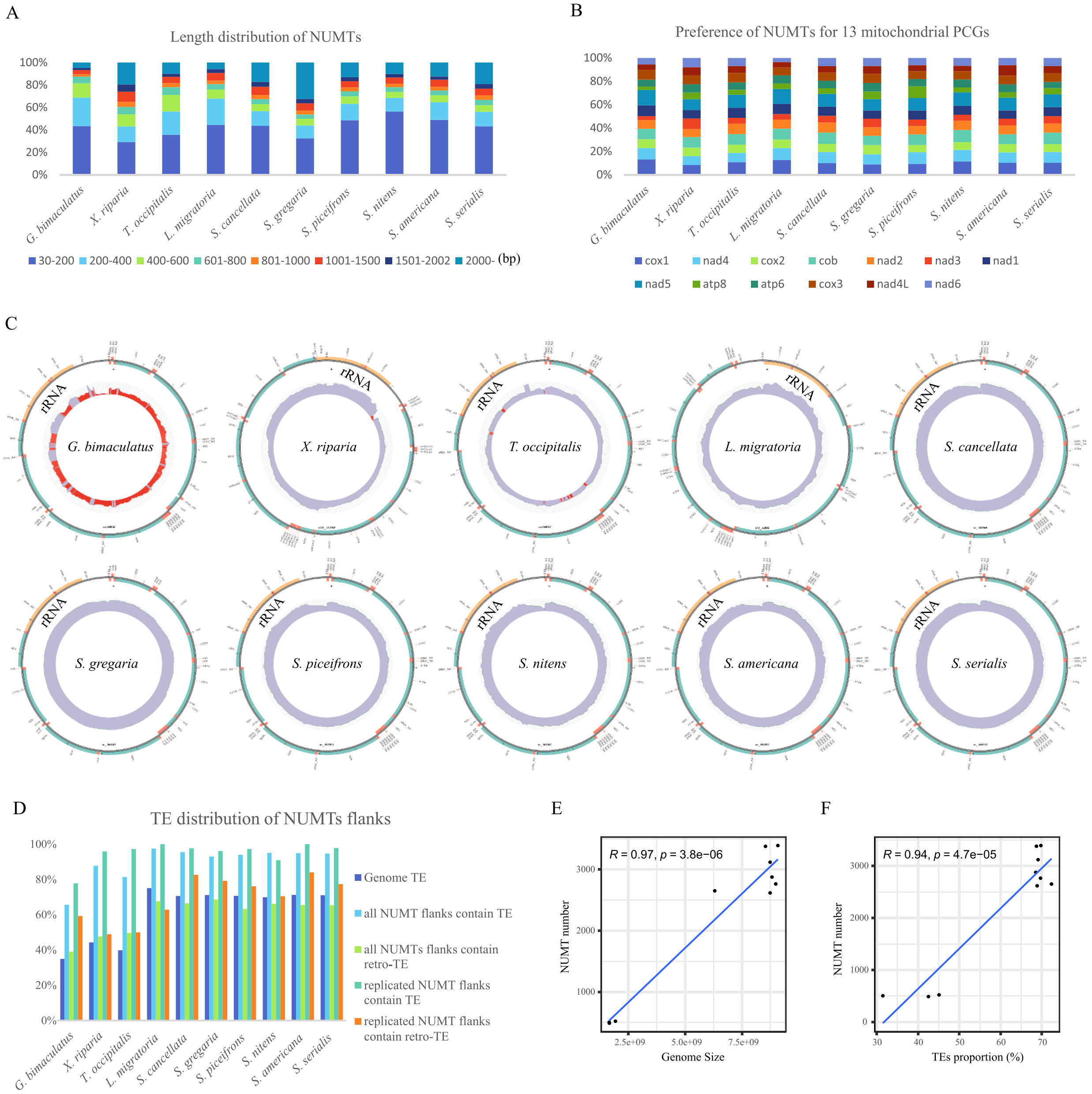

Fig. 2. NUMTs characteristic and insertion frequency analysis. (A) NUMTs length distribution statistic. (B) NUMT preference for 13 mitochondrial protein-coding genes. (C) Depth of NUMTs coverage on the mitochondrial genome. The outer circle is the mitochondrial skeleton and annotation information, and the inner circle is the coverage depth of each base in the mitochondria (standardized within the sample). The purple regions turn red if the depth is lower than the minimum value (default 20). The depth is larger than the upper quartile; the dark green outline turns purple. (D) TE distribution statistics of NUMTs flanking 2000 bp sequences. (E) Correlation analysis between NUMT number and genome size in species (R, Pearson Correlation Coefficient). (F) Correlation analysis between the number of NUMTs in species and the proportion of TEs in the genome.

was counted as one duplication event after insertion into the genome). We analyzed the 2,000 bp region flanking these duplicated NUMTs, and the results showed that the duplicated NUMT flanking regions contained a higher proportion of TEs and *retro*-TEs than other NUMT flanking regions (Fig. 2D).

3.3. The relationship between mitochondrial transcripts fraction and NUMTs

The comparison between the full-length mitochondrial transcriptomes and NUMTs showed that five mitochondrial transcripts identical with NUMTs (the sequence similarity is greater than 80 % and the length difference is within 20 bp, see Methods) were found in *X. riparia*, eight in *L. migratoria*, and 15 in *S. gregaria*. These NUMTs are considered to be nuclear integration events that may originate from mitochondrial transcripts, accounting for 1.02 %, 0.30 %, and 0.48 % of all mitochondrial nuclear transfer events, respectively. We mapped these matched mitochondrial transcripts and NUMTs onto the mitochondrial genome to explore the characteristics of these NUMTs (Fig. 3 and supplementary Fig. S1). Most of these NUMTs potentially formed from mitochondrial transcripts are complete gene fragments, and these matching transcripts are short mitochondrial secondary transcripts. Based on the presence of deletion mutations and stop codons in these transcript-identical NUMTs (Fig. 3), we excluded the possibility that these transcripts were derived from the nuclear genome. We speculated that these NUMTs may be caused by the integration of mitochondrial transcripts after reverse transcription, and the probability of this event occurring is 0.30 %-1.02 %. To explore the generality of the hypothesis

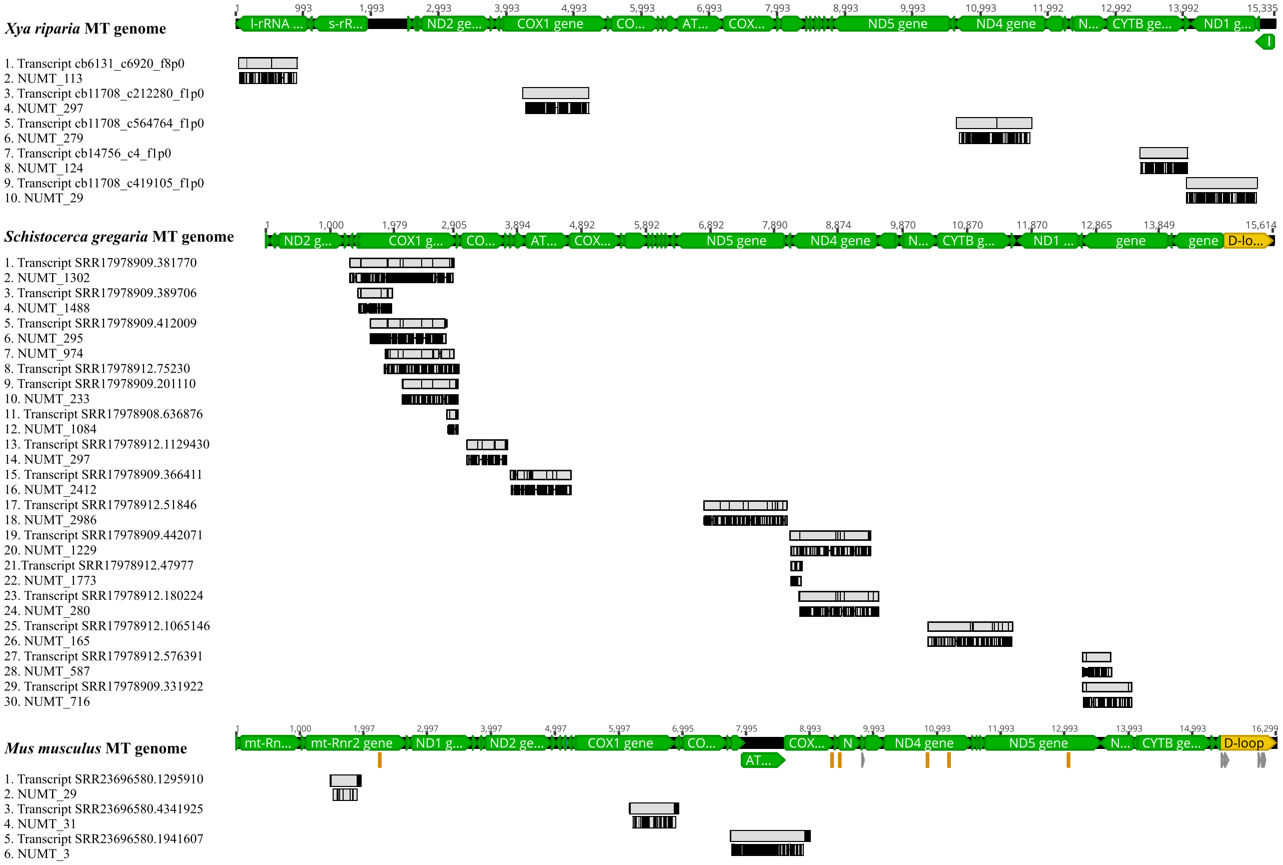
Xya riparia MT genome
1 993 1,993 2,993 3,993 4,993 5,993 6,993 7,993 8,993 9,993 10,993 11,992 12,992 13,992 15,335
l-rRNA ... s-rR... ND2 ge... COX1 gene CO... AT... COX... ND5 gene ND4 gene N... CYTB ge... ND1 g...
1. Transcript cb6131_c6920_f8p0
2. NUMT_113
3. Transcript cb11708_c212280_f1p0
4. NUMT_297
5. Transcript cb11708_c564764_f1p0
6. NUMT_279
7. Transcript cb14756_c4_f1p0
8. NUMT_124
9. Transcript cb11708_c419105_f1p0
10. NUMT_29
Schistocerca gregaria MT genome
1 1,000 1,979 2,905 3,894 4,892 5,892 6,892 7,890 8,874 9,870 10,870 11,870 12,865 13,849 15,614
ND2 g... COX1 g... CO... AT... COX... ND5 gene ND4 gene N... CYTB g... ND1 ... gene gene D-lo...
1. Transcript SRR17978909.381770
2. NUMT_1302
3. Transcript SRR17978909.389706
4. NUMT_1488
5. Transcript SRR17978909.412009
6. NUMT_295
7. NUMT_974
8. Transcript SRR17978912.75230
9. Transcript SRR17978909.201110
10. NUMT_233
11. Transcript SRR17978908.636876
12. NUMT_1084
13. Transcript SRR17978912.1129430
14. NUMT_297
15. Transcript SRR17978909.366411
16. NUMT_2412
17. Transcript SRR17978912.51846
18. NUMT_2986
19. Transcript SRR17978909.442071
20. NUMT_1229
21.Transcript SRR17978912.47977
22. NUMT_1773
23. Transcript SRR17978912.180224
24. NUMT_280
25. Transcript SRR17978912.1065146
26. NUMT_165
27. Transcript SRR17978912.576391
28. NUMT_587
29. Transcript SRR17978909.331922
30. NUMT_716
Mus musculus MT genome
1 1,000 1,997 2,997 3,997 4,997 5,997 6,995 7,995 8,993 9,993 10,993 11,993 12,993 13,993 14,993 16,299
mt-Rn... mt-Rnr2 gene ND1 g... ND2 ge... COX1 gene CO... AT... COX... N ND4 gene ND5 gene N... CYTB ge... D-loop
1. Transcript SRR23696580.1295910
2. NUMT_29
3. Transcript SRR23696580.4341925
4. NUMT_31
5. Transcript SRR23696580.1941607
6. NUMT_3

Fig. 3. NUMTs of mitochondrial transcript origin and transcripts mapping on the mitochondrial genome. The green axis represents the mitochondrial genome and annotation information. Gray boxes represent mitochondrial transcripts and NUMTs, whereas black vertical lines indicate mutation sites. The results for *Locusta migratoria* and *Homo sapiens* are shown in supplementary Fig. S1.

that some NUMTs originate from mitochondrial transcripts, we analyzed the mitochondrial transcriptomes of *Homo sapiens* and *Mus musculus*. The results showed that NUMTs consistent with mitochondrial transcripts exist in the genomes of *H. sapiens* and *M. musculus* (Fig. 3 and supplementary Fig. S1).

3.4. Numts evolution dynamics analysis

To infer the timing of NUMT insertion into the genome, we calculated the divergence (the pairwise distance) of NUMTs from the contemporary mitochondrial genomes. We observed that smaller genetic divergences between NUMTs and mitochondrial genomes indicated more recent transfers from the mitochondria to the nuclear genome. Conversely, larger genetic divergences suggested ancient NUMT insertion events. The p-distance values among ten species ranged from 0 to 0.2378 (Fig. 4A and supplementary Table S5). The variation range of NUMT p-distance within the species of the same genus was similar. The overall p-distances of NUMT for each species in the genus *Schistocerca* were significantly lower than *G. bimaculatus*. The p-distances of NUMTs in *L. migratoria* and *T. occipitalis* were not significantly different from those in *G. bimaculatus* (Fig. 4A). The p-distance of NUMTs in *X. riparia* was the largest, higher than that of other species.

In addition, we screened NUMTs containing the full-length COI and extracted NUMTs of COI. There was only one NUMT in the *G. bimaculatus* genome containing the complete COI, while *S. gregaria* had 339 full-length COI NUMTs. To assess the divergence between COI NUMTs and orthologous COI in each species, we calculated the p-distance, which ranged from 0 to 0.168 across the ten species (Fig. 4B and supplementary Table S6). We found that the overall divergence between COI NUMT and orthologous COI was lower than the divergence of all NUMTs. Overall, the p-distance distribution of COI NUMTs was consistent with the p-distance distribution of all NUMTs, with the p-distance of *X. riparia* being higher and that of *S. gregaria* (Fig. 4A and B).

The average value of the NUMT p-distance in each species was calculated, with the largest value of 0.114 for *X. riparia* and the smallest value of 0.059 for *S. gregaria* (supplementary Table S7). We found that the *X. riparia* with the smallest number of NUMTs had the highest p-distance, while the *Schistocerca* genomes with a higher number of NUMTs corresponded to a lower p-distance. The correlation analysis showed a significant negative correlation between the average p-distance among NUMTs and the number of insertions (Fig. 4C). These indicate that the lower the mean p-distance, the more frequent recent NUMT insertion events in the genome, and the higher the number of NUMTs in the genome.

In addition, we created histograms showing the accumulated number of NUMTs arranged according to the p-distance, which we refer to as the NUMT divergence landscape (Fig. 4D). This landscape shows the distribution of types of NUMTs inserted throughout an evolutionary timescale, using the p-distance as a proxy for time. When the peak of the landscape is skewed to the left, it indicates that the insertion of NUMTs has been more active in the recent. When the peak is skewed to the right, it indicates that a large number of NUMTs were inserted in the ancient period. We found that the peaks of the divergence landscapes of Acrididae with a larger number of NUMT insertions are skewed to the left, and the p-distances corresponding to the peaks are less than 10 % (Fig. 4D). *Xya riparia* had the smallest number of NUMT insertions, and the p-distance corresponding to the peak of its landscape reaches 15 % (Fig. 4D). The two species of Gryllidae had a smaller number of NUMTs insertions, and their landscapes contained two peaks located near the center (divergence distance > 10 %).

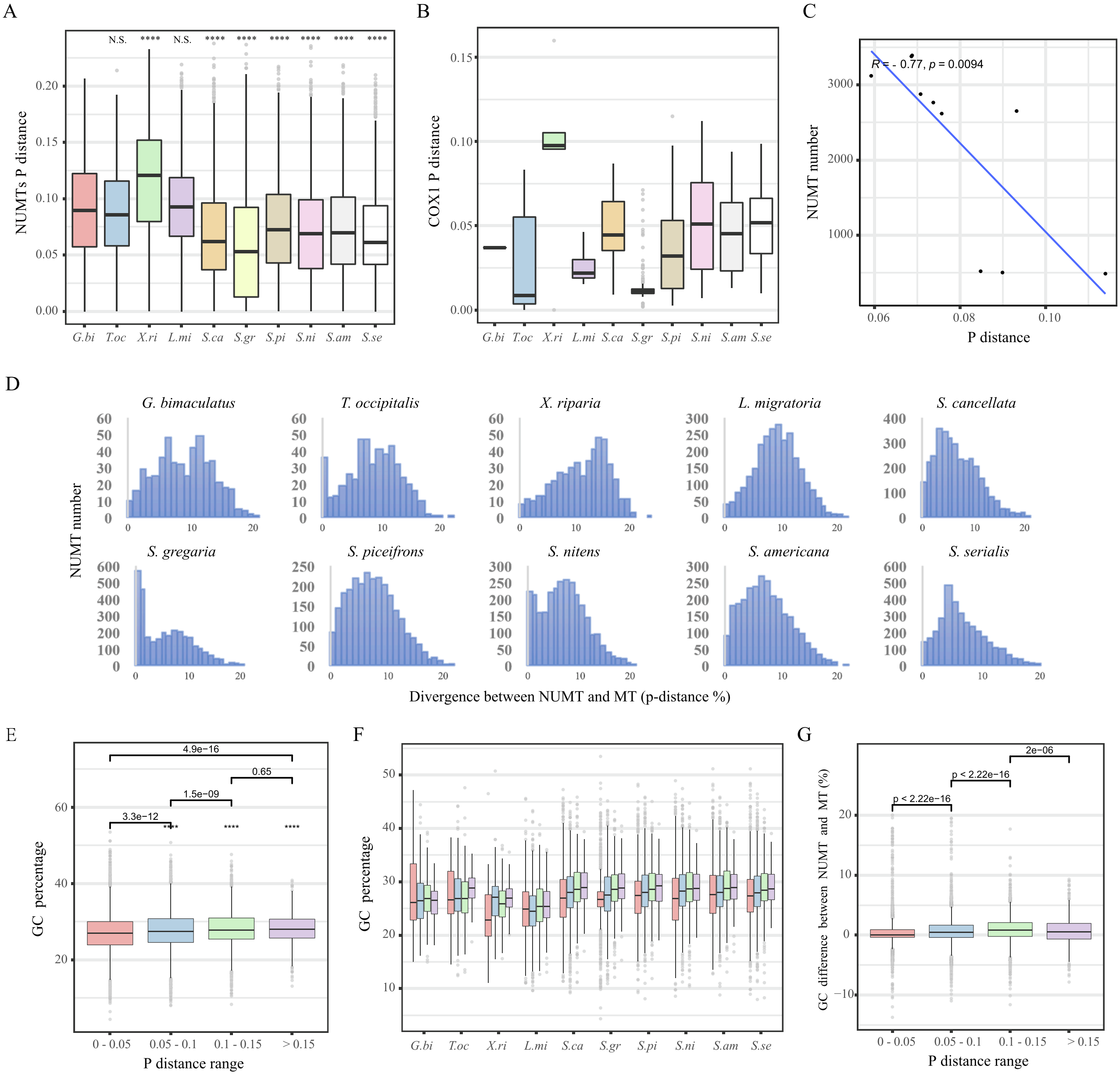

Fig. 4. Evolutionary dynamics of NUMTs and inference of ancient mitochondrial characteristics. The genetic divergence between NUMT and the mitochondrial genome is calculated by the pairwise distance (p-distance). (A) p-distance statistics for all NUMTs. The NUMTs p-distances of the nine species were analyzed for significance differences between them and *G. bimaculatus* (G.bi), respectively. Significance is denoted by * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; NS $p > 0.05$ (method = *t*-test). (B) p-distance statistics of NUMTs COI full-length sequence. (C) Correlation analysis between species' average p-distance and the number of NUMTs in the genome (method = Pearson correlation test). (D) Divergence landscape of NUMTs (Histogram, x-axis interval is 1 % p-distance). (E) Summary plot of GC content of NUMTs at different stages for ten species; the results for each species are in figure (F). (G) Summary plot of the difference in GC content between each NUMT and the corresponding mtDNA in the four stages. The ANOVA followed by Tukey's HSD test was performed for the multiple comparisons of p-distances in Fig. 4A,E,G. The values of Tukey's HSD test were shown in Table S12.

3.5. Evolutionary dynamics of NUMTs upon nuclear integration

We categorized the recovered NUMTs into four different stages of insertion according to the p-distances between NUMT and mtDNA, reflecting the age of NUMT insertions (stage 1: p-distance = 0–0.05; stage 2: p-distance = 0.05–0.1; stage 3: p-distance = 0.1–0.15; stage 4: p-distance > 0.15). Smaller p-distances indicate recent NUMT insertions, and larger p-distances indicate more ancient NUMT insertions. We also examined sequence characteristics of NUMTs by calculating the GC content. Mitochondrial genomes are AT-biased, while nuclear genomes are GC-biased. Thus, it is expected that recent NUMTs should retain relatively high AT content (or low GC content), while ancient NUMTs should accumulate more GC. As expected, we found that the GC content of NUMTs in stages 2, 3, and 4 was significantly higher than that in stage 1 ($p < 0.001$) (Fig. 4E,F and supplementary Table S8). The GC content gradually increased from stage 1 to stage 2 and then to stage 3, but no significant difference was found in the GC content of stage 4 and stage 3 ($p > 0.05$) (Fig. 4E,F). In addition, we also calculated the difference in GC content between NUMT and the corresponding mtDNA. If it is greater than 0, it means that the GC content of the NUMTs is increasing over evolutionary timescale. Judging from the overall results of the ten species, the average difference between NUMTs and mitochondrial GC content in the four stages is all greater than 0 (Fig. 4G and supplementary Fig. S2A). It demonstrates that once mtDNA is inserted into the nuclear genome, it no longer maintains its sequence integrity, and instead begins to follow the substitution rates and mutation patterns of the nuclear genome.

To explore the evolution of mitochondrial genome nucleotide sequences, we extracted all COI −containing NUMTs in each species for multiple sequence alignment. Using the COI CDS of *Schistocerca serialis cubense* as the reference sequence, the multiple sequence alignment results of 15 NUMTs of the indel occurring at 582–583 bp are shown in Fig. 5A. Compared with NUMTs, mitochondrial COI of ten species has a deletion mutation at 582–583 bp, and the number of deleted bases is an

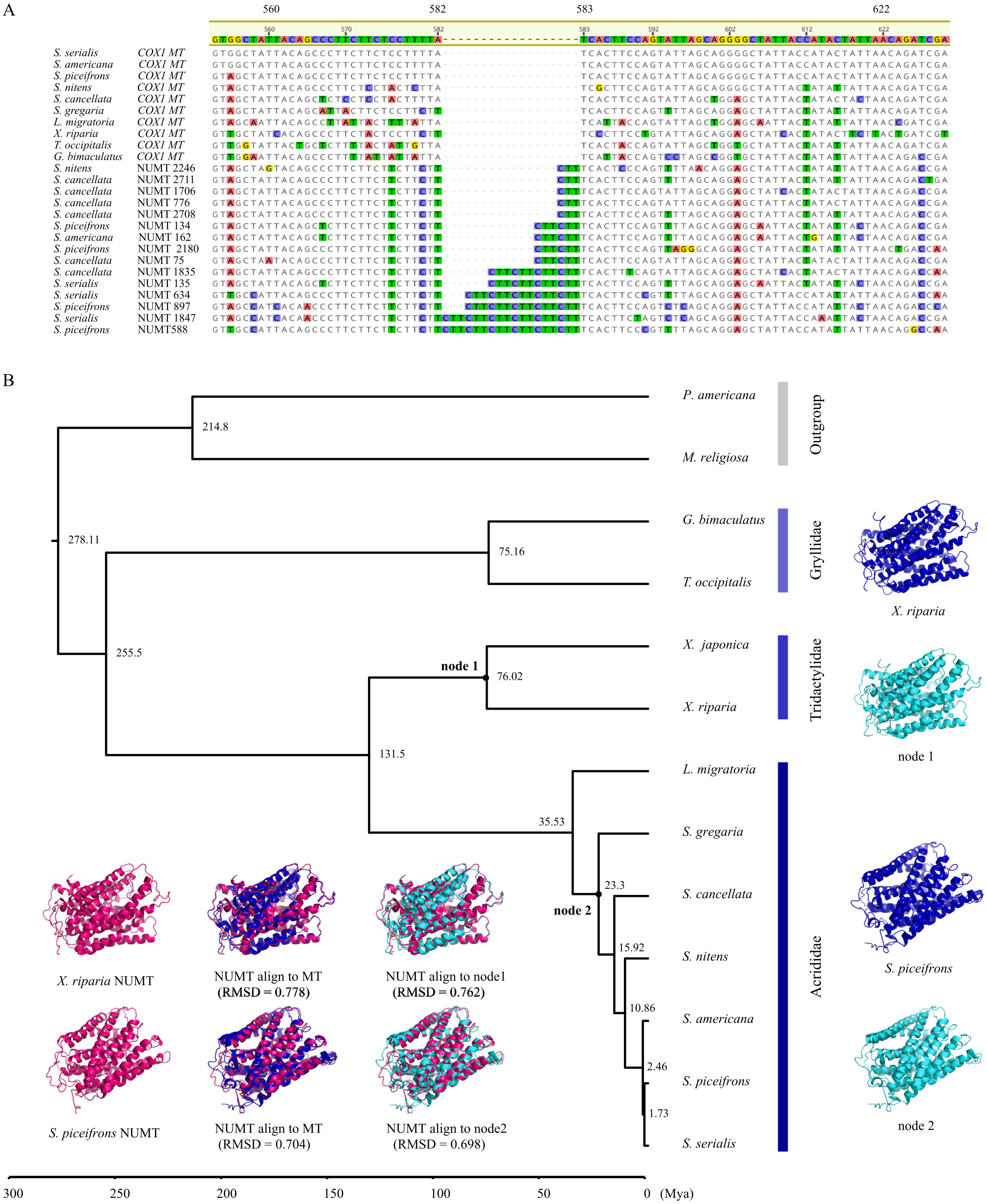

Fig. 5. Cytochrome *c* oxidase subunit 1 (COI) nucleotide sequence and protein structure evolution inference. (A) Multiple sequence alignment of mitochondrial COI and NUMT COI, using mitochondrial COI of *S. serialis* as the reference sequence. Some NUMTs suspected of having indels are shown in the figure. (B) Analysis of the protein structural evolution of COI. The maximum likelihood (ML) tree was constructed from 13 mitochondrial protein-coding genes. Species divergence time estimation based on ML trees was done by PAML, using three fossil nodes. The amino acid sequences of node1 and node2 are ancestral sequence reconstructions (ASR) of COI by PAML. The 3D structure of the blue protein represents the current mitochondrial COI, the 3D structure of the cyan protein represents the ancestral COI sequence, and the 3D structure of the magenta protein represents the oldest NUMT COI sequence. Similarity between protein structures was assessed by root-mean-square deviation (RMSD). The maximum likelihood tree with bootstrap values is shown in Fig. S4.

integral multiple of three. In addition, most mutations in NUMTs are synonymous mutations, such as amino acid 194 of COI (580–582 bp). These NUMTs all used CTT to encode leucine, while most current mitochondrial species use TTA to encode leucine (Fig. 5A).

Finally, we predicted the 3D structures of NUMTs, mtDNA, and ancestral COI sequences by AlphaFold2 (see Methods), exploring the 3D structural features of ancient mitochondrial coding genes provided in NUMTs under the assumption that the mutation rate in NUMTs is slow. First, we performed ancestral sequence reconstruction (ASR) based on the COI amino acid sequences of 13 species and predicted the ancestral COI amino acid sequences of the most recent common ancestor of *Xya* (node 1) and the most recent common ancestor of *Schistocerca* (node 2). Secondly, the two oldest full-length COI NUMTs (maximum p-distances) were selected (*X. riparia* NUMT265 and *Schistocerca piceifrons* NUMT389) and compared with the contemporary mitochondrial and the reconstructed ancestral COI protein structures, respectively. The comparison of protein structures showed that the conformation of COI NUMT was less different from the ancestral COI than the conformation of the contemporary COI (Fig. 5B). The conformational difference between the COI NUMT of *X. riparia* and the ancestral *Xya* COI (0.762, Root Mean Square Deviation) was smaller than the difference with the contemporary COI (0.778, RMSD). Similarly, the conformational difference between the COI NUMT and the ancestral *Schistocerca* COI (0.698, RMSD) was smaller than the difference with the contemporary COI (0.704, RMSD). These results suggest that NUMTs can provide more clues about the protein structures of the ancestral mitochondria.

3.6. NUMT as a molecular genetic marker for phylogenetic inference

We constructed a phylogenetic tree for a dataset of all COI NUMTs and all orthologous COI sequences in ten species. For each of the four species (*G. bimaculatus*, *T. occipitalis*, *X. riparia*, and *L. migratoria*) from different genera, all NUMT sequences formed a monophyletic group with the orthologous COI (Fig. 6). According to the category of NUMTs proposed by Song et al. (2013), we categorized these NUMTs as autaponumts, which were integrated after species divergence. We did not recover a clade consisting of only NUMTs of two cricket species (*G. bimaculatus* and *T. occipitalis*), suggesting that any ancient NUMT that were integrated into the nuclear genome of the common ancestor of these two cricket species must have mutated beyond recognition. Our taxon sampling included six closely related species in the genus *Schistocerca* and provided an opportunity to observe and identify diverse

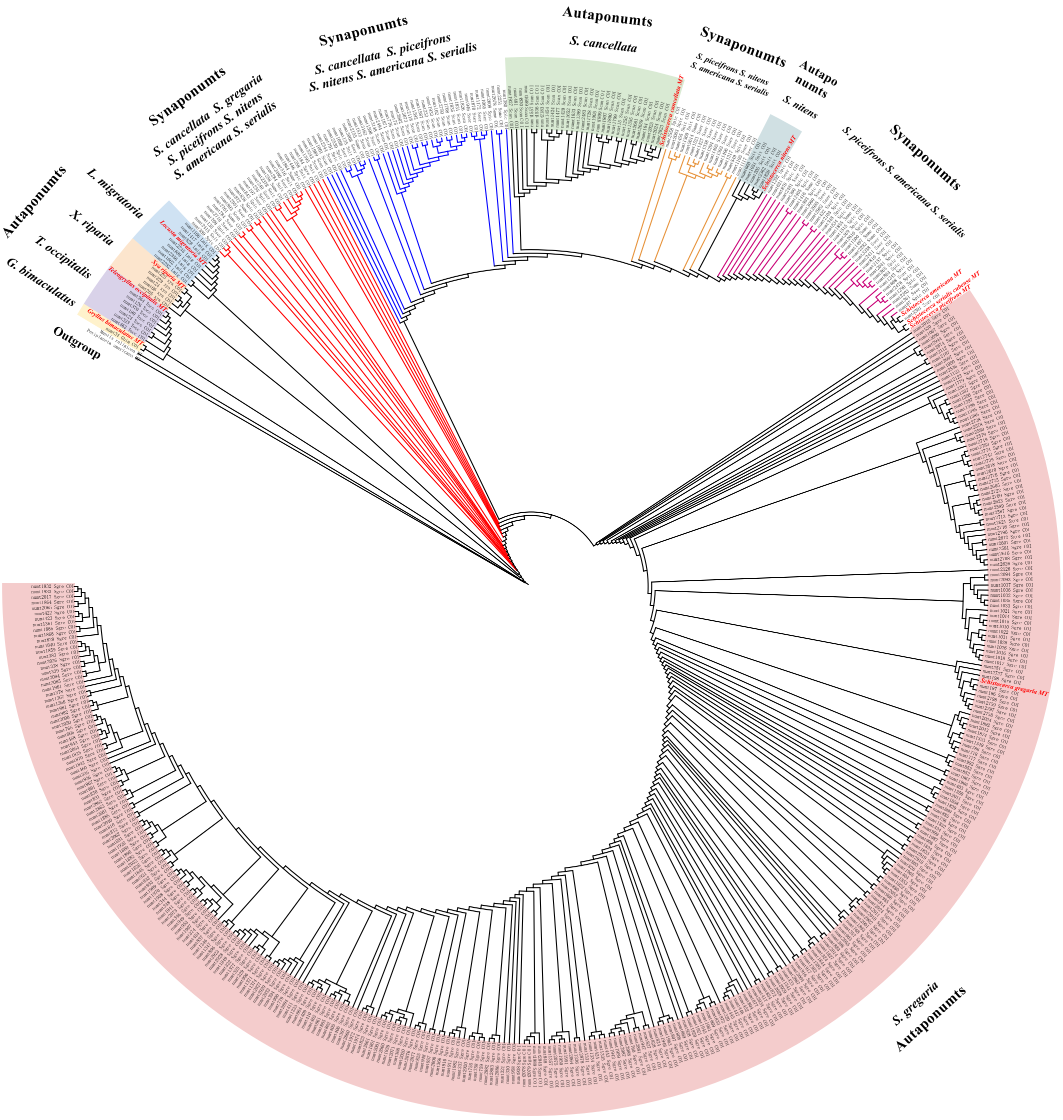

Fig. 6. A maximum likelihood tree for a dataset of all NUMTs COI and ten mitochondrial COI in ten species. The red label indicates the current mitochondrial COI of each species, and the monophyletic clade composed of autaponumts and their respective mitochondrial COI is represented by the colored area background on the label. The collapsed *S. gregaria* 338 Autaponumts tree was shown in Fig. S5, and the tree with branch lengths was shown in Fig. S6.

NUMTs. Both synaponumts and autaponumts were identified based on the positions of NUMTs in the overall phylogeny (Fig. 6). The distribution of synaponumts along the phylogeny was consistent with the divergence sequence of *Schistocerca*. Within *Schistocerca*, *S. gregaria* is the earliest diverging lineage. The first major clades of synaponumts (colored in red in Fig. 6) consisted of NUMTs of all six *Schistocerca* species, and they are place at the base of *Schistocerca*. These are ancient NUMTs that were inserted into the nuclear genome of the common ancestor of the six *Schistocerca* species which have been retained in the genomes of the contemporary species. The second major clades of *S. gregaria* autaponumts (red background labels in Fig. 6) emerged after the major clades of six species synaponumts. The third major clades of synaponumts (colored in blue in Fig. 6) consisted of NUMTs that were inserted at the common ancestor of five *Schistocerca* species after *S. gregaria* diverged. As expected, none of the NUMTs from *S. gregaria* grouped in these clades. This suggests that the nuclear genomes of contemporary species contain NUMTs that were integrated in different common ancestors throughout the phylogeny. We also found that the earlier species diverged in *Schistocerca*, the greater the number of autaponumts they contained.

4. Discussion

4.1. Challenges of NUMT identification in mitochondrial systematics

The fundamental assumption of PCR is that specific primers target specific orthologs. Conserved primers targeting mtDNA can amplify not only the target orthologous mtDNA fragments, but also NUMTs, which violates the assumption. In addition, the conserved primers can co-amplify microheteroplamsy caused by somatic mutations (Densmore et al., 1985), aging (Lin et al., 2002), biparental inheritance of mitochondria (Hoeh et al., 1991), or paternal leakage (Gyllensten et al., 1991; Kondo et al., 1990). In the absence of a complete genome, it becomes challenging to distinguish the orthologous mtDNA fragments from recently integrated NUMTs or microheteroplasmy. Our study shows that the sequence divergence between orthologous mtDNA and NUMTs ranges between 0 % and 23.78 % among these species. This finding implies that PCR-based mitochondrial systematics studies must be aware of inadvertent co-amplification or even preferential amplification on non-mtDNA sequences (Moulton et al., 2010).

NUMTs also interfere with species richness assessments (Hebert et al., 2023). Those NUMTs with > 2 % divergence introduce complexities to methods for species identification using mitochondrial markers, such as the COI region used for DNA barcoding (Hebert et al., 2003). Our results indicate many NUMTs with greater divergence in the DNA barcode region of COI (Fig. 4B), which may affect the evaluation of Orthoptera species richness.

4.2. Potential novel mechanism of NUMT insertion

Currently, the prevailing view is that NUMTs originate from escaped mtDNA fragments in mitochondria, which integrate into the nuclear genome at double-stranded DNA breaks (DSBs) via non-homologous end-joining (NHEJ) repair machinery (Blanchard and Schmidt, 1996; Hazkani-Covo and Covo, 2008). Orthopteran insects are known to have significantly enlarged genomes (Alfsnes et al., 2017), which contain a very large amount of TEs (Liu et al., 2022). These active TEs jump frequently in the genome, which may increase the frequency of double-strand break repair, resulting in a larger number of NUMTs in the Orthoptera genome. Our research shows that most NUMT insertion sites contain TEs on their flanks (Fig. 2D), and some studies have also shown that NUMTs are preferentially inserted into TE and repetitive sequence regions in the genome (Mishmar et al., 2004; Wang et al., 2020; Zhang et al., 2022). The above evidence supports that the double break repair of the nuclear genome caused by frequent TE insertion may promote NUMT integration. In addition, we found that the rRNA region of mtDNA was more frequently inserted into the nuclear genome than other mitochondrial genes (Fig. 2C). Since the transcript abundance of the mitochondrial rRNA region is high, this pattern may imply that these NUMTs could be inserted into the nuclear genome via RNA reverse transcription from mitochondrial rRNA. Overall, it seems that most NUMTs originated from random breaks in mitochondrial DNA, and a small number of NUMTs could have originated from reverse transcription of mitochondrial rRNA transcripts and other mitochondrial transcripts.

Notably, previous studies on grasshopper genomes revealed that retrotransposon transcript abundance was higher in the species with larger genomes than those with smaller genomes (Liu et al., 2022). The reverse transcriptase (RT) domain and integrase (INT) domain contained in the *retro*-TE transcript can help reverse transcription and integration of *retro*-TE transcript (Kazazian Jr, 2004; Piégu et al., 2015). Do the transposases (reverse transcriptase and integrase) carried in *retro*-TE help transfer mitochondrial fragments to the nuclear genome? We proposed that some mitochondrial transcripts are reverse transcribed into DNA fragments with the help of reverse transcriptase and inserted into the nuclear genome with the help of integrase. Our study of the full-length transcriptomes of three orthopteran species found several NUMTs that are consistent with mitochondrial transcripts (Fig. 3), and these NUMTs accounted for 0.30 %-1.02 % of the total NUMTs. Based on the research on mitochondrial transcripts, we propose a hypothesis that there is a potentially mechanism for mitochondrial nuclear transfer events: mitochondrial transcripts are reverse-transcribed into double-stranded DNA and then integrated into the nuclear genome. Although we currently do not have enough evidence that these mitochondrial transcripts have specific binding sites for transposase (reverse transcriptase and integrase), this integration process does not necessarily have to be integrated into the nuclear genome with the help of transposase. It can also be through DNA DSB repair machinery with or without the requirement of short microhomology (Blanchard and Schmidt, 1996; Hazkani-Covo and Covo, 2008). Schuster and Brennicke (1987) proposed the hypothesis that interorganellar transfer of genetic information may occur via RNA and subsequent local reverse transcription and genomic integration in higher plants (Schuster and Brennicke, 1987). They found fragments in the mitochondrial genome that were homologous to the chloroplast genome and the nuclear genome, and also found an open reading frame containing the central domain coding for the reverse transcriptase of the transposon- and retrovirus-encoded polyproteins in the middle of this homologous region, and verified the existence of this mRNA through RNA blot analysis. Based on these findings, we proposed a hypothesis that the transfer of genetic information from mitochondria to the nuclear genome may also involve the mechanism of integration after RNA reverse transcription. In addition, the mechanism of organellar genome fragment transfer via reverse transcription is universal, and we have also discovered NUMTs in *H. sapiens* and *M. musculus* that may integrate via reverse transcription (supplementary Fig. S1).

4.3. Insertion number and duplication events of NUMTs in the nuclear genome

The frequency of intracellular escape of mitochondrial DNA fragments and their integration into the nuclear genome is estimated to be 10^{-3} to 10^{-4} per cell per generation (Thorsness and Fox, 1990; Thorsness and Weber, 1996). Our results showed that any region of the mtDNA has the potential to be integrated into the nuclear genome as we found NUMTs of every single mitochondrial gene in all of the species examined, suggesting that the fragmentation and intracellular escape of mitochondrial DNA fragments may be a random process. We also found that rRNA regions were integrated slightly more frequently than the protein-coding genes (Fig. 2B,C). A similar pattern was found in the NUMTs of mammals (Tsuji et al., 2012).

The number of NUMTs in eukaryotic genomes is considered to be related to the frequency of DSB repair and the rate of NUMT loss (Hazkani-Covo et al., 2010). Some large-scale studies have found a strong positive correlation between genome size and the number of NUMTs (Hazkani-Covo et al., 2010; Zhang et al., 2020b). NUMTs are likely found in transposon-rich regions and are positively correlated with genomic TE content in mammal genomes (Uvizl et al., 2024). We found a significant positive correlation between the number of NUMTs with the genome sizes and TE content in the ten species genomes (Fig. 2F). However, our study is based only on ten species and thus it remains to be seen if this pattern holds up when more genome data become available. In addition, six of the ten species in our study were from the same genus, and the genomes of *Schistocerca* generally have high TE content. This correlation conclusion may be confounded by phylogenetic relatedness. The correlation between TE content and the number of NUMTs may disappear when analyzing only among more closely related species. Overall, we need more genomic data from Orthoptera insects to strengthen the evidence of a positive correlation between transposon content and the number of NUMTs.

Duplication is thought to be the main process responsible for the increase in the number of NUMTs in different taxa (Bensasson et al., 2003; Collura and Stewart, 1995; Triant and DeWoody, 2007). Our study showed that the number of post-insertion duplication events of NUMTs varies across taxa and even among closely related species (Fig. 1). We posit that the post-insertion duplication of NUMTs may be associated with TEs. We found that the proportion of NUMTs flanked by TEs is higher in the duplicated NUMTs compared to other NUMTs (Fig. 2D). We argue that the post-insertion duplication of NUMTs is a random event; the more NUMTs inserted and the higher the TE content in the genome, the greater the probability of NUMT duplication events occurring.

4.4. Inferring the evolutionary history of mitochondria based on NUMTs

The difference between NUMTs and contemporary mtDNA can be used to infer the time when NUMTs were inserted into the nuclear genome. In the NUMTs divergence landscape (Fig. 4D), we found that the peaks of NUMT accumulation in Gryllidae and Tridactylidae were at positions with p-distances > 10, while the peaks of Acrididae were at positions with smaller p-distances. These results indicate that a large number of NUMT insertions in Acrididae species occurred recently and that many NUMTs insertions in Acrididae may occur after species divergence. If we sampled multiple species of Gryllidae or Tridactylida, we might find the same pattern as we have found in Acrididae.

NUMTs, as molecular fossils of mitochondria and “frozen” snapshots representing ancient mitochondria, can provide us with some special clues about the evolution of mitochondria. When comparing the GC content changes of all NUMTs in different periods, the GC differences in different regions of the mitochondria (such as the coding region and repetitive regions of mitochondria) may be overlooked. We calculated the difference in GC content of each NUMT and its corresponding mtDNA fragment at different stages and found that the difference in GC content increased with the insertion time. It must be considered that the NUMTs were affected by the nuclear genome evolutionary pattern, the GC content of the mitochondrial genome remains stable during evolution, and the inserted NUMTs follow the nucleotide substitution pattern of the nuclear genome. Once mtDNA is inserted into the nuclear genome, it becomes non-functional due to the differences in genetic code, and is therefore released from any selective pressure to maintain the function of mitochondria. There is no empirical evidence that NUMTs have any functional role, and therefore the NUMTs will likely accumulate mutations randomly. The nuclear genome has evolved differently from the mitochondrial genome, and the mitochondrial genome has a higher AT content (supplementary Fig. S2B and Table S9) (Wolstenholme, 1992), while the integration of NUMTs into the nuclear genome reduces the bias towards A and T (Bensasson et al., 2001b). The older the insertion of NUMT and the longer it has followed nuclear genome evolution, the higher its GC content.

We found that the length of protein-coding genes decreased during mitochondrial genome evolution, which was manifested in a gradual reduction of three bases each time in the sequence and did not cause frameshift mutations (Fig. 5A). These may indicate that mitochondrial proteins continue to streamline their amino acid sequences during evolution. We need to consider another possibility: after NUMT insertion into the genome, the “CTT” motif may undergo tandem duplication (Fig. 5A), similar to the accumulation of repeats in genomic microsatellites (short tandem repeats) due to replication slippage (Ellegren, 2004; Field and Wills, 1998). In addition, we found some synonymous mutations in NUMTs (Fig. 5A), which are unlikely to be mutations after NUMTs were inserted into the genome (NUMTs may follow a random base substitutions pattern after insertion), but should be synonymous substitutions during mitochondrial evolution.

4.5. Phylogenetic placement of NUMTs provides insights into divergence patterns

Based on the abundance of NUMTs found in all ten orthopteran species (Fig. 1), it is reasonable to assume that nuclear integration of mtDNA has been an ongoing event since the common ancestor of Orthoptera diverged from other polyneopteran lineages to the speciation of contemporary species. NUMTs are only recognizable if they resemble the sequences of contemporary mtDNA. Although NUMTs persist in the nuclear genome throughout lineage diversification, they will lose their sequence similarity to mtDNA due to random mutation, given sufficient evolutionary time. The nucleotide substitution rate of 13 mitochondrial protein-coding genes in Orthoptera is 1.355 % substitutions per million years (Chang et al., 2020), and even faster in non-coding regions. In this study, we found the largest divergence between NUMTs and mtDNA was 23.78 %. This means that the NUMTs we have found in the current genome are only the subset of all NUMTs that can be aligned with contemporary mtDNA. Thus, it is likely that NUMTs that were integrated in the common ancestor of Orthoptera are no longer recognizable as mtDNA-like. Our study provides some insights into how long NUMTs persist in the nuclear genome as a recognizable form.

Each of the ten species has hundreds to thousands of NUMTs in their nuclear genomes (Fig. 1), most of which resemble the contemporary mtDNA sequences (Fig. 6). These are likely autaponumts, or NUMTs that were integrated after each species divergence (Song et al., 2013). The sequence divergence of these NUMTs from the orthologous mtDNA range between close to 0 % to more than 15 %, indicating that NUMTs begin to accumulate as soon as the species diverges while mtDNA continues to evolve. In the phylogenetic tree constructed from the NUMTs and ortholgous mtDNAs, the autaponumts of each species formed a monophyletic clade together with its mtDNA. In these clades, mtDNA is almost always located at the end position far away from the base of the clade (Fig. 6). The NUMTs at the base of these clades were integrated at the early stage of speciation and has followed the mutation rate of the nuclear genome for a long time. In addition, we found that some NUMTs within the clade of autaponumts formed monophyletic clades among themselves (Fig. 6), suggesting the occurrence of multiple nuclear integration events or post-integration duplication.

Our study included six closely related species within the genus *Schistocerca* and we were able to recover not only species-specific autaponumts, but also synaponumts that are shared among different species (Fig. 6). It is estimated that the genus is about 8 million years old and *S. gregaria* was the earliest-diverging lineage within the genus that diverged around 6 million years ago (Song et al., 2017). The fact that we were able to recover synaponumts that were shared among all six *Schistocerca* species indicates that NUMTs can stay recognizable at least for 8 million years. The synaponumts inserted in the common ancestor disappeared in some subsequent taxa as species diverged, and the ancestral synaponumts were not retained in every species. For example, all COI synaponumts found in the genomes of *Schistocerca* species are not present in equal numbers in the current genomes of the six species (Fig. 6). These ancestral inserted COI synaponumts may have received different random mutations after species divergence, and some may even no longer be recognizable as NUMTs. Within *Schistocerca*, *S. gregaria* had more autaponumts in its genome than any other species. It is worth noting that environmental conditions are thought to be associated with the increased transfer of mitochondria to the nuclear genome, with changes in temperature being an inducing factor (Thorsness and Fox, 1990; Wang et al., 2012). Species may face different environmental stresses after divergence, leading to inconsistent rates of nuclear transfer events. We were not able to identify any synaponumts between any *Schistocerca* species and *L. migratoria*, whose most recent common ancestor lived more than 50 million years ago (Song et al., 2018). This suggests that any NUMTs that were integrated in the common ancestor of *Schistocerca* and *Locusta* must have mutated beyond recognition. Likewise, we could not identify any synaponumts between *G. bimaculatus* and *T. oceanicus*, which belong to the same subfamily, Gryllinae.

5. Conclusions

In this study, we found divergences between NUMTs and current mtDNA in Orthoptera ranging from 0 % to 23.78 %. The results showed that the number of NUMT insertions was significantly positively correlated with the genome size and the content of transposable elements in the genome. We found that 39.09 %-68.65 % of the NUMTs flanking regions (2,000 bp) contain retrotransposons and more NUMTs originated from mitochondrial rDNA regions. Our study of the full-length transcriptomes of three species found NUMTs consistent with mitochondrial transcripts, and these NUMTs accounted for 0.30 %-1.02 % of the total NUMTs. We found a potential mechanism of NUMT insertion: mitochondrial transcripts are reverse transcribed into double-stranded DNA and then integrated into the genome. It is speculated that this mechanism may be related to the high transposase abundance in Orthoptera. Furthermore, we consider that the double break repair of the nuclear genome caused by frequent TE insertion may provide more opportunities for NUMTs integration. Based on NUMTs as "frozen" snapshots of mitochondria, the phylogenetic tree constructed using NUMTs and contemporary mtDNA, we provide insights into ancient evolutionary events such as species-specific "autaponumts" and "synaponumts" shared among different species, as well as post-integration duplication events.

Data and Resource Availability

The *X. riparia* raw genomic sequencing data has been deposited at the public NCBI under the SRA database SRR27686114. The Pacbio *iso*-seq CCS data of *X. riparia* have been deposited at the public NCBI under SRA database (accession number PRJNA1067744) (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1067744). The genome assembly data were downloaded from NCBI (the GenBank assembly accession numbers in Table S1) and the other raw sequencing data used in the study were obtained from the SRA database (accession numbers in Table S11). Additionally, the mitochondrial genomes, NUMTs, NUMT flanking 2000 bp sequence, and TE consensus sequence have been deposited in the figshare database (https://doi.org/10.6084/m9.figshare.25093205.v1). The major codes used in this research are available at the GitHub repository (https://github.com/Liuxuanzeng/NUMTs_evolution).

Funding

This work was supported by the National Natural Science Foundation of China (grant number 31872217), the Fundamental Research Funds for the Central Universities (grant numbers GK202206021, GK202101003) as well as the U.S. National Science Foundation (grant number DBI-2021795 to H.S.) and the United State Department of Agriculture (Hatch Grant TEX0-2–6584 to H.S).

CRediT authorship contribution statement

Xuanzeng Liu: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis. **Nian Liu:** Visualization, Methodology. **Xuan Jing:** Visualization. **Hashim Khan:** Software. **Kaiyan Yang:** Resources. **Yanna Zheng:** Investigation. **Yimeng Nie:** Data curation. **Hojun Song:** Writing – review & editing, Resources, Methodology, Funding acquisition. **Yuan Huang:** Writing – review & editing, Resources, Project administration, Methodology, Funding acquisition.

Declaration of Competing Interest

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

Acknowledgments

This work was supported by Excellent Graduate Training Program of Shaanxi Normal University (grant numbers LHRCYB23006).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2024.108221.

References

Alfsnes, K., Leinaas, H.P., Hessen, D.O., 2017. Genome size in arthropods; different roles of phylogeny, habitat and life history in insects and crustaceans. Ecology and Evolution 7, 5939–5947.

Arctander, P., 1995. Comparison of a mitochondrial gene and a corresponding nuclear pseudogene. Proceedings of the Royal Society of London. Series b: Biological Sciences 262, 13–19.

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A., Saunders, N.C., 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annual Review of Ecology and Systematics 489–522.

Bensasson, D., Zhang, D.-X., Hewitt, G.M., 2000. Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. Molecular Biology and Evolution 17, 406–415.

Bensasson, D., Petrov, D.A., Zhang, D.-X., Hartl, D.L., Hewitt, G.M., 2001a. Genomic gigantism: DNA loss is slow in mountain grasshoppers. Molecular Biology and Evolution 18, 246–253.

Bensasson, D., Zhang, D.-X., Hartl, D.L., Hewitt, G.M., 2001b. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends in Ecology & Evolution 16, 314–321.

Bensasson, D., Feldman, M.W., Petrov, D.A., 2003. Rates of DNA duplication and mitochondrial DNA insertion in the human genome. Journal of Molecular Evolution 57, 343–354.

Bethoux, O., Nel, A., Lapeyrie, J., Gand, G., Galtier, J., 2002. Raphogla rubra gen. n., sp. n., the oldest representative of the clade of modern Ensifera (Orthoptera: Tettigoniidea, Gryllidea). European Journal of Entomology 99, 111–116.

Blanchard, J.L., Schmidt, G.W., 1996. Mitochondrial DNA migration events in yeast and humans: integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns. Molecular Biology and Evolution 13, 537–548.

Brown, W.M., George Jr, M., Wilson, A.C., 1979. Rapid evolution of animal mitochondrial DNA. Proceedings of the National Academy of Sciences 76, 1967–1971.

Brown, W.M., Prager, E.M., Wang, A., Wilson, A.C., 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. Journal of Molecular Evolution 18, 225–239.

Buhay, J.E., 2009. "COI-like" sequences are becoming problematic in molecular systematic and DNA barcoding studies. Journal of Crustacean Biology 29, 96–110.

Chang, H., Qiu, Z., Yuan, H., Wang, X., Li, X., Sun, H., Guo, X., Lu, Y., Feng, X., Majid, M., 2020. Evolutionary rates of and selective constraints on the mitochondrial genomes of Orthoptera insects with different wing types. Molecular Phylogenetics and Evolution 145, 106734.

Collura, R.V., Stewart, C.-B., 1995. Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids. Nature 378, 485–489.

Densmore, L.D., Wright, J.W., Brown, W.M., 1985. Length variation and heteroplasmy are frequent in mitochondrial DNA from parthenogenetic and bisexual lizards (genus Cnemidophorus). Genetics 110, 689–707.

Ding, L., Sang, H., Sun, C., 2021. Genus-wide characterization of nuclear mitochondrial DNAs in bumblebee (Hymenoptera: Apidae) genomes. Insects 12, 963.

Ellegren, H., 2004. Microsatellites: simple sequences with complex evolution. Nature Reviews Genetics 5, 435–445.

Field, D., Wills, C., 1998. Abundant microsatellite polymorphism in Saccharomyces cerevisiae, and the different distributions of microsatellites in eight prokaryotes and S. cerevisiae, result from strong mutation pressures and a variety of selective forces. Proceedings of the National Academy of Sciences 95, 1647–1652.

Gao, F., Chen, C., Arab, D.A., Du, Z., He, Y., Ho, S.Y., 2019. EasyCodeML: A visual tool for analysis of selection using CodeML. Ecology and Evolution 9, 3891–3898.

Gellissen, G., Bradfield, J., White, B., Wyatt, G., 1983. Mitochondrial DNA sequences in the nuclear genome of a locust. Nature 301, 631–634.

Gyllensten, U., Wharton, D., Josefsson, A., Wilson, A.C., 1991. Paternal inheritance of mitochondrial DNA in mice. Nature 352, 255–257.

Haag-Liautard, C., Coffey, N., Houle, D., Lynch, M., Charlesworth, B., Keightley, P.D., 2008. Direct estimation of the mitochondrial DNA mutation rate in Drosophila melanogaster. PLoS Biology 6, e204.

Hanrahan, S.J., Johnston, J.S., 2011. New genome size estimates of 134 species of arthropods. Chromosome Research 19, 809–823.

Hawlitschek, O., Sadílek, D., Dey, L.-S., Buchholz, K., Noori, S., Baez, I.L., Wehrt, T., Brozio, J., Trávníček, P., Seidel, M., 2023. New estimates of genome size in Orthoptera and their evolutionary implications. Plos One 18, e0275551.

Hazkani-Covo, E., 2009. Mitochondrial insertions into primate nuclear genomes suggest the use of numts as a tool for phylogeny. Molecular Biology and Evolution 26, 2175–2179.

Hazkani-Covo, E., 2022. A burst of numt insertion in the Dasyuridae family during marsupial evolution. Frontiers in Ecology and Evolution 10, 844443.

Hazkani-Covo, E., Covo, S., 2008. Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. PLoS Genetics 4, e1000237.

Hazkani-Covo, E., Zeller, R.M., Martin, W., 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genetics 6, e1000834.

Hebert, P.D., Cywinska, A., Ball, S.L., DeWaard, J.R., 2003. Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series b: Biological Sciences 270, 313–321.

Hebert, P.D., Bock, D.G., Prosser, S.W., 2023. Interrogating 1000 insect genomes for NUMTs: A risk assessment for estimates of species richness. Plos One 18, e0286620.

Hlaing, T., Tun-Lin, W., Somboon, P., Socheat, D., Setha, T., Min, S., Chang, M.S., Walton, C., 2009. Mitochondrial pseudogenes in the nuclear genome of Aedes aegypti mosquitoes: implications for past and future population genetic studies. Bmc Genetics 10, 1–12.

Hoeh, W.R., Blakley, K.H., Brown, W.M., 1991. Heteroplasmy suggests limited biparental inheritance of Mytilus mitochondrial DNA. Science 251, 1488–1490.

Jing, X., Liu, X.Z., Yuan, H., Dai, Y., Zheng, Y.N., Zhao, L.N., Ma, L.B., Huang, Y., 2024. Evolutionary dynamics of genome size and transposable elements in crickets (Ensifera: Gryllidea). Systematic Entomology.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K., Von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14, 587–589.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30, 772–780.

Kleine, T., Maier, U.G., Leister, D., 2009. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. Annual Review of Plant Biology 60, 115–138.

Kondo, R., Satta, Y., Matsuura, E., Ishiwa, H., Takahata, N., Chigusa, S., 1990. Incomplete maternal transmission of mitochondrial DNA in Drosophila. Genetics 126, 657–663.

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution 35, 1547.

Letunic, I., Bork, P., 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research 49, W293–W296.

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100.

Li, H., 2021. New strategies to improve minimap2 alignment accuracy. Bioinformatics 37, 4572–4574.

Liang, B., Wang, N., Li, N., Kimball, R.T., Braun, E.L., 2018. Comparative genomics reveals a burst of homoplasy-free numt insertions. Molecular Biology and Evolution 35, 2060–2064.

Lin, M.T., Simon, D.K., Ahn, C.H., Kim, L.M., Beal, M.F., 2002. High aggregate burden of somatic mtDNA point mutations in aging and Alzheimer's disease brain. Human Molecular Genetics 11, 133–145.

Liu, X., Majid, M., Yuan, H., Chang, H., Zhao, L., Nie, Y., He, L., Liu, X., He, X., Huang, Y., 2022. Transposable element expansion and low-level piRNA silencing in grasshoppers may cause genome gigantism. BMC Biology 20, 1–16.

Liu, X., Zhao, L., Majid, M., Huang, Y., 2024. Orthoptera-TElib: a library of Orthoptera transposable elements for TE annotation. Mobile DNA 15, 5.

Lopez, J.V., Culver, M., Stephens, J.C., Johnson, W.E., O`Brien, S.J., 1997. Rates of nuclear and cytoplasmic mitochondrial DNA sequence divergence in mammals. Molecular Biology and Evolution 14, 277–286.

Majid, M., Yuan, H., 2021. Comparative Analysis of Transposable Elements in Genus Calliptamus Grasshoppers Revealed That Satellite DNA Contributes to Genome Size Variation. Insects 12, 837.

Meng, G., Li, Y., Yang, C., Liu, S., 2019. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic Acids Research 47, e63–e.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution 37, 1530–1534.

Mishmar, D., Ruiz-Pesini, E., Brandon, M., Wallace, D.C., 2004. Mitochondrial DNA-like sequences in the nucleus (NUMTs): Insights into our African origins and the mechanism of foreign DNA integration. Human Mutation 23, 125–133.

Moulton, M.J., Song, H., Whiting, M.F., 2010. Assessing the effects of primer specificity on eliminating numt coamplification in DNA barcoding: a case study from Orthoptera (Arthropoda: Insecta). Molecular Ecology Resources 10, 615–627.

Nguyen, L.-T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution 32, 268–274.

Perna, N.T., Kocher, T.D., 1996. Mitochondrial DNA: molecular fossils in the nucleus. Current Biology 6, 128–129.

Piégu, B., Bire, S., Arensburger, P., Bigot, Y., 2015. A survey of transposable element classification systems–a call for a fundamental update to meet the challenge of their diversity and complexity. Molecular Phylogenetics and Evolution 86, 90–109.

Podnar, M., Haring, E., Pinsker, W., Mayer, W., 2007. Unusual origin of a nuclear pseudogene in the Italian wall lizard: intergenomic and interspecific transfer of a large section of the mitochondrial genome in the genus Podarcis (Lacertidae). Journal of Molecular Evolution 64, 308–320.

Ricchetti, M., Fairhead, C., Dujon, B., 1999. Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. Nature 402, 96–100.

Ricchetti, M., Tekaia, F., Dujon, B., 2004. Continued colonization of the human genome by mitochondrial DNA. PLoS Biology 2, e273.

Richly, E., Leister, D., 2004. NUMTs in sequenced eukaryotic genomes. Molecular Biology and Evolution 21, 1081–1084.

Schiavo, G., Hoffmann, O.I., Ribani, A., Utzeri, V.J., Ghionda, M.C., Bertolini, F., Geraci, C., Bovo, S., Fontanesi, L., 2017. A genomic landscape of mitochondrial DNA insertions in the pig nuclear genome provides evolutionary signatures of interspecies admixture. DNA Research 24, 487–498.

Schiavo, G., Strillacci, M., Ribani, A., Bovo, S., Roman-Ponce, S., Cerolini, S., Bertolini, F., Bagnato, A., Fontanesi, L., 2018. Few mitochondrial DNA sequences are inserted into the Turkey (Meleagris gallopavo) nuclear genome: evolutionary analyses and informativity in the domestic lineage. Animal Genetics 49, 259–264.

Schimper, W.P., von Zittel, K.A., 1885. Handbuch der Palaeontologie. R. Oldenbourg.

Schuster, W., Brennicke, A., 1987. Plastid, nuclear and reverse transcriptase sequences in the mitochondrial genome of Oenothera: is genetic information transferred between organelles via RNA? The EMBO Journal 6, 2857–2863.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H., Flook, P., 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. Annals of the Entomological Society of America 87, 651–701.

Song, H., Buhay, J.E., Whiting, M.F., Crandall, K.A., 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proceedings of the National Academy of Sciences 105, 13486–13491.

Song, H., Moulton, M.J., Hiatt, K.D., Whiting, M.F., 2013. Uncovering historical signature of mitochondrial DNA hidden in the nuclear genome: the biogeography of S chistocerca revisited. Cladistics 29, 643–662.

Song, H., Amédégnato, C., Cigliano, M.M., Desutter-Grandcolas, L., Heads, S.W., Huang, Y., Otte, D., Whiting, M.F., 2015. 300 million years of diversification: elucidating the patterns of orthopteran evolution based on comprehensive taxon and gene sampling. Cladistics 31, 621–651.

Song, H., Foquet, B., Mariño-Pérez, R., Woller, D.A., 2017. Phylogeny of locusts and grasshoppers reveals complex evolution of density-dependent phenotypic plasticity. Scientific Reports 7, 6606.

Song, H., Mariño-Pérez, R., Woller, D.A., Cigliano, M.M., 2018. Evolution, diversification, and biogeography of grasshoppers (Orthoptera: Acrididae). Insect Systematics and Diversity 2, 3.

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

Sunnucks, P., Hales, D.F., 1996. Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus Sitobion (Hemiptera: Aphididae). Molecular Biology and Evolution 13, 510–524.

Thalmann, O., Hebler, J., Poinar, H.N., Pääbo, S., Vigilant, L., 2004. Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes. Molecular Ecology 13, 321–335.

Thalmann, O., Serre, D., Hofreiter, M., Lukas, D., Eriksson, J., Vigilant, L., 2005. Nuclear insertions help and hinder inference of the evolutionary history of gorilla mtDNA. Molecular Ecology 14, 179–188.

Thompson, J.D., Gibson, T.J., Higgins, D.G., 2003. Multiple sequence alignment using ClustalW and ClustalX. Current protocols in bioinformatics, 2.3. 1-2.3. 22.

Thorsness, P.E., Fox, T.D., 1990. Escape of DNA from mitochondria to the nucleus in Saccharomyces cerevisiae. Nature 346, 376–379.

Thorsness, P.E., Weber, E.R., 1996. Escape and migration of nucleic acids between chloroplasts, mitochondria, and the nucleus. International Review of Cytology 165, 207–234.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., Martin, W., 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nature Reviews Genetics 5, 123–135.

Triant, D.A., DeWoody, J.A., 2007. Extensive mitochondrial DNA transfer in a rapidly evolving rodent has been mediated by independent insertion events and by duplications. Gene 401, 61–70.

Triant, D.A., Hayes, L.D., 2011. Molecular approaches in behavioural research: a cautionary note regarding mitochondrial transfers to the nucleus (numts). Animal Behaviour 3, 601–606.

Tsuji, J., Frith, M.C., Tomii, K., Horton, P., 2012. Mammalian NUMT insertion is non-random. Nucleic Acids Research 40, 9073–9088.

Uvizl, M., Puechmaille, S.J., Power, S., Pippel, M., Carthy, S., Haerty, W., Myers, E.W., Teeling, E.C., Huang, Z., 2024. Comparative genome microsynteny illuminates the fast evolution of nuclear mitochondrial segments (NUMTs) in mammals. Molecular Biology and Evolution 41, msad278.

Vrijenhoek, R., 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol Mar Biol Biotechnol 3, 294–299.

Wang, J.-X., Liu, J., Miao, Y.-H., Huang, D.-W., Xiao, J.-H., 2020. Tracking the distribution and burst of nuclear mitochondrial DNA sequences (NUMTs) in Fig Wasp Genomes. Insects 11, 680.

Wang, D., Lloyd, A.H., Timmis, J.N., 2012. Environmental stress increases the entry of cytoplasmic organellar DNA into the nucleus in plants. Proceedings of the National Academy of Sciences 109, 2444–2448.

Wang, D., Timmis, J.N., 2013. Cytoplasmic organelle DNA preferentially inserts into open chromatin. Genome Biology and Evolution 5, 1060–1064.

Wei, W., Schon, K.R., Elgar, G., Orioli, A., Tanguy, M., Giess, A., Tischkowitz, M., Caulfield, M.J., Chinnery, P.F., 2022. Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. Nature 611, 105–114.

Wolstenholme, D.R., 1992. Animal mitochondrial DNA: structure and evolution. International Review of Cytology 141, 173–216.

Xiang, C.Y., Gao, F., Jakovlić, I., Lei, H.P., Hu, Y., Zhang, H., Zou, H., Wang, G.T., Zhang, D., 2023. Using PhyloSuite for Molecular Phylogeny and Tree-Based Analyses. iMeta 2, e87.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24, 1586–1591.

Yuan, H., Huang, Y., Mao, Y., Zhang, N., Nie, Y., Zhang, X., Zhou, Y., Mao, S., 2021. The evolutionary patterns of genome size in Ensifera (Insecta: Orthoptera). Frontiers in Genetics 12, 693541.

Zhang, G.-J., Dong, R., Lan, L.-N., Li, S.-F., Gao, W.-J., Niu, H.-X., 2020b. Nuclear integrants of organellar DNA contribute to genome structure and evolution in plants. International Journal of Molecular Sciences 21, 707.

Zhang, D., Gao, F., Jakovlić, I., Zou, H., Zhang, J., Li, W.X., Wang, G.T., 2020a. PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Molecular Ecology Resources 20, 348–355.

Zhang, G., Geng, D., Guo, Q., Liu, W., Li, S., Gao, W., Wang, Y., Zhang, M., Wang, Y., Bu, Y., 2022. Genomic landscape of mitochondrial DNA insertions in 23 bat genomes: characteristics, loci, phylogeny, and polymorphism. Integrative Zoology 17, 890–903.

Zhang, D.-X., Hewitt, G.M., 1996a. Nuclear integrations: challenges for mitochondrial DNA markers. Trends in Ecology & Evolution 11, 247–251.

Zhang, D.X., Hewitt, G.M., 1996b. Highly conserved nuclear copies of the mitochondrial control region in the desert locust Schistocerca gregaria: some implications for population studies. Molecular Ecology 5, 295–300.

Zischler, H., Geisert, H., von Haeseler, A., Pääbo, S., 1995. A nuclear’fossil’of the mitochondrial D-loop and the origin of modern humans. Nature 378, 489–492.