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

Mitochondrial DNA transfer to the nuclear genome results in the emergence of nuclear DNA sequences of mitochondrial origin (NUMTs), which are recognized as significant factors in genome evolution. This study addresses the gaps in our understanding *Ellobius lutescens* and *Ellobius talpinus* mitochondrial DNA (mtDNA) variations, focusing on nuclear mitochondrial DNA segments NUMTs and their impact on evolutionary dynamics and species barcoding. Using advanced assembly techniques, we obtained circularized mtDNA sequences for both species. Detailed genomic analyses explored NUMT abundance, composition, and distribution in relation to repeated elements. Mitochondrial single nucleotide variants (SNVs) were identified in *E. lutescens* and *E. talpinus* samples. *E. lutescens* exhibits unique mtDNA structural features, with a longer and more heavily melted mtDNA compared to other rodents. NUMT analysis revealed species-specific differences, clustering patterns, and repeated element associations. *E. talpinus* displayed notable SNV coverage variations, suggesting NUMT association. Our findings illuminate the intricate mtDNA variations in mole voles, emphasizing the importance of advanced sequencing for accurate analysis. NUMT complexity poses challenges in evolutionary studies and species barcoding, and requires careful consideration in molecular analyses. This study enhances our understanding of genomic intricacies in these rodents and, contributes valuable insights into their evolutionary dynamics.

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, complete mitochondrial genomes of four *Ellobius* species have been sequenced: *Ellobius lutescens, Ellobius fuscocapillus, Ellobius talpinus*[12], and *Ellobius tancrei*[13]. Their mitochondrial DNAs (mtDNAs) are typical of the subfamily Arvicolinae but possess a unique feature of high GC content[14].

A captivating facet of genomic exploration involves nuclear mitochondrial DNA segments (NUMTs), which are fragments of mitochondrial DNA integrated into the nuclear genome.

Discovered after sequencing of the first human mitochondrial genome in 1983[15], NUMTs were believed to traverse the nucleus through various mechanisms, such as homologous recombination[16], non-homologous end joining[17], microhomology-mediated end joining[18], and retrotransposition[19]. Although NUMTs are prevalent across eukaryotic genomes, the largest number of NUMTs has been described in mammals[18,20–27]. These repeated sequences can be utilized for various research purposes, such as population genetics[28], past demographic event detection[29], inference of hybridization events[30], and biogeographic analysis[31].

Investigation of NUMTs in *Ellobius* species presents a unique perspective ащк deciphering the genomic complexities of these subterranean rodents. In this study, our primary objective was to explore the complex landscape of NUMTs within *E. lutescens* and *E. talpinus*, focusing on their diverse structural patterns and their potential impact on evolutionary dynamics. To achieve this, we utilized advanced assembly methodologies and conducted an extensive bioinformatic analysis, to provide a comprehensive understanding of the NUMT scenario in mole voles. This study provides valuable insights into the intricate interplay between mitochondrial and nuclear genomes, shedding light on the intricate dynamics of these fascinating 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, adapter sequences and PCR duplicates were removed using fastp 0.23.4[32] 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 mitogenome assembly. We used the assembler GetOrganelle v1.7.7.0[33] with the option -F animal_mt to assemble the mitogenomes. All available mitochondrial sequences of *Ellobius* from GenBank were used as the seed sequences. Protein-coding genes, non-coding regions, and RNA genes were annotated using MITOS2[34]. To detect mtDNA variants, the reads used for assembly were aligned against newly assembled mitochondria using the BWA-MEM algorithm 0.7.17-r1188[35]. Variant calling was performed using Ivar 1.4.2[36].

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). The search was performed using BLASTn version 2.14.1, with an e-value $\leq 5\times10^{-3}$. A set of High-scoring Segment Pairs (HSPs) were subsequently filtered to prevent possible mitochondrial contamination. HSPs with identity greater than 99% were removed. The 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 carried out annotation. For the identification and classification of repetitive sequences, we utilized RepeatModeler2 version 2.0.4, with the option -LTRStruct [37] for *de novo* and *ab initio* repeat identification. Similar sequences of repeated elements were clustered using CD-HIT 4.8.1 [38]. We used RepeatMasker version 4.1.5 [39] with the parameters -e rmblast -s -no_is to find 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 [40] 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 carried out using bedtools 2.31.1 [41] and custom scripts avaliable at https://github.com/Saluev/…

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* (Table 2).

Table 2. 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	D-loop	
								size	GC-content
ELUT NCBI	MT483992.1	16540	31.1	25.9	14.3	28.4	42.8	1172	41.5
ELUT 1		18056	31	26.1	14.4	28.5	42.9	2382	41.4
ELUT 2		18123	30.9	26.1	14.5	28.5	43	2440	41.6
ELUT 3		18179	30.9	26.1	14.5	28.5	43	2440	41.6
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

The newly assembled *E. lutescens* mitogenome was considerably longer than that utilized in the study by Bondareva et al.[12], with nucleotide differences ranging from 1,516 to 1,639 bp. The presence of assembled hypervariable D-loop region duplication, consisting of 2-3 duplicated regions and a light-strand replication origin that had not been previously assembled, accounts for the difference in our assembly compared with previous studies. We successfully assembled a complete circular sequence of mitochondrial DNA without any gaps (Figure 1).

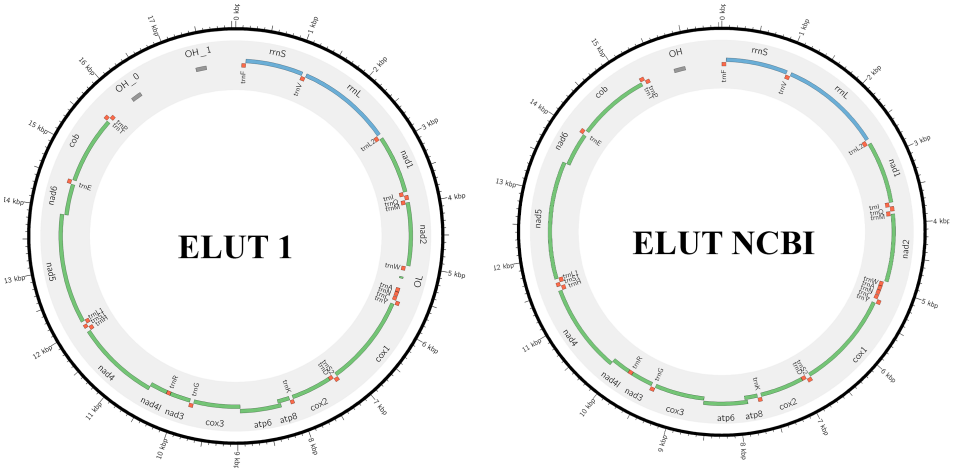

Figure 1. Visualization of the assembly and annotation of the complete mitochondrial DNA of E. lutescens (ELUT 1) obtained in this work and available in the NCBI database (ELUT NCBI)

The length of the assembly (18,056-18,179 bp) and the elevated GC composition (42.9-43%%) suggest that the mitochondrial DNA sequence of *E. lutescens* is the longest and most heavily melting among all living rodents for which an assembly of complete mitochondrial genomes is available in the NCBI Nucleotide database by 10/24/23, as depicted in Figure 2.

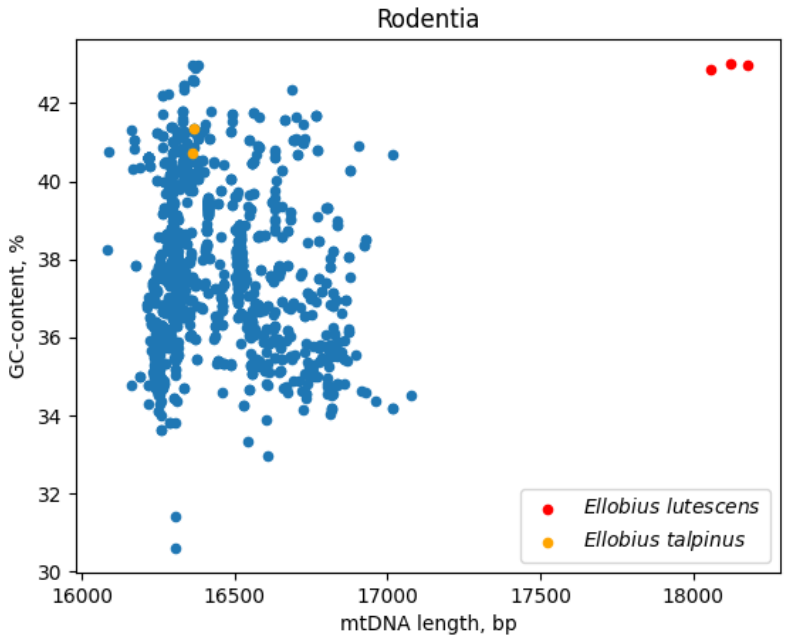

Figure 2. A scatterplot comparing the GC composition and length of complete mitochondrial DNAs from NCBI's database (blue dots). It also displays the GC composition and length of E. lutescens (red dots) and E. talpinus (yellow dots), which were assembled in this study.

Bioinformatic analysis of the mitochondrial genome in the nuclear genomes of *E. lutescens* and *E. talpinus* revealed the presence of 208 and 485 fragments, respectively. The total length of NUMTs in the *E. talpinus* and *E. lutescens* genomes were 134049 and 199619 bp, respectively, which accounted for 0.0057% and 0.0088% of the assembly length, or 0.0045% and 0.0068% of the nuclear genome, respectively, based on previous genome amount estimations [42]. The most frequently occurring fragments of the mtDNA sequence in the nuclear genome are non-coding mtDNA sequences such as D-loop, rrnS, and rrnL ribosomal RNA. The lowest number of mtDNA fragments found in the nuclear genome was that of tRNAs (Figure 3).

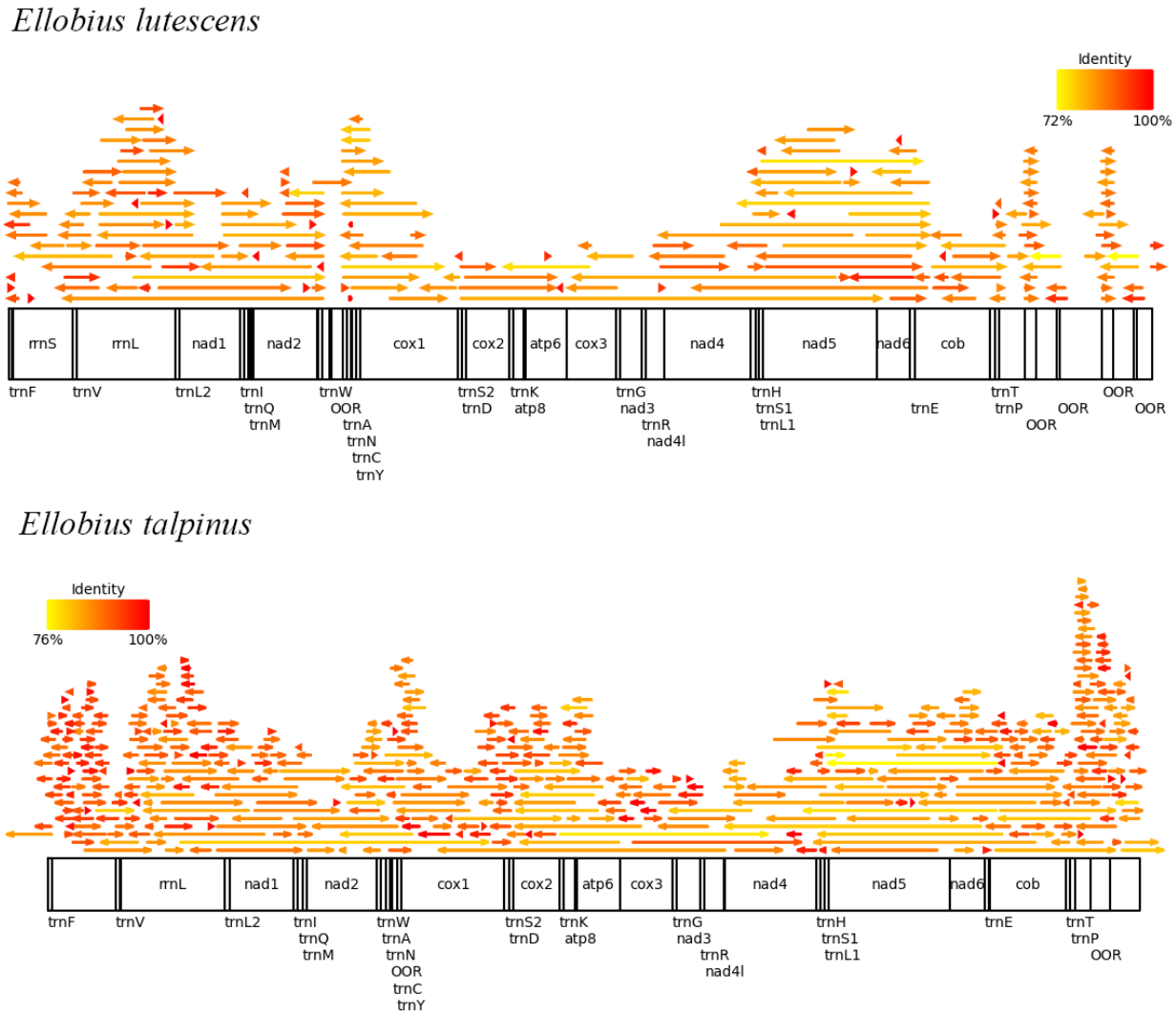

Figure 3. Visual representation of NUMT alignments on the mitochondrial DNA map of Ellobius lutescens and Ellobius talpinus. The arrangement of the sequences is demonstrated using arrows, with the direction of the sequence indicated. The colour scheme, which range from yellow to red, conveys the degree of identity.

33.3% and 22.8% of NUMTs in *E. lutescens* and *E. talpinus* organised in clusters, containing homologous mtDNA sequences interrupted with spacer sequences.

Genomic context of NUMTs

Analysis of NUMTs in relation to genomic annotations revealed that the majority of NUMTs were located in the intergenic regions of the genome, with a notable proportion of NUMTs detected in gene regions (Figure 4). 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 gene.

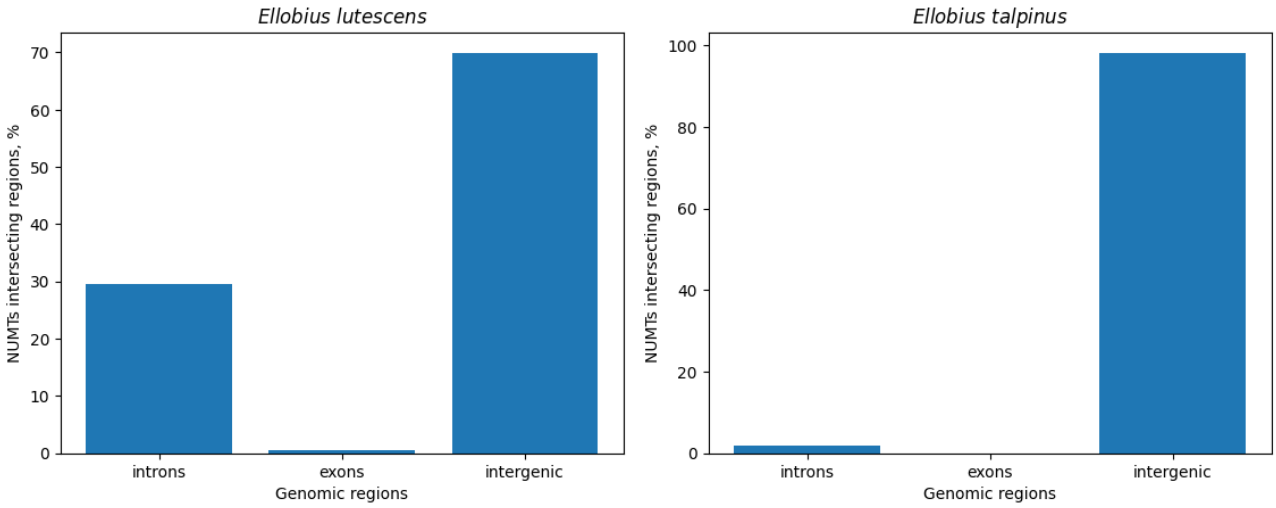

Figure 4: Location of NUMTs in the genomic regions of two mole vole species.

A noticeable prevalence of repeated elements was detected in the flanking regions of the NUMTs, with an abundance of LINE, SINE, and LTR superfamily elements within these regions (Figure 5). Analysis of *E. talpinus* revealed a distinct pattern, with a higher frequency of LINE, SINE, and LTR superfamily elements within NUMT clusters than in *E. lutescens*. LINE elements were the most prevalent, constituting 56.67% of the total repetitive elements in the flanking regions of the NUMTs in *E. talpinus*. SINE elements followed closely in abundance, representing a substantial fraction of the total repeated elements, at 43.33% in *E. talpinus* and 44.00% in *E. lutescens*. Furthermore, LTR elements were observed to a considerable extent, comprising 30.67% and 26.00% of the total repetitive elements in the flanking regions of the NUMTs in *E. talpinus* and *E. lutescens*, respectively.

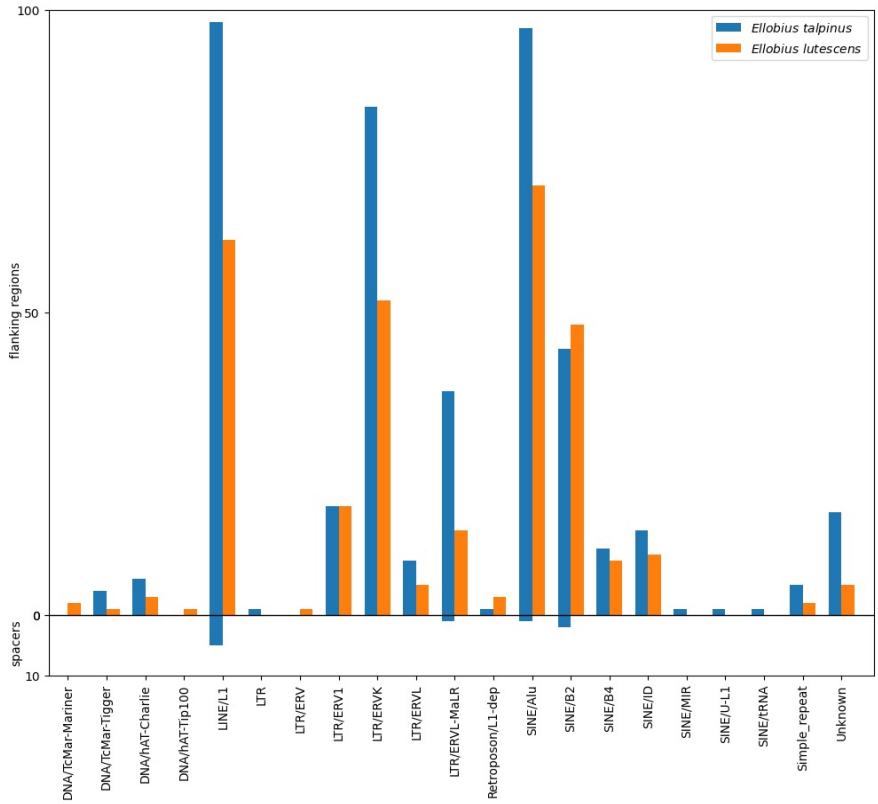

Figure 5: Content of repeated elements in the internal and flanking regions of NUMT clusters

Nucleotide variants and coverage analysis

Several mitochondrial single nucleotide variants (SNV) were identified in *E. lutescens* samples, and among these, four possessed a considerable degree of coverage (Figure 6).

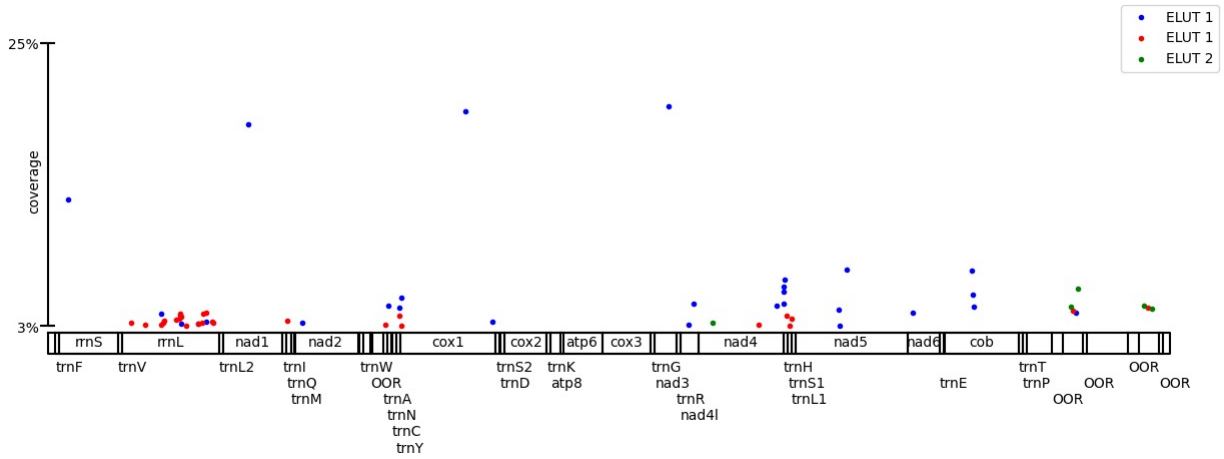

Figure 6: SNV composition in mitochondrial genomes of Ellobius lutescens

An unusual observation was made regarding the *E. talpinus* samples. Specifically, for 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 3). Conversely, for ETAL 1, only a small number of nucleotide variants were observed, and these were mostly of low coverage.

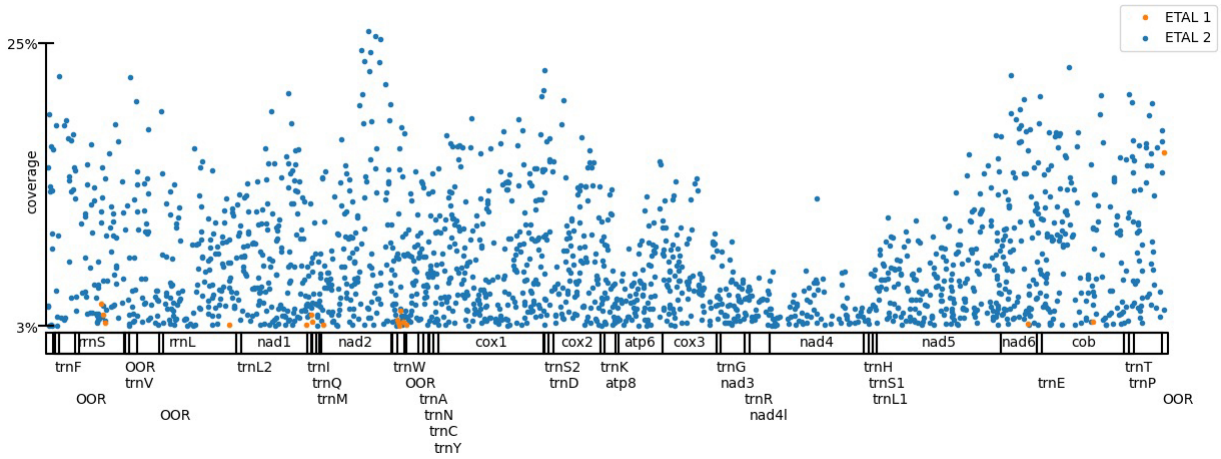

Figure 7: SNV composition in mitochondrial genomes of Ellobius talpinus

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*. While our primary focus is on NUMTs, it is important to acknowledge the crucial first step in assembling completely circularized mitochondrial DNA sequences. A detailed exploration of NUMTs uncovers intricate genomic complexities, offering insights into the dynamic interaction between mitochondrial and nuclear genomes in these closely related species. 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.

Structural Differences in Mitochondrial Genomes: Insights from Ellobius lutescens

To comprehensively study the NUMT genome composition, it is necessary to obtain *E. lutescens* and *E. talpinus* complete sequence mtDNA. Our assembly approach revealed that *E. lutescens* has a considerably larger mtDNA size, which is primarily attributed to the assembly of hypervariable D-loop region duplications. The D-loop (also known as the control region) is the most variable part of the mitochondrial genome [43]. Several proposed mechanisms for the formation of duplications in the mitochondrial genome include slipped strand mispairing facilitated by repeat units or sequences forming secondary structures during DNA replication, over-running of the termination signal, initiation of replication at secondary structure sites other than the origin, and unequal crossing over during recombination between two mitochondrial genomes within a single mitochondrion [44]. This structural feature underscores the dynamic nature of mtDNA and its potential for significant variation within species.

The length and GC composition of *E. lutescens* mtDNA suggest that it is the longest and most heavily melted among living rodents, potentially influencing mitochondrial function and contributing to physiological adaptation. Some representatives of the Arvicolinae subfamily, including species of the genus *Ellobius*, lead an underground lifestyle associated with a large number of morphological, behavioral, and physiological modifications [45]. To date, several studies have been conducted to search for genomic adaptations to the underground lifestyle, and several signatures of adaptation have been discovered in nuclear genes [45] and mitochondrial DNA [46,47]. Although the GC composition was not found to be associated with adaptations to a subterranean

lifestyle [47], obtaining high-quality mitochondrial assemblies for a larger number of species may allow the detection of such associations.

The successful assembly of a complete circular sequence without gaps further validates the accuracy and completeness of our results. This study underscores the importance of employing advanced assembly methods to comprehensively understand mtDNA diversity. The identification of unique structural features in the mtDNA of *E. lutescens* contributes to our understanding of mitochondrial genome organization and evolution, underscoring the ongoing need for advancements in genomic sequencing and assembly methodologies.

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 their 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 [48], fruit flies [49], bats [25], marsupial shrews [50] and primates [23].

Analysis of the composition of non-coding mtDNA sequences in the nuclear genome revealed that they were the most prevalent fragments. This finding is consistent with previous observations in some mammals [21,23,51] and suggests that non-coding mtDNA sequences are more easily integrated into the nuclear genome because of the lack of selective constraints. Conversely, tRNAs, which are essential for protein translation, were the least abundant type of NUMT fragment, indicating that their integration and persistence in the nuclear genome may be hindered.

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 processes such as non-homologous end-joining, replication-dependent DNA double-strand break repair and/or retrotransposition [19,52].

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 intergenic regions and are enriched in introns, a characteristic feature of most studied NUMTs [18,21,23,51,53]. This observation suggests that NUMTs in these regions are subject to less selective pressure than those within gene regions. 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 [19,21,23], implies 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, suggested subtle heteroplasmy. However, distinguishing genuine 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* due to the prevalence of NUMT sequences. Significant SNV coverage in ETAL 2 (Table 1) suggests a NUMT association, supported by distribution patterns resembling NUMT clusters (Figure 3). This highlights the challenges in accurate mitochondrial DNA analysis and underscores the need for careful consideration in species with high NUMT prevalence. In ETAL 1 (Table 1), a notable absence of SNVs from NUMTs is observed. Plausible reasons include low coverage and the muscle tissue bias towards heteroplasmic mitochondrial variants due to its higher mtDNA content. Choosing 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 [54–64]. Therefore, it is essential to account for NUMTs in molecular studies and to employ additional sequencing techniques and methodologies to ensure the fidelity of species barcoding.

Conclusion

Our study represents the first identification of nuclear mitochondrial DNA segments (NUMTs) in mole voles, shedding light on their presence in these species. Despite the shared localization of NUMTs in intergenic and intronic regions accompanied by repeated elements in flanking regions, our investigation uncovered substantial species-specific variations in length and composition within *Ellobius lutescens* and *Ellobius talpinus*. These findings highlight the unique genomic landscapes of these species and emphasize the critical role of NUMTs in shaping their evolutionary trajectories. As NUMTs continue to pose challenges in mitochondrial genomic analysis, particularly in the context of species barcoding, our research contributes to the comprehensive exploration of mole vole genomics. Overall, this study enhances our understanding of the intricate dynamics governing the genomic architecture of mole voles, highlighting the significance of NUMTs in this context.

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. 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).
16. 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).
17. Puertas, M. J. & González-Sánchez, M. Insertions of mitochondrial DNA into the nucleus—effects and role in cell evolution. *Genome* **63**, 365–374 (2020).

18. Hazkani-Covo, E., Zeller, R. M. & Martin, W. Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes. *PLOS Genetics* **6**, e1000834 (2010).
19. Tsuji, J., Frith, M. C., Tomii, K. & Horton, P. Mammalian NUMT insertion is non-random. *Nucleic Acids Research* **40**, 9073–9088 (2012).
20. Ricchetti, M., Tekaia, F. & Dujon, B. Continued Colonization of the Human Genome by Mitochondrial DNA. *PLOS Biology* **2**, e273 (2004).
21. Calabrese, F. M. *et al.* NumtS colonization in mammalian genomes. *Sci Rep* **7**, 16357 (2017).
22. Grau, E. T. *et al.* Survey of mitochondrial sequences integrated into the bovine nuclear genome. *Sci Rep* **10**, 2077 (2020).
23. 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).
24. 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).
25. 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).
26. 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).
27. 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).
28. Lang, M. *et al.* Polymorphic NumtS trace human population relationships. *Hum Genet* **131**, 757–771 (2012).
29. 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).
30. 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).

31. 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).
32. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
33. Jin, J.-J. *et al.* GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology* **21**, 241 (2020).
34. Donath, A. *et al.* Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. *Nucleic Acids Research* **47**, 10543–10552 (2019).
35. Atamaniuk, J. *et al.* Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **8**, 927–941 (2015).
36. 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).
37. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *PNAS* **117**, 9451–9457 (2020).
38. 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).
39. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, Unit 4.10 (2009).
40. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435–W439 (2006).
41. Quinlan, A. R. BEDTools: the Swiss-army tool for genome feature analysis. *Curr Protoc Bioinformatics* **47**, 11.12.1-11.12.34 (2014).
42. Gamperl, R., Ehmann, Ch. & Bachmann, K. Genome size and heterochromatin variation in rodents. *Genetica* **58**, 199–212 (1982).
43. McMillan, W. O. & Palumbi, S. R. Rapid Rate of Control-Region Evolution in Pacific Butterflyfishes (Chaetodontidae). *J Mol Evol* **45**, 473–484 (1997).
44. Schirtzinger, E. E. *et al.* Multiple independent origins of mitochondrial control region duplications in the order Psittaciformes. *Molecular Phylogenetics and Evolution* **64**, 342–356

(2012).

45. Bondareva, O. *et al.* How voles adapt to subterranean lifestyle: Insights from RNA-seq. *Frontiers in Ecology and Evolution* **11**, (2023).
46. 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).
47. Bondareva, O. *et al.* Signatures of Adaptation in Mitochondrial Genomes of Palearctic Subterranean Voles (Arvicolinae, Rodentia). *Genes* **12**, 1945 (2021).
48. 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).
49. Rogers, H. H. & Griffiths-Jones, S. Mitochondrial Pseudogenes in the Nuclear Genomes of Drosophila. *PLOS ONE* **7**, e32593 (2012).
50. Hazkani-Covo, E. A Burst of Numt Insertion in the Dasyuridae Family During Marsupial Evolution. *Frontiers in Ecology and Evolution* **10**, (2022).
51. Wei, W. *et al.* Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. *Nature* **611**, 105–114 (2022).
52. 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).
53. Richly, E. & Leister, D. NUMTs in Sequenced Eukaryotic Genomes. *Molecular Biology and Evolution* **21**, 1081–1084 (2004).
54. 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).
55. Buhay, J. E. “COI-like” Sequences Are Becoming Problematic in Molecular Systematic and DNA Barcoding Studies. *Journal of Crustacean Biology* **29**, 96–110 (2009).
56. 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).

57. 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).
58. 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).
59. Ermakov, O. A. *et al.* Implications of Hybridization, NUMTs, and Overlooked Diversity for DNA Barcoding of Eurasian Ground Squirrels. *PLOS ONE* **10**, e0117201 (2015).
60. Ahmed, N. S. & Jaffar Ali, H. A. Numts: an impediment to DNA barcoding of Polyclinids, Tunicata. *Mitochondrial DNA Part A* **27**, 3395–3398 (2016).
61. 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).
62. 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).
63. Schultz, J. A. & Hebert, P. D. N. Do pseudogenes pose a problem for metabarcoding marine animal communities? *Molecular Ecology Resources* **22**, 2897–2914 (2022).
64. 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 work 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; 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.