Genomic Underground: Unraveling NUMTs in Mole Voles

Dmitry Prokopov[1*], Tigran Saluev[2], Svetlana Romanenko[3], Irina Bakloushinskaya[4], Alexander Graphodatsky[3]

[1] Independent researcher, United Kingdom, London

[2] Independent researcher, Georgia, Tbilisi

[3] Institute of Molecular and Cellular Biology, Russian Academy of Sciences, Siberian Branch, Novosibirsk, 630090, Russia

[4] Koltzov Institute of Developmental Biology of Russian Academy of Sciences, 26 Vavilov Street, Moscow, 119334, Russia

[*] dimprokopov@gmail.com

Abstract

Nuclear mitochondrial DNA segments (NUMTs) are pervasive elements of eukaryotic genomes. This study focuses on *Ellobius talpinus* and *Ellobius lutescens*, for which we assembled full mitochondrial DNA sequences. Our study identified NUMTs encompassing approximately 0.0052% and 0.0086% of genome assembly length in *E. talpinus* and *E. lutescens*, respectively. These NUMTs collectively spanned a total length of 122,294 bp in *E. talpinus* and 194,875 bp in *E. lutescens*. Notably, the majority of NUMTs in both species were short, with lengths of less than 500 bp. In *E. talpinus*, the data indicated the presence of comparatively recent NUMT insertions. More than half of the NUMTs in each species are organized into clusters, primarily situated in intergenic regions or within introns. RNA genes are the most frequently occurring fragments within these NUMTs. Furthermore, our analysis identified LINE, SINE, and LTR retrotransposons within and flanking NUMT clusters. Our results demonstrate the intricate dynamics of NUMT integration and distribution in *Ellobius* species and provide insights into their genomic architecture and evolutionary history. This study contributes to the broader understanding of mitochondrial DNA contributions to nuclear genomes and underscores the complexity of distinguishing between mtDNA and nuclear DNA in genomic studies.

Introduction

Mole voles (*Ellobius*, Arvicolinae, Cricetidae) are highly adapted to underground life social subterranean rodents that inhabit south-eastern Europe south of Western Siberia, Central Asia, and Mongolia[1–3]. Their genomic landscape, marked by unique features such as a distinct sex chromosome system[4–8] and highly rearranged genomes with numerous inversions and translocations[9–11], alludes to a rapid phase of genomic evolution within these subterranean rodents.

To date, the complete mitochondrial genomes of four *Ellobius* species have been sequenced: *Ellobius lutescens*, *Ellobius fuscocapillus*, *Ellobius talpinus*[12], and *Ellobius tancrei*[13], providing a foundation for understanding their mtDNA structure and evolutionary dynamics. Although their mitochondrial DNAs (mtDNAs) are typical of the Arvicolinae subfamily, they exhibit a unique feature of high GC content[14]. Recent studies[15,16] have highlighted adaptive evolution in the mitochondrial genes of *Ellobius* species, particularly ATP8 and CYTB. These findings show varied selection pressures and link mtDNA evolution to rodent subterranean adaptations.

A notable aspect of genomic research is the study of nuclear mitochondrial DNA segments (NUMTs), which are fragments of mitochondrial DNA integrated into the nuclear genome. Discovered after sequencing the first human mitochondrial genome in 1983[17], NUMTs were believed to traverse the nucleus through various mechanisms, such as homologous recombination[18], non-homologous end joining[19], microhomology-mediated end joining[20], and retrotransposition[21]. Although NUMTs are prevalent across eukaryotic genomes, the largest number of NUMTs has been described in mammals[20,22–29]. These repeated sequences can be utilized for various research purposes, such as population genetics[30], past demographic event detection[31], inference of hybridization events[32], and biogeographic analysis[33].

The presence of NUMTs poses difficulties in studying mitochondrial DNA (mtDNA), which may lead to the amplification of nuclear mtDNA copies. A recent study identified 195 NUMTs in the *E. talpinus* nuclear genome[34]. The high number of NUMTs highlights the challenges in differentiating mtDNA sequences and warns about the amplification of cryptic NUMTs in future studies on *E. talpinus*.

In this study, we conducted a thorough analysis of publicly available genomic data for *E. lutescens* and *E. talpinus*, using cutting-edge bioinformatics techniques. These techniques include mitochondrial DNA assembly, annotation, and NUMT detection. Additionally, we performed a comprehensive genome annotation to investigate the distribution, frequency, and structural patterns of NUMTs in these species.

Materials and Methods

Mitochondrial DNA assembly and annotation

Genomic reads were downloaded from the NCBI SRA database under accession numbers listed in Table 1.

Table 1. List of reads used in this study.

Sample name	Species	SRA Accession	Tissue	Size (in Gb)	Study
ELUT 1	*Ellobius lutescens*	SRR10542531	Muscle	4.4	[12]
ELUT 2		SRR3497433	Liver	35.7	[7]
ELUT 3		SRR3497460	Liver	20.1	[7]
ETAL 1	*Ellobius talpinus*	SRR10542532	Muscle	3.5	[12]
ETAL 2		SRR3497471	Liver	37.3	[7]

To ensure quality control, low-quality sequences, adapter sequences and PCR duplicates were removed using fastp 0.23.4[35] with the following parameters: --detect_adapter_for_pe -D --dup_calc_accuracy 6 -5 -3 -r -l 20. High-quality trimmed deduplicated reads were used for the mitogenome assembly. We used the assembler GetOrganelle v1.7.7.0[36] with the option -F animal_mt to assemble mitogenomes. All available mitochondrial sequences of *Ellobius* from GenBank were used as the seed sequences. *de novo* assembly was manually checked and corrected. Protein-coding genes, non-coding regions, and RNA genes were annotated using MITOS2[37]. To detect mtDNA variants, the reads used for assembly were aligned against the newly assembled mitochondria using the BWA-MEM algorithm 0.7.17-r1188[38]. Variant calling was performed using the Ivar 1.4.2[39].

DNA Samples

Tissue samples of animals from a laboratory colony maintained in IDB RAS were used: four samples of *E. talpinus*, and four samples were used of *E. lutescens*. Tissue samples were deposited in the Large-Scale Research Facility “Collection of Wildlife tissues for genetic research” of the Core Centrum of the Koltzov Institute of Developmental Biology Russian Academy of Sciences (CWT IDB RAS), state registration number 3579666. The treatment of animals adhered to conventional international protocols in accordance with the guidelines for humane endpoints for animals used in biomedical research. The Ethics Committee for Animal Research of the Koltzov Institute of Developmental Biology RAS, which has approved all experimental protocols in compliance with the Regulations for Laboratory Practice in the Russian Federation (most recent protocol being № 37-25.06.2020), ensured that every possible measure was taken to minimize the animals' suffering during capturing and sampling. Information on the origin founders of the colonies and sex of samples is presented in Table 2.

Table 2. Samples description

№	Species	sex	Tissue	Location
1	*Ellobius lutescens*	female	kidney	Armenia, near Vokhchabert settlement
2	*Ellobius lutescens*	female	kidney	Armenia, near Vokhchabert settlement
3	*Ellobius lutescens*	male	kidney	Armenia, near Atis settlement
4	*Ellobius lutescens*	male	kidney	Armenia, near Atis settlement
5	*Ellobius talpinus*	female	liver	Russia, Orenburg Oblast, near Alekseevka settlement
6	*Ellobius talpinus*	female	kidney	Russia, Saratov Oblast, near Khvalynsk
7	*Ellobius talpinus*	male	liver	Russia, Orenburg Oblast, near Alekseevka settlement
8	*Ellobius talpinus*	male	kidney	Russia, Saratov Oblast, near Khvalynsk

DNA isolation was performed from tissues stored in ethanol using a Thermo Scientific™ GeneJET Genomic DNA Purification Kit (Thermo Fisher Scientific, USA) according to the manufacturer's protocol.

PCR

Unique specific primers were selected using Primer3 2.6.1 program[40]. The list of primer pairs is presented in Supplementary Table 1.

The PCR mixture contained 1× Taq buffer, 0.3 U Taq polymerase, 2 mM $MgCl_2$, 200 μM dNTPs, 1 μM of each primer, and 20 ng of DNA.

To amplify the major non-coding region (NCR), located between mt-tRNA genes of phenylalanine and proline of *E. lutescens*, a protocol enabling the amplification of long DNA sequences (long PCR) was employed. Initial denaturation was performed at 94 °C for 3 min. Subsequently, the following cycling protocol was applied (35 cycles): 94 °C for 25 s, 68 °C for 10 s, and 68 °C for 2 min 30 s. The final elongation step was at 72 °C for 2 min and 30 s.

For the remaining amplification experiments, the program was as follows: initial denaturation at 94 °C for 3 min, followed by a cycling protocol (35 cycles) at 94 °C for 25 s, 68 °C for 10 s, and 68 °C for 50 s. The final elongation step was performed at 72 °C for 40 s.

Amplicons were visualized by electrophoresis on a 1.5% agarose gel using molecular markers S-8100 and S-8250 (BioLabMix).

NUMTs detection in genome assembly

A similarity search was conducted on the nuclear and newly assembled mitochondrial genomes of two *Ellobius* species: male *E. lutescens* (ASM168507v1) and female *E. talpinus* (Etalpinus_0.1). Two mitochondrial genomes from each species were concatenated head-to-tail to avoid edge effects.

The search was performed using BLASTn version 2.15, with an e-value ≤ 5×10^{-3}. NUMT sequences were concatenated into clusters if they followed each other on the scaffold.

Genome annotation

The genomes of the mole voles from the NCBI Assembly database were not annotated; hence, we performed genomic annotation. For the identification and classification of repetitive sequences, we utilized RepeatModeler2 version 2.0.4, with the option -LTRStruct[41] for *de novo* and *ab initio* repeat identification. Similar sequences of repeated elements were clustered using CD-HIT 4.8.1[42]. We used RepeatMasker version 4.1.5[43] with the parameters -e rmblast -s -no_is to identify repetitive sequences in the whole-genome sequences. The coordinates of the repeats detected were subsequently intersected with the spacer sequences present in the NUMT clusters, as well as the flanking regions of these clusters, with lengths of 1,000 base pairs (bp).

Protein-coding gene structures were predicted *ab initio* using Augustus 3.4.0[44] with the options --strand=both --noInFrameStop=false --protein=on --codingseq=on --introns=on --start=on --stop=on --cds=on --singlestrand=false --UTR=off –genemodel=complete.

Manipulation with output tables and genomic intervals was performed using bedtools 2.31.1[45] and custom scripts available at https://github.com/Saluev/ellobius.

Results

Assembled mtDNA

The GetOrganelle pipeline successfully generated completely circularized mitochondrial DNA sequences for all available NCBI SRA sample reads of *E. lutescens* and *E. talpinus*.

The assembled mitochondrial genomes were similar to those previously reported by Bondareva et al.[12], although there were significant structural differences in *E. lutescens*. The newly assembled *E. lutescens* mitogenomes were considerably longer than those assembled in the study by Bondareva et al.[12], with nucleotide differences ranging from 1,516 to 1,639 bp. The presence of light-strand replication origin (O_L) and duplications in the control region, consisting of 2-3 duplicated origin for H-strand DNA replication (O_H) sequences that had not been previously assembled, accounts for the difference in our assembly compared with previous studies.

To assess the copy number and size of the control region for both species, we performed PCR on the flanking regions and the internal segment of the O_H. The observed amplicon length for *E. talpinus* perfectly matched the expected size determined by gel electrophoresis. However, for *E. lutescens,* the O_H amplicon was 2-3 times smaller than expected, and the copy number of this region was reduced by 2-3 times, suggesting assembly artifacts. Molecular validation of the O_H length allowed for manual trimming of the sequence, retaining only a single copy of the O_H sequence.

The assembly process was conducted for both ELUT3 and ELUT2 samples, originating from the same parental individuals[7]. This analysis revealed that ELUT2 and ELUT3 had identical maternal

mitochondrial DNA, reflecting their shared maternal lineages. The characteristics of the assembled mitogenomes are listed in Table 3.

Table 3. Characteristics of the mitochondrial DNA of E. lutescens (ELUT) and E. talpinus (ETAL) from the NCBI database (coloured in gray) and this article.

Name	Accession number	mtDNA length	A%	T%	G%	C%	GC-content	Control region	
								size	GC-content
ELUT NCBI	MT483992.1	16540	31.1	25.9	14.3	28.4	42.8	1172	41.5
ELUT 1		16843	31.2	25.8	14.4	28.6	43	1169	41.4
ELUT 2 and ELUT3		16853	31.1	25.8	14.5	29	43.1	1169	41.5
ETAL NCBI	NC_054160.1	16367	32.4	26.3	13.6	27.7	41.3	976	39.8
ETAL 1		16367	32.4	26.3	13.6	27.7	41.3	976	39.8
ETAL 2		16361	32.8	26.5	13.3	27.4	40.7	969	38.9

NUMTs identification

Bioinformatics analysis of the mitochondrial genome in the nuclear genomes of *E. lutescens* and *E. talpinus* revealed the presence of 167 and 463 fragments, respectively. The total length of NUMTs in the *E. talpinus* and *E. lutescens* genomes was 122 294 and 194 875 bp, respectively, which accounted for 0.0052% and 0.0086% of the genome assembly length, or 0.0041% and 0.0068% of the nuclear genome, respectively, based on previous genome amount estimations[46]. The majority of NUMTs exhibited a length of less than 500 bp (mean: 732 bp, median: 471 bp in *E. lutescens* and mean: 421 bp, median: 244 bp in *E. talpinus*), as depicted in Supplementary Figure 1. The degree of sequence identity between mtDNA and NUMT exhibits considerable variation, encompassing values spanning from 71.5% to 100%. The most frequently observed identity was registered at 80-84% (mean: 87%, median: 85%) in *Ellobius lutescens*, while 90-94% (mean: 90%, median: 90%) was recorded in *Ellobius talpinus* (Supplementary Figure 2).

The most frequently occurring fragments of the mtDNA sequence in the nuclear genome were RNA-encoding sequences (263 and 447 genes in NUMTs for *E. talpinus* and *E. lutescens*, respectively) and protein-coding genes (153 and 372). The lowest number of mtDNA fragments

found in the nuclear genome was found in non-coding regions (Figure 1, Supplementary Figure 3). The most frequent genes in NUMTs were rrnL (40 fragments in *E. lutescens* and 80 in *E. talpinus*) and control regions (33 and 80 fragments, respectively).

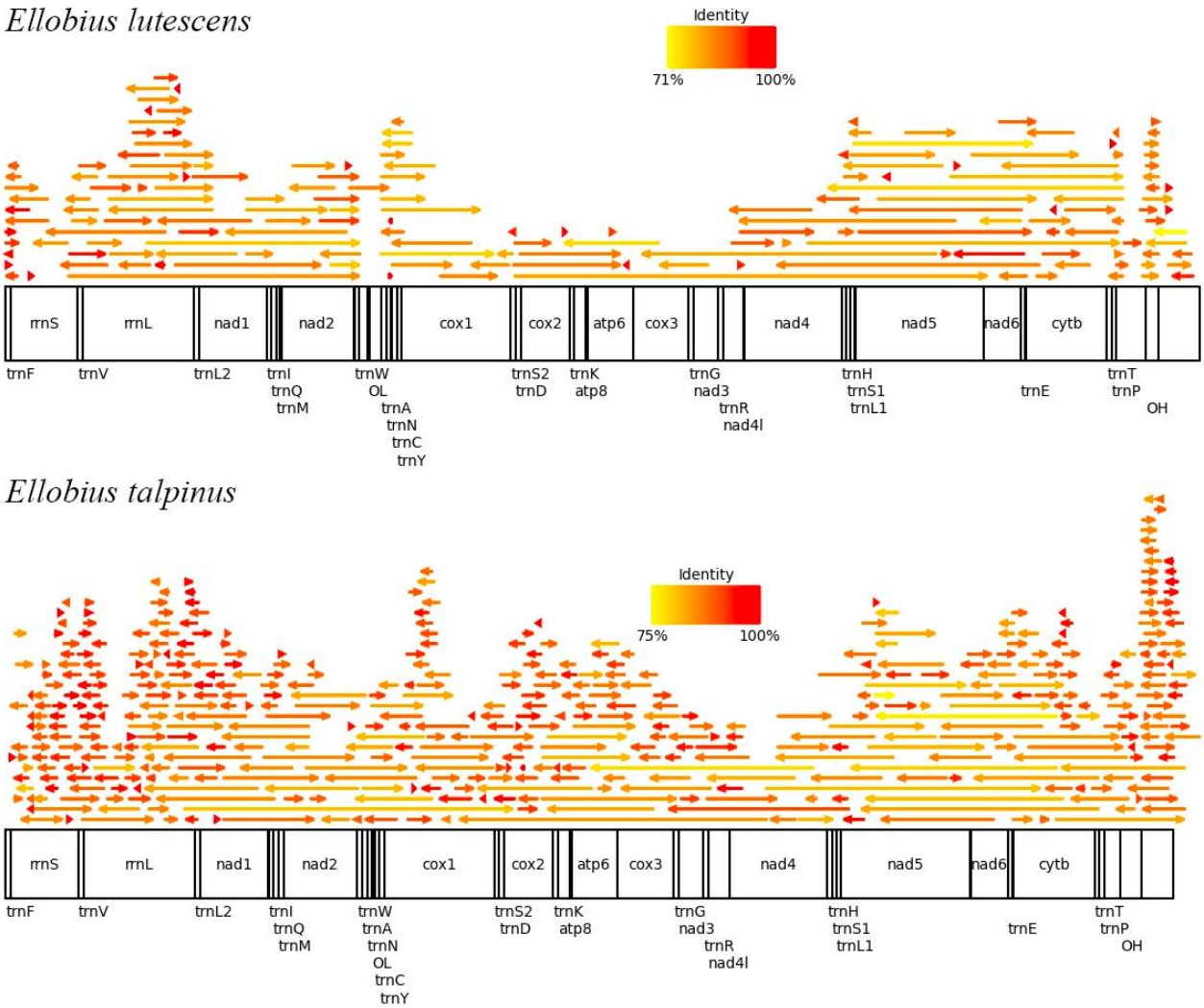


Figure 1. Visual representation of NUMT alignments on the mitochondrial DNA map of Ellobius lutescens and Ellobius talpinus. The arrangement of the sequences is demonstrated using arrows, with the direction of the sequence indicated. The colour scheme, which range from yellow to red, conveys the degree of identity.

51.5% and 45.6% of NUMTs in *E. lutescens* and *E. talpinus* organized in clusters (86 and 211 clusters in the genome, respectively), containing homologous mtDNA sequences interrupted by spacer sequences.

Genomic context of NUMTs

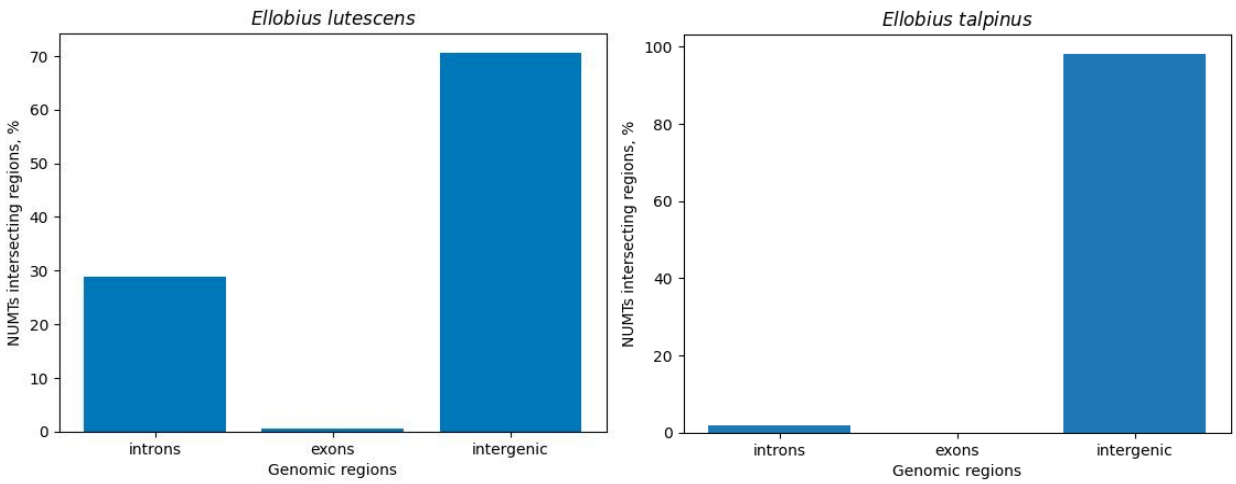


Figure 2: Location of NUMTs in the genomic regions of two mole vole species.

Analysis of NUMTs in relation to genomic annotations revealed that the majority of NUMTs were located in intergenic regions of the genome, with a notable proportion of NUMTs detected in gene regions (Figure 2). Within these gene regions, introns are the most common location for NUMTs, followed by exons. In the genome of *E. lutescens*, only one NUMT was identified within the coding region of a gene, specifically the second exon of the PDSS2 truncated pseudogene.

We investigated the occurrence of repetitive sequences both within and around NUMT clusters. (Figure 3).

In *E. lutescens*, 166 repetitive sequences were found within the NUMT clusters, with LTRs (22.3%), SINEs (43.4%, mainly SINE/Alu), and LINEs (23.5%, all LINE/L1). The flanking regions contained 277 sequences, showing a higher prevalence of LTRs (34.7%, mainly LTR/ERVK) and

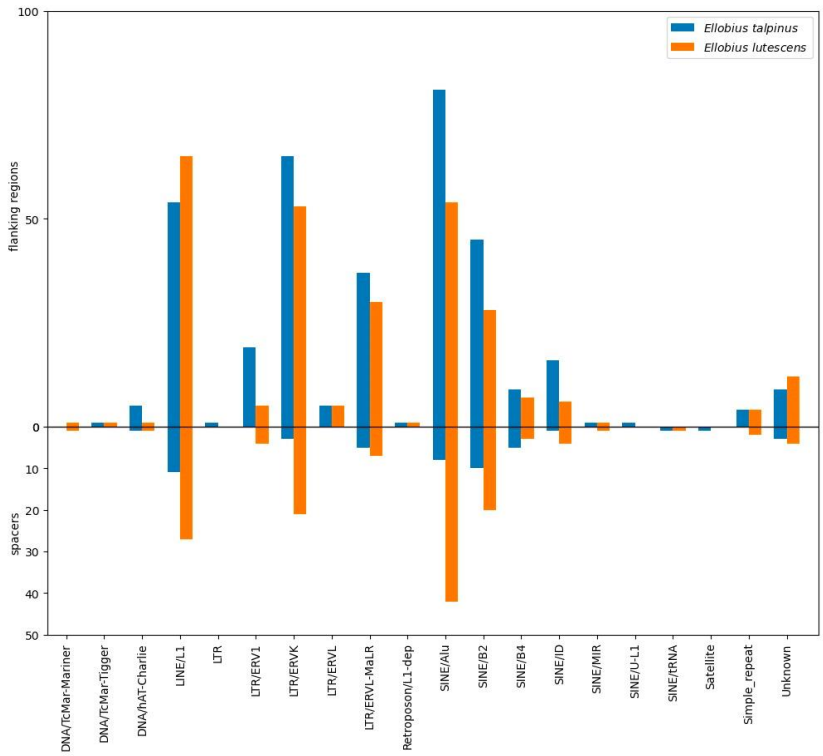


Figure 3 : Content of repeated elements in the internal and flanking regions of NUMT clusters.

SINEs (33.6%).

In *E. talpinus*, NUMT clusters contained 82 repeats, with a composition of LTRs (29.3%, mostly LTR/ERVL-MaLR and LTR/ERVK), SINEs (54.9%, predominantly SINE/Alu and SINE/B2), and LINEs (17.1%, all LINE/L1). The flanking areas had 322 sequences, with a larger proportion of LTRs (42.5%, mainly LTR/ERVK) and SINEs (43.8%).

Nucleotide variants and coverage analysis

Several mitochondrial single-nucleotide variants (SNV) were identified in *E. lutescens* samples, of which five possessed a considerable degree of coverage (Figure 4).

An unusual observation was made regarding the *E. talpinus* samples. Specifically, in one sample (ETAL 2), a large number of nucleotide variants were detected with high coverage, which was similar to the distribution pattern of the NUMT clusters (Figure 1). Conversely, for ETAL 1, only a small number of nucleotide variants were observed, and these were mostly of low coverage (Figure 4).

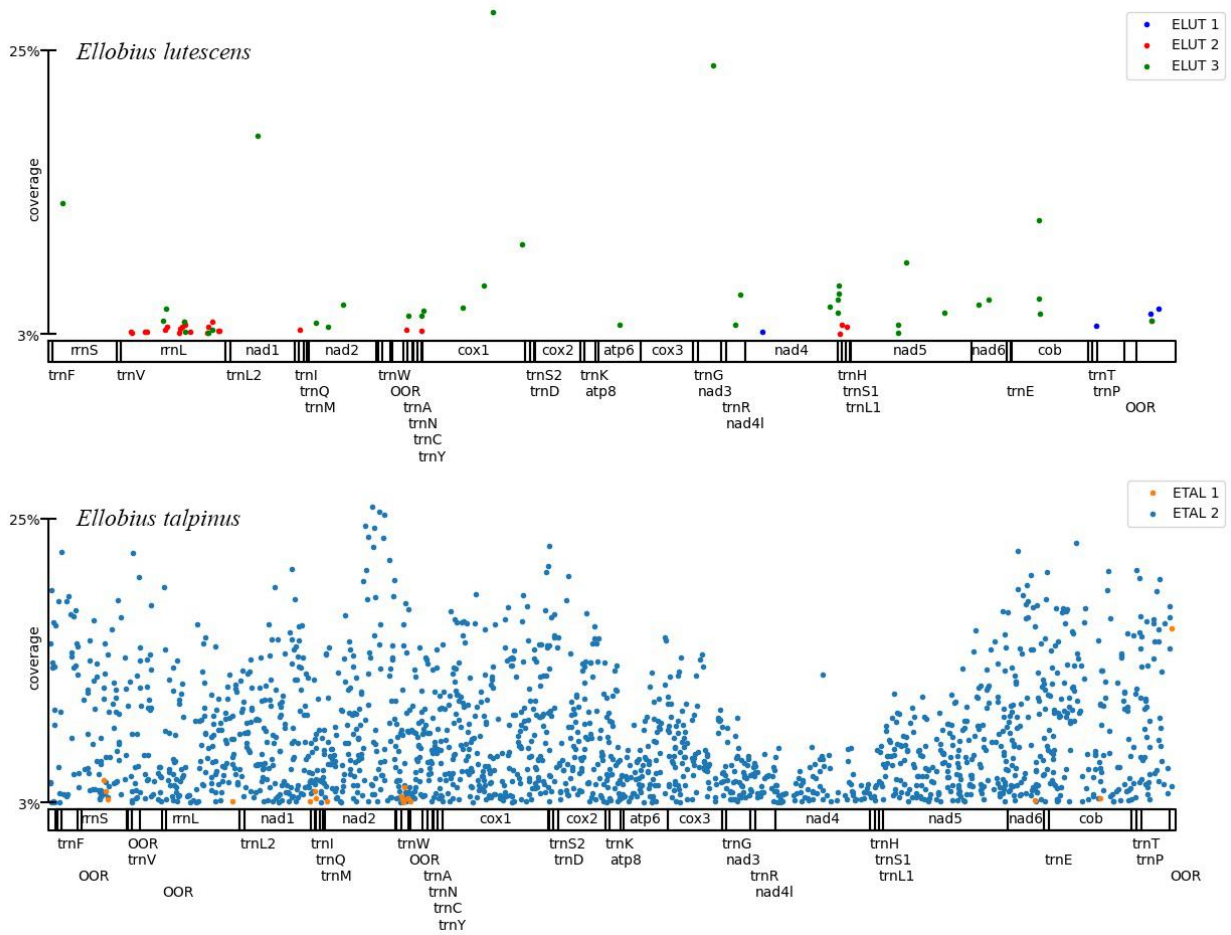


Figure 4: SNV composition in mitochondrial genomes of Ellobius lutescens and Ellobius talpinus.

Discussion

Our study provides an in-depth look at the distinctive patterns of mitochondrial DNA in the nuclear genomes of *E. lutescens* and *E. talpinus*. A detailed exploration of NUMTs uncovers genomic complexities, offering insights into the dynamic interaction between mitochondrial and nuclear genomes. Furthermore, our analysis extends beyond mere identification, as we uncovered the composition, organization, and genomic context of NUMTs, offering novel insights into the evolutionary dynamics of these rodent species.

Mitochondrial Genome Assembly Challenges: Lessons from *Ellobius lutescens*

To comprehensively study the NUMT genome composition, it is necessary to obtain *E. lutescens* and *E. talpinus* complete sequence mtDNA. The GetOrganelle pipeline enabled the successful assembly of circular mitochondrial DNA in both species. Previously, the mitochondrial genome of *E. lutescens* (MT483992.1) had two gaps of unknown nucleotide composition, with a total of 39 nucleotides. The new genomic assembly fills these gaps, resulting in the first complete mitochondrial assembly of this species.

However, the new assembly included extended tandem duplications in the control region (O_H sequence), which were repeated two-three times. Analysis of the duplicated sequence length using PCR suggested that this duplication was an artifact of the assembly process. It is probable that this duplication occurred because of the use of short reads and the De Bruin graph for sequence assembly[47]. Although GetOrganelle has demonstrated the highest accuracy among all published plastid assembly pipelines in benchmarks[48–50], this observation underscores the importance of the molecular validation of *in silico* results.

NUMT Dynamics Unveiled: Contrasting Abundance, Length, and Composition in *Ellobius lutescens* and *Ellobius talpinus*

Our analysis of NUMTs in *E. lutescens* and *E. talpinus* revealed noteworthy differences in abundance, total length, and composition. Specifically, *E. talpinus* displayed a higher number and total length of NUMTs, indicating greater propensity for NUMT integration. This disparity may be attributed to variations in the mechanisms and rates of NUMT transfer between mitochondrial and nuclear genomes. Previously, high NUMT polymorphisms in the genomes, in terms of both copy number and total length, were also observed in some closely related species of fig wasps[51], fruit flies[52], bats[27], marsupial shrews[53] and primates[25].

The structuring of NUMTs into clusters with homologous mtDNA sequences interspersed by spacer sequences suggests multiple rounds of insertion or duplication, which may have been facilitated by non-homologous end-joining, replication-dependent DNA double-strand break repair and/or

retrotransposition[21,54]. The presence of NUMT clusters has also been described in some plants[55] and yeast[56].

The predominance of short NUMTs could also be attributed to integration mechanism into the nuclear genome. During the transfer of mitochondrial DNA to the nuclear genome, longer fragments may be more prone to degradation or truncation. Additionally, the breakage and repair processes that facilitate the integration of mtDNA into the nuclear genome may favor shorter segments because of their easier incorporation and less disruptive impact on nuclear genome stability.

Analysis of the composition of RNA-encoded sequences in NUMTs revealed that they were the most prevalent fragments in both genomes. This led us to propose a hypothesis that the higher concentration of RNA genes, including tRNAs and rRNAs, compared to protein-coding genes, increases the likelihood of integration into the nuclear genome. The abundance of these RNA molecules in cytoplasm[57,58] might lead to preferential transfer or integration of these sequences into the nuclear genome because of their availability.
The higher level of identity in *E. talpinus* NUMTs suggests a more recent transfer of these sequences to the nuclear genome. The high identity implies less time for mutations to accumulate in these NUMTs since their integration, suggesting a dynamic and ongoing process of mitochondrial DNA incorporation into the nuclear genome in *E. talpinus*.

Decoding the NUMT Landscape: Genomic Signatures and Repetitive Element Associations

Examination of the genomic context of NUMTs provided noteworthy findings. NUMTs are predominantly located within the intergenic regions and are enriched in introns, a characteristic feature of most studied NUMTs[20,23,25,59,60]. This distribution suggests that NUMTs in these regions face less selective pressure than those within coding regions. The presence of NUMTs within introns may be due to lower functional constraints in these regions, allowing for their integration without significantly disrupting gene function.

Additionally, the presence of repeated elements, particularly LINE, SINE, and LTR, in the flanking regions of NUMTs, which has also been observed in various species[21,23,25], suggests their possible involvement in NUMT insertion and mobilization.

Deciphering SNV Complexity: Insights and Challenges in Mitochondrial Genomic Analysis

Mitochondrial single nucleotide variant (SNV) analysis in *E. lutescens* and *E. talpinus* samples revealed intriguing patterns. In *E. lutescens*, the detection of several SNVs, with four exhibiting high coverage, suggests subtle heteroplasmy. However, distinguishing mitochondrial variations

from nuclear mitochondrial DNA poses challenges, necessitating advanced sequencing approaches, especially long-read data.

Caution is crucial when interpreting SNV data in *E. talpinus* because of the prevalence of NUMT sequences. Significant SNV coverage in ETAL 2 (Fugure 4) suggests a NUMT association, supported by distribution patterns resembling NUMT clusters (Figure 2). This highlights the challenges in accurate mitochondrial DNA analysis and underscores the need for careful consideration in species with high NUMT prevalence. A notable absence of SNVs from the NUMTs was observed in ETAL 1 (Table 1). Plausible reasons include low coverage and muscle tissue bias towards heteroplasmic mitochondrial variants owing to its higher mtDNA content. Choosing the muscle tissue for barcoding may enhance reliability by minimizing NUMT-related amplification bias. NUMT co-amplification is known to occur in species that have experienced a recent NUMT insertion and poses a major challenge to the interpretation of DNA barcoding results[61–71]. A recent study revealed the co-amplification of NUMT-derived sequences within a control region fragments in *E. talpinus*, which could result in erroneous outcomes in population analysis[34]. Therefore, it is essential to account for NUMTs in molecular studies and employ additional sequencing techniques and methodologies to ensure the fidelity of species barcoding.

Conclusion

In this study of *Ellobius lutescens* and *Ellobius talpinus*, we systematically characterized NUMTs, revealing distinct patterns in their abundance, length, and genomic distribution. Our analysis indicates significant differences between the two species, with *E. talpinus* exhibiting more recent NUMT insertions. The prevalence of short NUMTs, often clustered and predominantly located in non-coding regions, suggests complex integration mechanisms. The identification of LINE, SINE, and LTR elements near NUMT clusters implies their role in NUMT mobilization. Our work highlights the challenges in distinguishing mitochondrial DNA from NUMTs in genomic studies, underscoring the need for careful data interpretation and advanced sequencing approaches. This research contributes to the understanding of NUMT dynamics in mole voles and their potential impact on genomic evolution.

References

1. Wilson, D. & Solari, S. Mammal species of the world: a taxonomic and geographic reference. *Journal of Mammalogy* (2007).
2. Kryštufek, B. & Shenbrot, G. *Voles and Lemmings (Arvicolinae) of the Palaearctic Region*. (University of Maribor, University Press, 2022). doi:10.18690/um.fnm.2.2022.

3. Smorkatcheva, A. V. & Lukhtanov, V. A. Evolutionary association between subterranean lifestyle and female sociality in rodents. *Mamm Biol* **79**, 101–109 (2014).
4. Just, W. *et al.* The sex determination in Ellobius lutescens remains bizarre. *Cytogenet Genome Res* **96**, 146–153 (2002).
5. Just, W. *et al.* Ellobius lutescens: Sex Determination and Sex Chromosome. *Sexual Development* **1**, 211–221 (2007).
6. Matveevsky, S., Bakloushinskaya, I. & Kolomiets, O. Unique sex chromosome systems in Ellobius: How do male XX chromosomes recombine and undergo pachytene chromatin inactivation? *Sci Rep* **6**, 29949 (2016).
7. Mulugeta, E. *et al.* Genomes of Ellobius species provide insight into the evolutionary dynamics of mammalian sex chromosomes. *Genome Research* **26**, 1202–1210 (2016).
8. Gil-Fernández, A. *et al.* Sex differences in the meiotic behavior of an XX sex chromosome pair in males and females of the mole vole Ellobius tancrei: turning an X into a Y chromosome? *Chromosoma* **130**, 113–131 (2021).
9. Romanenko, S. A. *et al.* Chromosomal evolution of Arvicolinae (Cricetidae, Rodentia). II. The genome homology of two mole voles (genus Ellobius), the field vole and golden hamster revealed by comparative chromosome painting. *Chromosome Res* **15**, 891–897 (2007).
10. Matveevsky, S., Kolomiets, O., Bogdanov, A., Hakhverdyan, M. & Bakloushinskaya, I. Chromosomal Evolution in Mole Voles Ellobius (Cricetidae, Rodentia): Bizarre Sex Chromosomes, Variable Autosomes and Meiosis. *Genes* **8**, 306 (2017).
11. Bogdanov, A., Tambovtseva, V., Matveevsky, S. & Bakloushinskaya, I. Speciation on the Roof of the World: Parallel Fast Evolution of Cryptic Mole Vole Species in the Pamir-Alay—Tien Shan Region. *Life* **13**, 1751 (2023).
12. Bondareva, O. V. *et al.* The complete mitochondrial genomes of three Ellobius mole vole species (Rodentia: Arvicolinae). *Mitochondrial DNA Part B* **5**, 2485–2487 (2020).
13. Wang, X. *et al.* Phylogenomics reveals the evolution, biogeography, and diversification history of voles in the Hengduan Mountains. *Commun Biol* **5**, 1–12 (2022).
14. Abramson, N. I., Bodrov, S. Yu., Bondareva, O. V., Genelt-Yanovskiy, E. A. & Petrova, T. V. A mitochondrial genome phylogeny of voles and lemmings (Rodentia: Arvicolinae): Evolutionary and taxonomic implications. *PLoS One* **16**, e0248198 (2021).
15. Bondareva, O. *et al.* Signatures of Adaptation in Mitochondrial Genomes of Palearctic Subterranean Voles (Arvicolinae, Rodentia). *Genes* **12**, 1945 (2021).
16. Bondareva, O. V., Potapova, N. A., Konovalov, K. A., Petrova, T. V. & Abramson, N. I. Searching for signatures of positive selection in cytochrome b gene associated with

subterranean lifestyle in fast-evolving arvicolines (Arvicolinae, Cricetidae, Rodentia). *BMC Ecology and Evolution* **21**, 92 (2021).

17. Tsuzuki, T., Nomiyama, H., Setoyama, C., Maeda, S. & Shimada, K. Presence of mitochondrial-DNA-like sequences in the human nuclear DNA. *Gene* **25**, 223–229 (1983).
18. Bensasson, D., Feldman, M. W. & Petrov, D. A. Rates of DNA Duplication and Mitochondrial DNA Insertion in the Human Genome. *J Mol Evol* **57**, 343–354 (2003).
19. Puertas, M. J. & González-Sánchez, M. Insertions of mitochondrial DNA into the nucleus—effects and role in cell evolution. *Genome* **63**, 365–374 (2020).
20. Hazkani-Covo, E., Zeller, R. M. & Martin, W. Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes. *PLOS Genetics* **6**, e1000834 (2010).
21. Tsuji, J., Frith, M. C., Tomii, K. & Horton, P. Mammalian NUMT insertion is non-random. *Nucleic Acids Research* **40**, 9073–9088 (2012).
22. Ricchetti, M., Tekaia, F. & Dujon, B. Continued Colonization of the Human Genome by Mitochondrial DNA. *PLOS Biology* **2**, e273 (2004).
23. Calabrese, F. M. *et al.* NumtS colonization in mammalian genomes. *Sci Rep* **7**, 16357 (2017).
24. Grau, E. T. *et al.* Survey of mitochondrial sequences integrated into the bovine nuclear genome. *Sci Rep* **10**, 2077 (2020).
25. Dayama, G., Zhou, W., Prado-Martinez, J., Marques-Bonet, T. & Mills, R. E. Characterization of nuclear mitochondrial insertions in the whole genomes of primates. *NAR Genomics and Bioinformatics* **2**, lqaa089 (2020).
26. Féménia, M., Charles, M., Boulling, A. & Rocha, D. Identification and characterisation of mitochondrial sequences integrated into the ovine nuclear genome. *Animal Genetics* **52**, 556–559 (2021).
27. Zhang, G. *et al.* Genomic landscape of mitochondrial DNA insertions in 23 bat genomes: characteristics, loci, phylogeny, and polymorphism. *Integrative Zoology* **17**, 890–903 (2022).
28. Anello, M., García Folco, G. & Di Rocco, F. Detection and characterization of nuclear mitochondrial DNA (NUMTs) in the alpaca genome. *Small Ruminant Research* **224**, 106986 (2023).
29. Patterson, E. C. *et al.* Defining cat mitogenome variation and accounting for numts via multiplex amplification and Nanopore sequencing. *Forensic Science International: Genetics* **67**, 102944 (2023).
30. Lang, M. *et al.* Polymorphic NumtS trace human population relationships. *Hum Genet* **131**, 757–771 (2012).

31. Miraldo, A., Hewitt, G. M., Dear, P. H., Paulo, O. S. & Emerson, B. C. Numts help to reconstruct the demographic history of the ocellated lizard (Lacerta lepida) in a secondary contact zone. *Molecular Ecology* **21**, 1005–1018 (2012).
32. Pérez, T., Rodríguez, F., Fernández, M., Albornoz, J. & Domínguez, A. Ancient mitochondrial pseudogenes reveal hybridization between distant lineages in the evolution of the Rupicapra genus. *Gene* **628**, 63–71 (2017).
33. Lucas, T., Vincent, B. & Eric, P. Translocation of mitochondrial DNA into the nuclear genome blurs phylogeographic and conservation genetic studies in seabirds. *Royal Society Open Science* **9**, 211888 (2022).
34. Kuprina, K., Smorkatcheva, A., Rudyk, A. & Galkina, S. Numerous insertions of mitochondrial DNA in the genome of the northern mole vole, Ellobius talpinus. *Mol Biol Rep* **51**, 36 (2023).
35. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
36. Jin, J.-J. *et al.* GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology* **21**, 241 (2020).
37. Donath, A. *et al.* Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. *Nucleic Acids Research* **47**, 10543–10552 (2019).
38. Atamaniuk, J. *et al.* Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **8**, 927–941 (2015).
39. Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biology* **20**, 8 (2019).
40. Untergasser, A. *et al.* Primer3—new capabilities and interfaces. *Nucleic Acids Res* **40**, e115 (2012).
41. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *PNAS* **117**, 9451–9457 (2020).
42. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
43. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, Unit 4.10 (2009).
44. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435–W439 (2006).
45. Quinlan, A. R. BEDTools: the Swiss-army tool for genome feature analysis. *Curr Protoc Bioinformatics* **47**, 11.12.1-11.12.34 (2014).
46. Gamperl, R., Ehmann, Ch. & Bachmann, K. Genome size and heterochromatin variation in rodents. *Genetica* **58**, 199–212 (1982).

47. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**, 36–46 (2012).
48. Mahar, N. S., Satyam, R., Sundar, D. & Gupta, I. A systematic comparison of human mitochondrial genome assembly tools. *BMC Bioinformatics* **24**, 341 (2023).
49. Giorgashvili, E. *et al.* Software Choice and Sequencing Coverage Can Impact Plastid Genome Assembly–A Case Study in the Narrow Endemic Calligonum bakuense. *Frontiers in Plant Science* **13**, (2022).
50. Freudenthal, J. A. *et al.* A systematic comparison of chloroplast genome assembly tools. *Genome Biol* **21**, 254 (2020).
51. Wang, J.-X., Liu, J., Miao, Y.-H., Huang, D.-W. & Xiao, J.-H. Tracking the Distribution and Burst of Nuclear Mitochondrial DNA Sequences (NUMTs) in Fig Wasp Genomes. *Insects* **11**, 680 (2020).
52. Rogers, H. H. & Griffiths-Jones, S. Mitochondrial Pseudogenes in the Nuclear Genomes of Drosophila. *PLOS ONE* **7**, e32593 (2012).
53. Hazkani-Covo, E. A Burst of Numt Insertion in the Dasyuridae Family During Marsupial Evolution. *Frontiers in Ecology and Evolution* **10**, (2022).
54. Ju, Y. S. *et al.* Frequent somatic transfer of mitochondrial DNA into the nuclear genome of human cancer cells. *Genome Res.* **25**, 814–824 (2015).
55. Michalovova, M., Vyskot, B. & Kejnovsky, E. Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. *Heredity* **111**, 314–320 (2013).
56. Sacerdot, C. *et al.* Promiscuous DNA in the nuclear genomes of hemiascomycetous yeasts. *FEMS Yeast Research* **8**, 846–857 (2008).
57. Held, J. P. & Patel, M. R. Functional conservation of mitochondrial RNA levels despite divergent mtDNA organization. *BMC Research Notes* **13**, 334 (2020).
58. Ribas de Pouplana, L. The mitochondrial tRNA conundrum. *Nat Rev Mol Cell Biol* **21**, 361–361 (2020).
59. Richly, E. & Leister, D. NUMTs in Sequenced Eukaryotic Genomes. *Molecular Biology and Evolution* **21**, 1081–1084 (2004).
60. Wei, W. *et al.* Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. *Nature* **611**, 105–114 (2022).
61. Song, H., Buhay, J. E., Whiting, M. F. & Crandall, K. A. 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 (2008).

62. Buhay, J. E. “COI-like” Sequences Are Becoming Problematic in Molecular Systematic and DNA Barcoding Studies. *Journal of Crustacean Biology* **29**, 96–110 (2009).
63. Moulton, M. J., Song, H. & Whiting, M. F. 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 (2010).
64. Kim, S.-J., Lee, K. Y. & Ju, S.-J. Nuclear mitochondrial pseudogenes in Austinograea alayseae hydrothermal vent crabs (Crustacea: Bythograeidae): effects on DNA barcoding. *Molecular Ecology Resources* **13**, 781–787 (2013).
65. Jordal, B. H. & Kambestad, M. DNA barcoding of bark and ambrosia beetles reveals excessive NUMTs and consistent east-west divergence across Palearctic forests. *Molecular Ecology Resources* **14**, 7–17 (2014).
66. Ermakov, O. A. *et al.* Implications of Hybridization, NUMTs, and Overlooked Diversity for DNA Barcoding of Eurasian Ground Squirrels. *PLOS ONE* **10**, e0117201 (2015).
67. Ahmed, N. S. & Jaffar Ali, H. A. Numts: an impediment to DNA barcoding of Polyclinids, Tunicata. *Mitochondrial DNA Part A* **27**, 3395–3398 (2016).
68. da Cruz, M. O. R. *et al.* DNA barcoding of the rodent genus Oligoryzomys (Cricetidae: Sigmodontinae): mitogenomic-anchored database and identification of nuclear mitochondrial translocations (Numts). *Mitochondrial DNA Part A* **30**, 702–712 (2019).
69. Françoso, E. *et al.* Conserved numts mask a highly divergent mitochondrial-COI gene in a species complex of Australian stingless bees Tetragonula (Hymenoptera: Apidae). *Mitochondrial DNA Part A* **30**, 806–817 (2019).
70. Schultz, J. A. & Hebert, P. D. N. Do pseudogenes pose a problem for metabarcoding marine animal communities? *Molecular Ecology Resources* **22**, 2897–2914 (2022).
71. Hebert, P. D. N., Bock, D. G. & Prosser, S. W. J. Interrogating 1000 insect genomes for NUMTs: A risk assessment for estimates of species richness. *PLOS ONE* **18**, e0286620 (2023).

Funding

This study was supported by a research Grant of the Russian Science Foundation (RSF 19-14-00034-П).

Contributions

D.P.: conceptualization; D.P.: methodology; T.S.: software, D.P. and S.R.: validation, D.P. and T.S.: formal analysis; D.P., I.B., T.S., and S.R.: investigation; A.G.: resources, D.P. and T.S.: data curation; D.P.: writing – original draft preparation, D.P., I.B., T.S., S.R., and A.G.: writing – review and editing; D.P., and T.S.: visualization; D.P.: supervision; S.R., and A.G.: project administration; A.G.: funding acquisition. All authors reviewed and approved the final manuscript.

Corresponding author

Correspondence to Dmitry Prokopov.

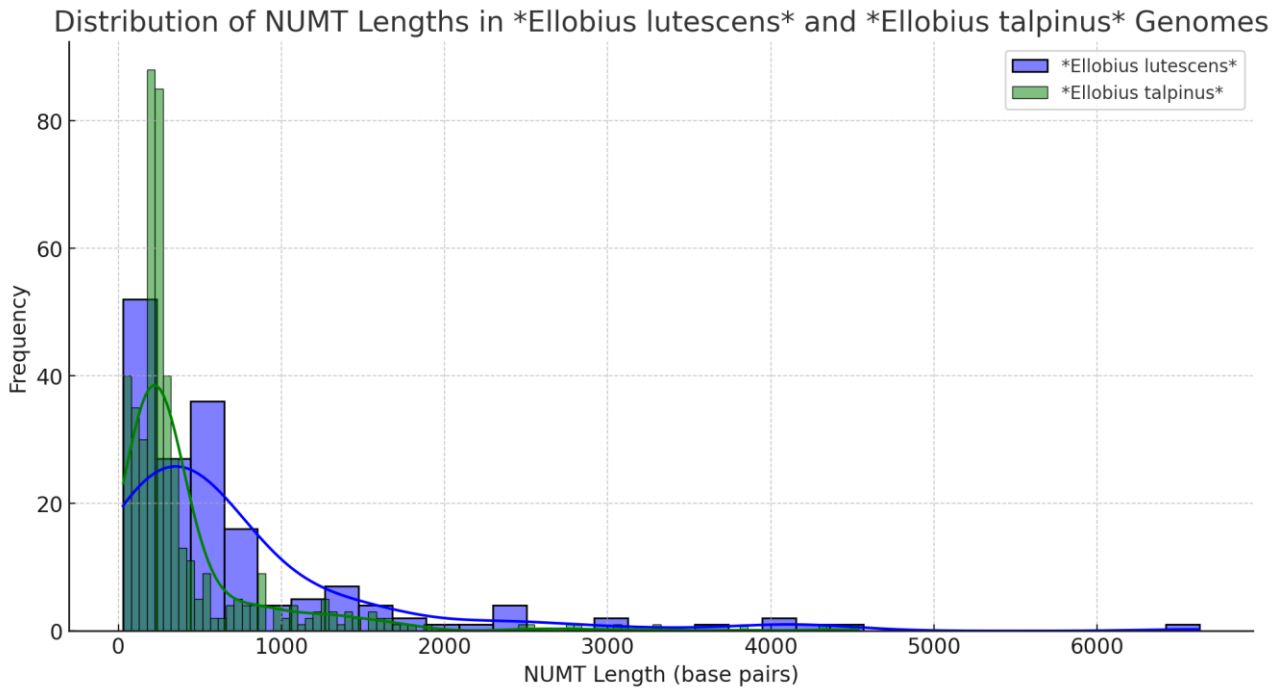
Competing interests

The author(s) declare no competing interests.

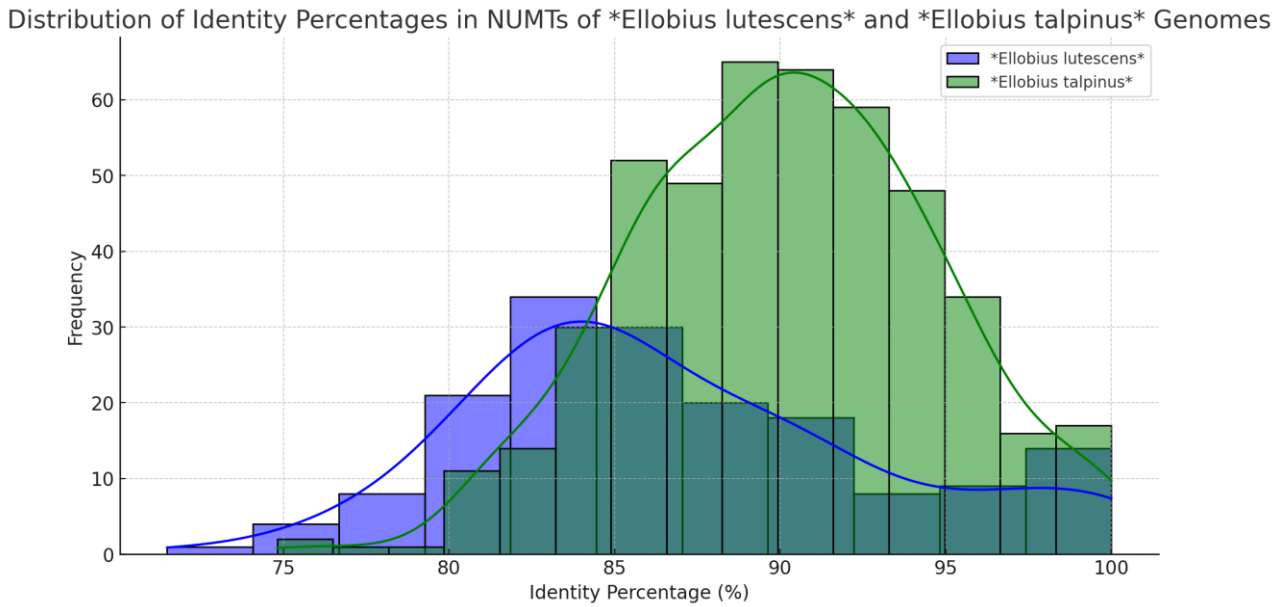
Supplementary files

Genomic Underground: Unraveling NUMTs in Mole Voles

Dmitry Prokopov, Tigran Saluev, Svetlana Romanenko, Irina Bakloushinskaya, Alexander Graphodatsky

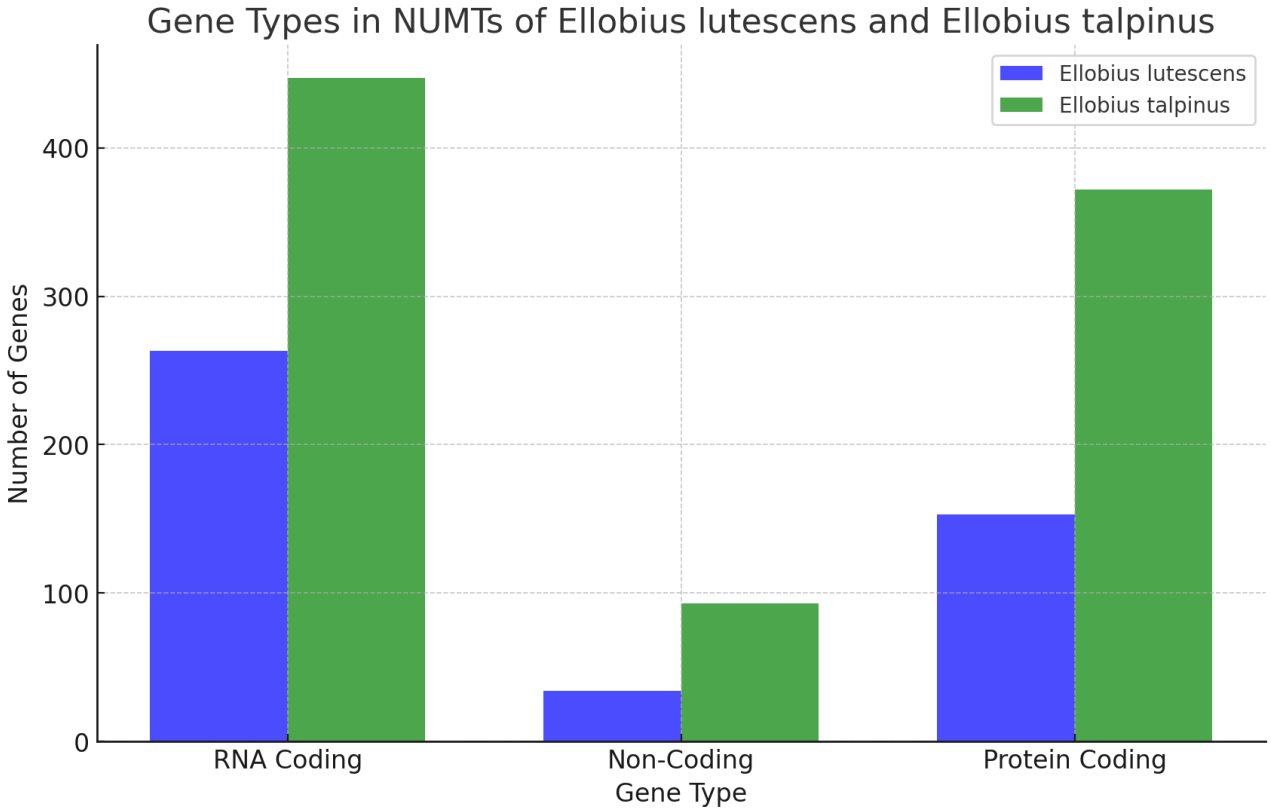


Supplementary Figure 1. Distribution of NUMT lengths in the genomes of Ellobius lutescens (in blue) and Ellobius talpinus (in green). The x-axis represents the length of NUMTs in base pairs, and the y-axis shows the frequency of each length range.



Supplementary Figure 2. Distribution of identity percentages for NUMTs in the genomes of Ellobius lutescens (in blue) and Ellobius talpinus (in green). The x-axis represents the percentage of identity, and the y-axis shows the frequency of each identity range.

310



Supplementary Figure 3. The types of genes present in NUMTs of Ellobius lutescens and Ellobius talpinus. The chart shows the number of RNA coding genes (including tRNA and rRNA), non-coding genes (OL and control region), and protein-coding genes in NUMTs for each species.

Supplementary Table 1. List of primer pairs

Primer pair	F-Tm	RTm	F	R
ELUT_D	68.26	68.27	TGCTCATTCCTCGATCCAAACGTCTATC	CAAACCTTTGTGTTTATGGGATGCTTGC
ELUT_D_inn	68.57	68.40	CTTGATCTACCATCCTCCGTGAAACCAAC	AATGGCCCTGAAGAAAGAACCAGATGTC
ETAL_D	68.53	68.55	TCTTCTCTTCCATACGACTATCCCTGTCCC	GGAACACTCGAATGGGCGAAGGAC
ETAL_D_inn	68.28	68.03	AATCCCTTGATCTATCATCCTCCGTGAAAC	TTGACGGCTATGCTGAGTGATACCATACC
ETAL_N1	54.37	55.25	ATTCATAGGCTTAGGGCTAGTG	CCACCCTATCCCTCACACTAG
ETAL_N2	55.33	56.09	TATTGGGTAACCAGCTATCACC	CCCGAAACCAAACGAGC
ETAL_N3	55.28	55.68	TATTCGCCGTAATAGCAACAG	GGGTTGTTGGATCCTGTTTC
ETAL_N4	54.56	54.90	AACGTTGGTTTAAATCCTTCC	ATATGAGAAATAGGTTGGAGATGAG
ETAL_N5	55.23	55.31	GTCGTGGTATGCCGGATAG	AGTTGCTCACTTCCACTACGTC
ETAL_N6	55.40	55.59	GAGTGAGAAGATAAAGGCTAGGG	TCAGCAGGCTCTAGTAAACCC