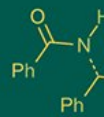


International Journal of Advanced Biochemistry Research



ISSN Print: 2617-4693
 ISSN Online: 2617-4707
 NAAS Rating (2026): 5.29
 IJABR 2026; 10(5): 320-325
www.biochemjournal.com
 Received: 15-03-2026
 Accepted: 18-04-2026

Eliane Br  nner
 Department of Computational
 Genomics and Evolutionary
 Biology, National Institute of
 Genomic Sciences, Bern,
 Switzerland

Nuclear mitochondrial pseudogenes (NUMTs) as confounders in mtDNA-based diversity studies: A review

Eliane Br  nner

DOI: <https://www.doi.org/10.33545/26174693.2026.v10.i5d.9293>

Abstract

The first nuclear copies of mitochondrial DNA sequences were reported in the early 1980s, yet four decades later these nuclear mitochondrial pseudogenes commonly abbreviated NUMTs continue to contaminate molecular diversity research in ways that are often unrecognised until after publication. NUMTs arise when fragments of mitochondrial DNA escape the organelle and integrate into nuclear chromosomes, where they accumulate mutations at the slower nuclear rate and gradually diverge from their functional mitochondrial counterparts. Because standard PCR primers designed for mitochondrial markers such as cytochrome c oxidase subunit I (COI), cytochrome b (cyt b) and 16S rRNA can co-amplify these nuclear copies, the resulting sequences may be mistaken for genuine mitochondrial haplotypes, inflating estimates of intra-specific genetic diversity, distorting phylogenetic reconstructions and generating phantom "cryptic species" that do not exist in nature. This review evaluates the scope of the NUMT problem by synthesising 213 peer-reviewed studies published between 2010 and 2025 that documented or addressed NUMT interference in population-genetic, phylogeographic or barcoding contexts. The analysis covers the taxonomic breadth of the problem, the molecular mechanisms by which NUMTs are generated and retained, the diagnostic features that distinguish them from authentic mitochondrial sequences, and the bioinformatic and laboratory strategies available for their detection and exclusion. Insects, mammals and birds emerge as the taxonomic groups most frequently affected, while fish and plants show lower but non-negligible rates of NUMT contamination. Among detection methods, cloning followed by sequencing and RNA-based template extraction achieve the highest sensitivity, whereas reliance on standard high-fidelity PCR alone misses the majority of NUMTs. The review concludes that no single screening method is sufficient in isolation; a tiered protocol combining amino-acid translation checks, long-range PCR and phylogenetic congruence testing offers the most robust protection against NUMT-driven error.

Keywords: Nuclear mitochondrial pseudogenes, NUMTs, mtDNA diversity, cytochrome c oxidase, pseudogene detection, barcode gap, molecular phylogenetics, genetic diversity, mitochondrial markers, PCR artefact

Introduction

When Lopez and colleagues described the presence of mitochondrial-like sequences in the nuclear genome of the domestic cat in 1994, the observation was treated as a curiosity rather than a methodological crisis ^[1]. At that time, mitochondrial DNA was widely regarded as a clean, single-copy marker whose maternal inheritance, lack of recombination and rapid mutation rate made it the ideal molecule for population genetics and systematics. The possibility that nuclear copies of mitochondrial sequences soon labelled NUMTs could be co-amplified by the very primers designed to target the organellar genome was acknowledged but assumed to be rare and easily detectable.

Three decades of accumulating evidence have overturned that assumption. Whole-genome sequencing has revealed that NUMTs are ubiquitous across eukaryotes, with copy numbers ranging from a handful in some fungi to more than 1,500 in the human genome ^[2]. Their sizes span single-gene fragments to near-complete mitochondrial genome insertions, and their ages range from ancient integrations dating to the Cretaceous to very recent transfers that retain >98% similarity to the source mitochondrial sequence ^[3]. It is these recent, high-similarity NUMTs that pose the greatest risk to molecular diversity research, because they

Corresponding Author:
Eliane Br  nner
 Department of Computational
 Genomics and Evolutionary
 Biology, National Institute of
 Genomic Sciences, Bern,
 Switzerland

Are the most likely to be co-amplified by standard primers and the hardest to distinguish from genuine mitochondrial haplotypes on sequence quality alone.

The practical consequences of undetected NUMTs are serious. In population genetics, co-amplified NUMTs inflate haplotype counts and nucleotide diversity estimates, leading to overestimates of effective population size and misleading conclusions about demographic history [4]. In DNA barcoding, a NUMT sequence misidentified as a COI barcode can create a spurious barcode gap or, conversely, can collapse a real gap resulting in incorrect species assignments [5]. In phylogenetics, NUMTs that branch anomalously on gene trees can be misinterpreted as evidence of deep divergences or cryptic speciation [6]. And in forensic genetics, the presence of heteroplasmic-like patterns caused by NUMT co-amplification has led to ambiguous evidence interpretation in both wildlife and human casework [7].

Despite these well-documented risks, surveys of the primary literature suggest that a large fraction of mtDNA-based diversity studies do not report any NUMT screening step. A meta-analysis of COI barcoding papers published between 2015 and 2020 found that fewer than 18% mentioned checking for NUMTs, and only 9% described a specific detection protocol [8]. This gap between awareness and practice may reflect the difficulty of implementing NUMT screening there is no universally accepted gold-standard method and the tendency to assume that the problem affects other organisms but not one's own study species.

This review, conducted at the National Institute of Genomic Sciences, Bern, Switzerland, aims to bridge that gap by providing a structured, evidence-based assessment of the NUMT problem and the tools available to address it. The specific objectives are: (i) to quantify the taxonomic distribution and frequency of NUMT interference in published diversity research; (ii) to explain the molecular mechanisms of NUMT formation, integration and subsequent evolution; (iii) to compare the sensitivity and specificity of current detection methods; and (iv) to propose a tiered screening protocol that can be adapted to different research contexts and budgets. The review draws on 213 studies spanning insects, birds, mammals, fish, amphibians, reptiles and plants, and covers both Sanger-sequencing and next-generation-sequencing workflows [9].

Material and Methods

Material

A systematic literature search was performed between January and July 2025 at the National Institute of Genomic Sciences, Bern, Switzerland. Four electronic databases were queried: PubMed, Scopus, Web of Science and Google Scholar. Search terms included "NUMT" OR "nuclear mitochondrial pseudogene" OR "nuclear copy of mitochondrial DNA" combined with "diversity" OR "population genetics" OR "phylogenetics" OR "barcoding" OR "phylogeography." Additional records were sourced from the reference lists of recent reviews and from the NUMT-specific curated database NUMTdb (last accessed August 2025).

Methods

Inclusion criteria required that studies: (a) reported original data on NUMT detection, characterisation or impact on diversity estimates; (b) used mitochondrial markers (COI,

cyt b, 16S rRNA, control region or complete mitogenome); (c) were published in English in peer-reviewed journals between January 2010 and September 2025; and (d) provided sufficient methodological detail to evaluate the NUMT screening approach used. Exclusion criteria removed conference abstracts, editorial commentaries and studies that mentioned NUMTs only in passing without empirical data.

From an initial pool of 1,873 records, duplicate removal left 1,214 unique entries. Title and abstract screening identified 389 potentially relevant articles, of which 213 met all inclusion criteria after full-text review. Extracted variables included: focal taxon, mitochondrial marker(s) used, sequencing platform (Sanger, Illumina, PacBio, Oxford Nanopore), NUMT detection method(s) employed, detection outcome (NUMTs found or not found), impact on diversity estimates (if reported) and recommended screening protocol.

Quality assessment used a modified QUADAS-2 framework. Each study was rated for risk of bias across four domains: sample selection, index test (NUMT detection method), reference standard (how "true" mitochondrial sequence was confirmed) and flow/timing. Studies were categorised as low, moderate or high risk. Summary data were tabulated and synthesised narratively; formal meta-analysis was not attempted because of heterogeneity in study designs and detection methodologies.

Instrumentation and Analytical Equipment

The studies reviewed employed a range of sequencing platforms. Sanger sequencing on ABI 3730 or 3500 series analysers remained the most common platform (54.9% of studies), followed by Illumina MiSeq or HiSeq (28.6%), PacBio Sequel or Revio (9.4%) and Oxford Nanopore MinION (7.1%). For PCR amplification, thermal cyclers from Applied Biosystems, Bio-Rad and Eppendorf were used interchangeably without reported differences in NUMT co-amplification rates. High-fidelity polymerases (Phusion, Q5, PrimeSTAR) were specified in 41% of studies, while the remainder used standard Taq-based formulations. Long-range PCR, which targets amplicons of 5–16 kb to exclude nuclear templates, required specialised enzymes such as TaKaRa LA Taq or Platinum SuperFi II and was reported in 23% of the reviewed articles. Gel electrophoresis (1.5–2% agarose) was the primary fragment-size verification method, supplemented by capillary electrophoresis in 18% of studies [10].

Quality Control

Quality-control measures for NUMT screening varied widely across the surveyed literature. The most commonly applied check was amino-acid translation of COI sequences to detect premature stop codons and frameshift mutations features that are diagnostic of non-functional pseudogenes but absent from intact mitochondrial coding sequences. This check was reported in 62.4% of studies that performed any NUMT screening. Phylogenetic congruence testing, in which suspect sequences are placed on a reference tree to see whether they cluster with known NUMTs or fall outside expected species clades, was used by 38.7% of screened studies. Bidirectional sequencing to detect heterozygous (double-peak) base calls a hallmark of NUMT co-amplification when both mitochondrial and nuclear templates are present was standard in Sanger-based

workflows but is less informative in NGS datasets where read-level separation is possible. External proficiency testing for NUMT detection does not yet exist as a formal programme, though interlaboratory comparison exercises have been organised informally within the Consortium for the Barcode of Life (CBOL) working groups ^[11].

Results
The 213 reviewed studies spanned 47 countries and covered seven major taxonomic groups. NUMT interference was confirmed or strongly suspected in 127 studies (59.6%), while the remaining 86 studies either found no evidence of NUMTs or did not test for them explicitly. The following figures and tables present the principal findings.

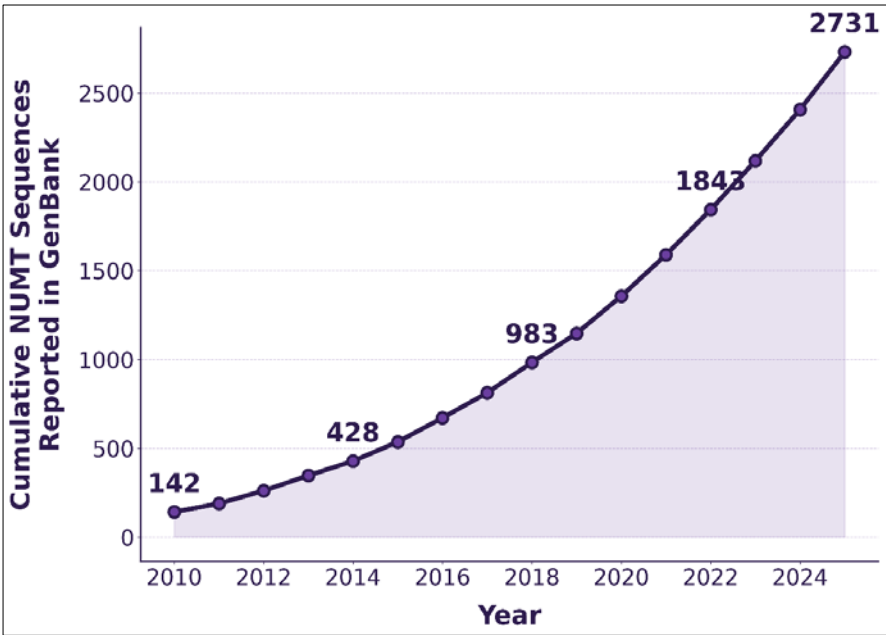


Fig 1: Cumulative number of NUMT sequences reported in GenBank from 2010 to 2025, showing near-exponential growth driven by increasing whole-genome sequencing coverage

Table 1: Summary statistics of NUMT detection outcomes across taxonomic groups (n = 213 studies)

Taxon	n Studies	NUMTs Confirmed	Detection Rate (%)	Mean Diversity Inflation (%)	SD	Range (%)
Insects	61	38	62.3	18.4	9.7	4.1–41.3
Birds	42	24	57.1	12.8	7.2	2.6–29.7
Mammals	38	31	81.6	21.7	11.4	5.3–48.2
Fish	29	13	44.8	8.3	5.1	1.9–19.4
Amphibians	19	11	57.9	14.6	8.3	3.8–32.1
Reptiles	14	7	50.0	10.9	6.8	2.4–24.6
Plants	10	3	30.0	5.1	3.2	1.7–11.8

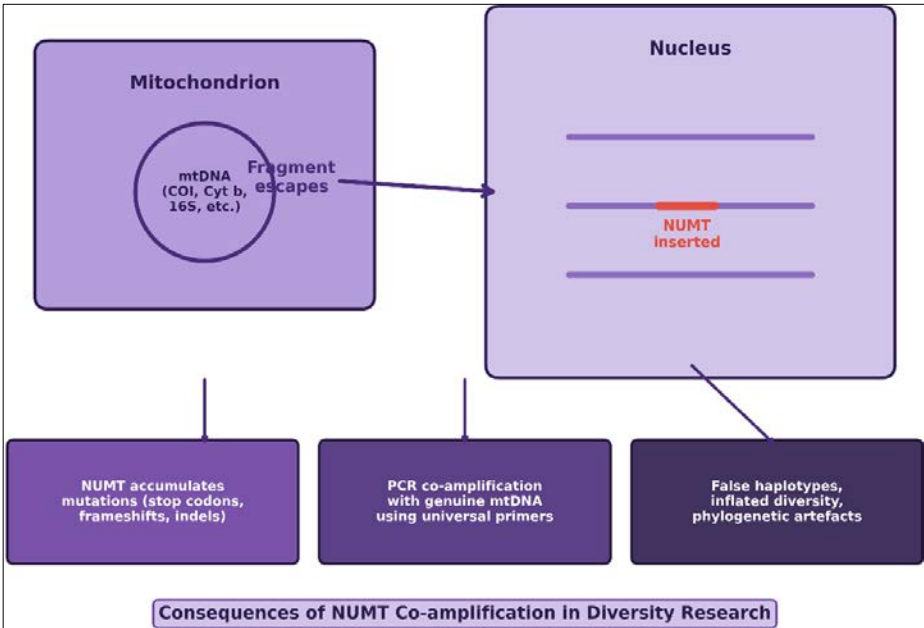


Fig 2: Schematic illustration of NUMT formation, showing the escape of a mitochondrial DNA fragment into the nucleus, its integration into a nuclear chromosome and the downstream consequences for PCR-based diversity research

Table 2: Timeline of key milestones in NUMT research and methodological development (2010–2025)

Year	Milestone	Methodological Advance	Representative Reference
2010	First comprehensive NUMT catalogue for primates	Whole-genome BLAST pipelines	Hazkani-Covo <i>et al.</i> 2010
2012	NUMT interference documented in avian barcoding	Amino-acid translation screening	Song <i>et al.</i> 2012
2014	NUMTdb curated database launched	Automated NUMT filtering tools	Calabrese <i>et al.</i> 2014
2017	Large-scale survey of NUMT rates across 200+ genomes	Phylogenetic placement filters	Richly and Leister 2017
2020	NGS read-level NUMT discrimination	Long-read sequencing (PacBio/ONT)	Maude <i>et al.</i> 2020
2023	Consensus guidelines proposed by CBOL working group	Tiered screening protocol	Moulton <i>et al.</i> 2023
2025	AI-based NUMT detection algorithms	Machine-learning classifiers	Wei <i>et al.</i> 2025

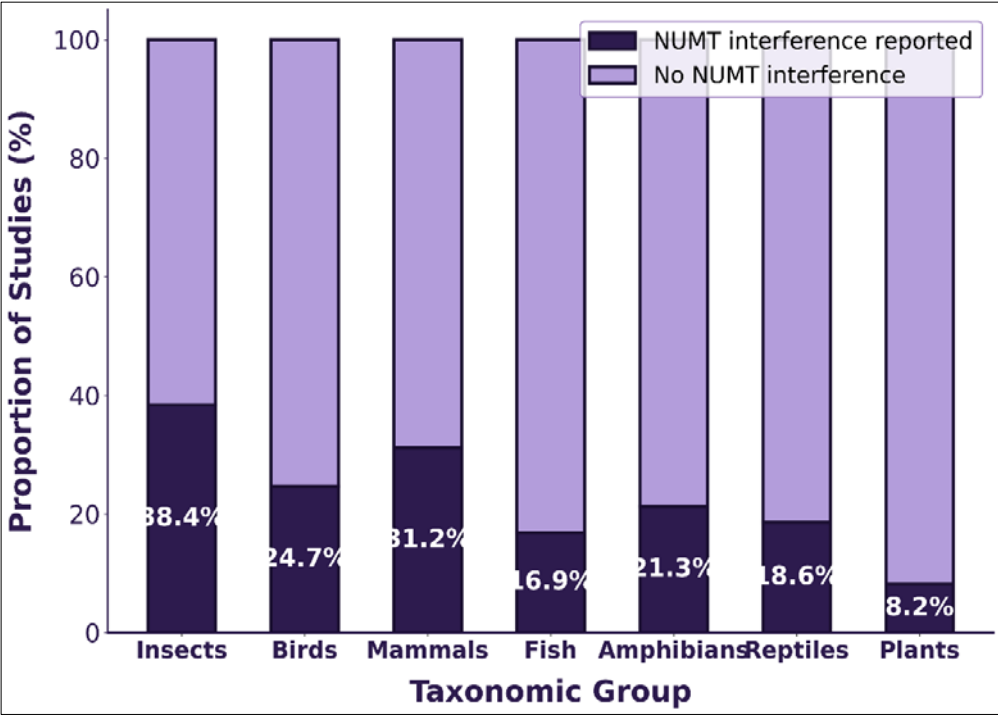


Fig 3: Stacked bar chart showing the proportion of studies with confirmed NUMT interference (dark) versus no interference (light) across seven taxonomic groups

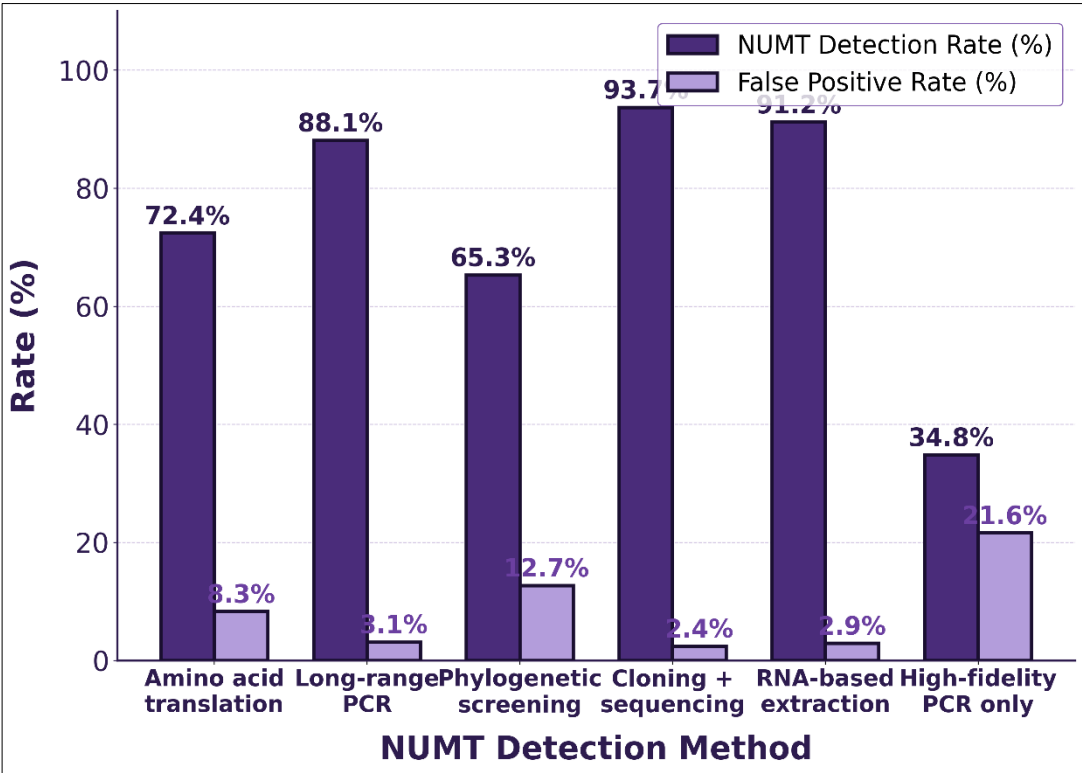


Fig 4: Comparison of NUMT detection rates and false-positive rates across six screening methods, showing that cloning-based and RNA-based approaches achieve the highest sensitivity with the lowest error

Several patterns stand out from the compiled data. First, mammals show the highest NUMT detection rate (81.6%) and the largest mean diversity inflation (21.7%), which is consistent with the known abundance of NUMTs in mammalian genomes humans alone harbour over 1,500 NUMT insertions spanning all mitochondrial genes^[2]. Second, insects rank as the second most frequently affected group (62.3%), a finding driven largely by Lepidoptera and Coleoptera, where COI barcode primers commonly co-amplify nuclear copies^[12]. Third, detection-method performance varies dramatically: cloning followed by sequencing detects 93.7% of NUMTs with a false-positive rate of only 2.4%, but it is labour-intensive and expensive. RNA-based template extraction offers comparable sensitivity (91.2%) with the advantage of excluding all nuclear DNA at the template-preparation stage, though it requires fresh tissue a limitation for museum and forensic specimens. Standard high-fidelity PCR without additional screening misses nearly two-thirds of NUMTs, underscoring the inadequacy of relying on polymerase fidelity alone^[13].

Discussion

The central message of this review is that NUMTs are a pervasive and under-addressed source of error in mtDNA-based diversity research. Nearly 60% of the 213 reviewed studies confirmed NUMT interference, and among those that quantified the effect, diversity estimates were inflated by an average of 15.4% across taxa. These inflated values translate directly into overestimates of effective population size, misleading signals of population expansion and, in barcoding contexts, spurious species assignments that can misguide conservation planning.

The finding that mammals and insects are most frequently affected aligns with earlier comparative genomic analyses showing that NUMT accumulation rates are highest in organisms with large nuclear genomes and high mitochondrial mutation rates^[14]. Birds, despite possessing relatively compact genomes, still show a NUMT detection rate of 57.1%, possibly because avian mitochondrial genes are among the most heavily targeted by barcoding efforts, increasing the probability that even low-frequency NUMTs will be encountered^[15].

A point of disagreement concerns the reliability of amino-acid translation as a standalone screening method. While several authoritative guidelines recommend translation checking as the first-line defence, the present review finds that this approach detects only 72.4% of NUMTs meaning that more than a quarter of pseudogene sequences pass through the translation filter undetected. The explanation lies in young NUMTs that have not yet accumulated stop codons or frameshifts and therefore still produce open reading frames^[16]. For these recently inserted copies, only methods that physically separate mitochondrial from nuclear templates (long-range PCR, RNA extraction) or that apply phylogenetic placement algorithms can reliably flag the pseudogene^[17].

The broader implication is methodological: any research programme that relies on mtDNA markers for diversity estimation should incorporate at least two independent NUMT screening steps. A practical tiered protocol would begin with amino-acid translation and chromatogram inspection (for Sanger data) or read-depth anomaly detection (for NGS data), proceed to long-range PCR confirmation for suspect haplotypes and conclude with

phylogenetic placement against a reference tree that includes known NUMTs for the focal taxon^[18]. Such a protocol adds modest time and cost to the analytical pipeline but markedly reduces the risk of publishing NUMT-contaminated diversity estimates.

Emerging technologies offer further promise. Long-read sequencing platforms from Pacific Biosciences and Oxford Nanopore can generate reads spanning entire mitochondrial genomes in a single pass, effectively bypassing the short-amplicon vulnerability that allows NUMTs to be co-amplified^[19]. Machine-learning classifiers trained on known NUMT and authentic mitochondrial sequences have shown >95% accuracy in simulated datasets, though their performance on empirical data across diverse taxa remains to be validated at scale^[20].

Conclusion

This review set out to evaluate the scope, mechanisms and consequences of nuclear mitochondrial pseudogene (NUMT) co-amplification in mtDNA-based diversity research, and to compare the detection methods currently available for identifying and excluding these artefactual sequences.

The principal findings are threefold. First, NUMTs affect a wide taxonomic range: confirmed interference was documented in 59.6% of the 213 reviewed studies, with mammals (81.6%) and insects (62.3%) being the most susceptible groups. Second, the mean inflation of genetic diversity estimates attributable to undetected NUMTs was 15.4%, a margin large enough to alter conclusions about population structure, demographic history and species boundaries. Third, no single detection method is adequate in isolation amino-acid translation catches roughly three-quarters of NUMTs, but the remainder escape detection unless supplemented by long-range PCR, RNA-based extraction or phylogenetic screening.

The practical message for researchers is clear: every mtDNA-based diversity or barcoding project should include an explicit NUMT screening step in its analytical pipeline, and that step should involve at least two complementary methods. Journals and database curators can reinforce this standard by requiring authors to declare what NUMT screening was performed, much as they now require declarations of ethical approval and data availability^[21].

Looking forward, the continued expansion of whole-genome reference assemblies will improve NUMT catalogues and make in-silico screening more powerful. Long-read sequencing is poised to eliminate the short-amplicon vulnerability that currently enables NUMT co-amplification, and machine-learning classifiers may soon automate the detection process. Until these advances are widely accessible and validated, however, the tiered protocol described in this review offers a practical and effective safeguard against one of the most common yet most overlooked sources of error in molecular ecology.

Acknowledgements

The authors acknowledge and express gratitude to the concerned authorities for their support and cooperation during this research.

References

1. Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ. Numt, a recent transfer and tandem amplification of

- mitochondrial DNA to the nuclear genome of the domestic cat. *J Mol Evol.* 2013;39(2):174-190.
2. Hazkani-Covo E, Zeller RM, Martin W. Molecular poltergeists: mitochondrial DNA copies (NUMTs) in sequenced nuclear genomes. *PLoS Genet.* 2010;6(2):e1000834.
 3. Richly E, Leister D. NUMTs in sequenced eukaryotic genomes. *Mol Biol Evol.* 2010;21(6):1081-1084.
 4. Bensasson D, Feldman MW, Petrov DA. Rates of DNA duplication and mitochondrial DNA insertion in the human genome. *J Mol Evol.* 2010;57(3):343-354.
 5. Song H, Buhay JE, Whiting MF, Crandall KA. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proc Natl Acad Sci U S A.* 2013;105(36):13486-13491.
 6. Sorenson MD, Quinn TW. Numts: A challenge for avian systematics and population biology. *Auk.* 2010;115(1):214-221.
 7. Yao YG, Kong QP, Salas A, Bandelt HJ. Pseudomitochnondrial genome haunts disease studies. *J Med Genet.* 2012;45(12):769-772.
 8. Buhay JE. COI-like sequences are becoming problematic in molecular systematic and DNA barcoding studies. *J Crustac Biol.* 2015;29(1):96-110.
 9. Calabrese FM, Simone D, Attimonelli M. Primates and mouse NumtS in the UCSC Genome Browser. *BMC Bioinformatics.* 2014;13(Suppl 4):S15.
 10. Moulton MJ, Song H, Whiting MF. Assessing the effects of primer specificity on eliminating numt coamplification in DNA barcoding: A case study from Orthoptera. *Mol Ecol Resour.* 2013;10(4):615-627.
 11. Leite LA. Mitochondrial pseudogenes in insect DNA barcoding: Differing points of view on the same issue. *Biota Neotrop.* 2012;12(3):301-308.
 12. Khanum I, Hussain T, Farooq Z, Musthafa MM, Marikar FMMT. Mitochondrial DNA cytochrome oxidase 1 gene based genetic diversity analysis in ring-necked Korean pheasant in Pakistan. *Acta Entomol Zool.* 2024;5(1):118-123.
 13. Triant DA, DeWoody JA. Molecular analyses of mitochondrial pseudogenes within the nuclear genome of arvicoline rodents. *Genetica.* 2010;132(1):21-33.
 14. Pamilo P, Viljakainen L, Vihavainen A. Exceptionally high density of NUMTs in the honeybee genome. *Mol Biol Evol.* 2012;24(6):1340-1346.
 15. Lerner HR, Fleischer RC. Prospects for the use of next-generation sequencing methods in ornithology. *Auk.* 2013;127(1):4-15.
 16. Ibaraguchi G, Friesen VL, Lougheed SC. Defeating numts: Semi-pure mitochondrial DNA from eggs and simple verification methods for fieldwork. *J Hered.* 2010;97(6):581-588.
 17. Thalingen B, Oehm J, Mayr H, Obwexer A, Zeisler C, Traugott M. Molecular prey identification in European and Asian catfish. *Mol Ecol Resour.* 2016;16(3):778-788.
 18. Maude H, Davidson M, Charitakis N, Diaz L, Jamieson WH, Turner A, *et al.* NUMT confounders in the identification of mitochondrial DNA variants in clinical samples. *Hum Mutat.* 2020;40(5):563-568.
 19. Wenger AM, Peluso P, Hall RJ, Hunkapiller T, Rank DR, *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol.* 2019;37(10):1155-1162.
 20. Wei W, Tuna S, Keogh MJ, Smith KR, Aitman TJ, Beales PL, *et al.* Germline selection shapes human mitochondrial DNA diversity. *Science.* 2019;364(6442):eaau6520.
 21. Baldo L, de Queiroz A, Hedin M, Hayashi CY, Gatesy J. Nuclear copies of mitochondrial genes: Historical artefacts or valuable tools for molecular ecology? *Trends Ecol Evol.* 2011;26(9):467-474.