Probability-based sequence comparison finds the oldest ever nuclear mitochondrial DNA segments in mammalian genomes

Muyao Huang[1[0009−0007−1413−8165]] and Martin C. Frith[1[0000−0003−0998−2859]]⋆

The Department of Computational Biology and Medical Sciences, The Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan

Abstract. The insertion of mitochondrial genome-derived DNA sequences into the nuclear genome is a frequent event in organismal evolution, resulting in nuclear-mitochondrial DNA segments (NUMTs), which serve as a significant driving force for genome evolution. Once incorporated into the nuclear genome, some NUMTs can be conserved for extended periods, adapting to perform novel cellular functions. However, current mainstream methods for detecting NUMTs are inefficient at identifying ancient and highly degraded NUMTs, leading to their prevalence and impact being underestimated. These ancient NUMTs likely play a far greater role in genetic functions than previously recognized, including contributing to the acquisition of functional exons. This study focuses on identifying ancient NUMTs in mammalian genomes using enhanced high-sensitivity sequence comparison methods. A sensitive and accurate NUMT-searching pipeline was established, predicting 1,013 NUMTs in the human reference genome, 398 (39%) of which are newly detected compared to the UCSC reference human NUMTs database. Notably, 93 pre-Eutherian human NUMTs were identified, representing significantly older NUMTs than previously reported, with origins dating back at least 100 million years. The most ancient mammalian NUMTs could even date back over 160 million years, inserted into the nuclear genome of the common ancestor of therian mammals. This study provides a comprehensive exploration of the quantity and evolutionary history of mammalian NUMTs, paving the way for future research on endosymbiotic impact on the evolution of nuclear genomes.

Keywords: Nuclear-mitochondrial DNA segments (NUMTs) · Genome evolution · Mammals.

1 Introduction

The insertion of mitochondrial DNA fragments into the nucleus generates Nuclear-mitochondrial DNA segments (NUMTs), a frequent process in eukaryotes that

⋆ Corresponding author: mcfrith@edu.k.u-tokyo.ac.jp

influences nuclear genome evolution. However, the mechanisms behind this process are largely unexplored. Recent advances in sequencing have enabled comprehensive studies of NUMTs in animals and plants [2,10,20,14]. Once inserted, NUMTs often lose their original mitochondrial functions due to the lack of evolutionary constraints [20]. These various fates complicate the identification of NUMTs and their functions, making it challenging to determine their roles as adaptive elements of the nuclear genome.

NUMTs were once considered non-functional and harmless, mainly contributing to non-coding sequences and pseudogenes. However, recent studies have revealed more significant impacts, such as NUMTs disrupting gene expression or, in some cases, causing diseases [1,16]. NUMTs can also be recruited as novel exons, providing new coding sequences to existing nuclear genes [11]. This suggests that NUMTs may contribute to the acquisition of functional exons, which was previously hard to detect due to the limitations of traditional sequence comparisons.

Currently, the primary approaches for identifying NUMTs include wet lab-based and computational methods [18]. Computational approaches can be further classified based on the target of study: (1) the identification of common NUMTs, which represent ancestral insertions present in the reference genome, and (2) the detection of rare polymorphic NUMTs specific to individuals. In recent years, advancements in next-generation sequencing technologies have increased the focus on recent NUMT insertion events and polymorphic NUMTs, spurring the development of improved computational methods for their identification. In contrast, little progress has been made in annotating ancestral NUMTs in the reference genome. Databases of common human NUMTs remain outdated, with annotations still based on hg19 [13]. Comprehensive annotation of ancestral NUMTs, however, remains crucial for two key reasons:

1. Accurate annotation is essential for detecting polymorphic NUMTs, as it enables researchers to identify annotated ancestral NUMTs in next-generation sequencing data and exclude them from further analyses.
2. It advances studies on the evolutionary history of NUMTs, as previous annotations often overlook ancient NUMTs due to limitations in detection methods.

In previous studies, the main method for identifying NUMTs in the reference genome is by comparing the whole mitochondrial genome to the nuclear genome, using local sequence alignment tools [13,15]. However, this approach may fail to detect ancient and highly degenerated NUMTs due to the less sensitive nature of DNA:DNA comparison compared to Protein:DNA comparison, which uses protein similarity scoring matrices [12]. On the other hand, though protein sequence comparisons are more sensitive than DNA-level comparisons when they are applied to NUMTs identification, they will miss the alignments related to non-protein-coding sequences in the mitochondrial genome [14]. Therefore, a combined approach utilizing both protein-to-DNA and DNA-to-DNA comparisons may offer a more accurate and sensitive method for identifying NUMTs,

but further improvements are needed to enhance NUMTs detection, especially for ancient insertions.

In this study, we propose an optimized and accurate NUMTs-searching pipeline based on LAST combining Protein:DNA and DNA:DNA matching methods. The primary reasons for choosing LAST to identify NUMTs can be summarized in two points:

1. Compared to the traditional alignment tools that rely on fixed, pre-determined scoring matrices, LAST uses a standard iterative approach to derive scoring matrices tailored to the unique characteristics of the input sequences [5]. This adaptability ensures that the alignment process accounts for sequence-specific patterns and evolutionary constraints, thereby enhancing the accuracy of NUMT detection.
2. In LAST, the Protein:DNA matching allows frameshift within matches and uses a 64x21 substitution matrix rather than a standard 20x20 matrix to improve the sensitivity [19]. This improvement enables LAST to identify more subtle similarities, enhancing its sensitivity for detecting NUMTs, particularly those that have undergone significant sequence divergence.

Our approach is well-suited for detecting highly degenerated NUMTs. By using the optimized NUMTs-searching pipeline, we completed the annotation of NUMTs in the human reference genome hg38, identified NUMTs in 15 additional available reference genomes, and traced ancient NUMTs across mammalian clades, providing estimates of their insertion times.

2 Materials and Methods

2.1 Data collection

16 annotated reference genomes, including mitochondrial and nuclear sequences from various mammalian orders, were downloaded from NCBI. The species analyzed represent 9 orders: Artiodactyla (2 species), Carnivora (3 species), Lagomorpha (1 species), Marsupial (2 species), Monotremata (2 species), Pilosa (1 species), Primates (3 species), Proboscidea (1 species), and Rodentia (1 species). Each genome was assembled to chromosome-level quality in its latest version (May 2024). Unplaced scaffolds were excluded from this study to ensure accurate NUMT identification. 13 mitochondrial proteins for each species were downloaded separately from NCBI.

2.2 Optimized NUMTs-searching pipeline

NUMTs-searching workflow NUMTs were detected using a novel NUMTs-searching pipeline, which combined nuclear genome-mitochondrial genome and nuclear genome-mitochondrial protein comparisons based on LAST version 1607 [9]. In the nuclear (NU) genome-mitochondrial (MT) genome comparison, the nuclear genome sequence is set as the query, and the mitochondrial genome is

set as the reference. LAST makes efforts to find alignments for every coordinate in the nuclear genome by pre-learning the substitution and gap rates between the mitochondrial genome and the nuclear genome using `last-train` [5] with the following commands:

```
lastdb -P8 -c mitogenodb $mitogenoFASTA

last-train -S0 --pid=70 --sample-number=0 -P8
mitogenodb $nuclearFASTA > nu2mitogeno.train

lastal -H1 -J1 -R00 -P8 -p nu2mitogeno.train
mitogenodb $nuclearFASTA > nu2mitogeno.maf
```

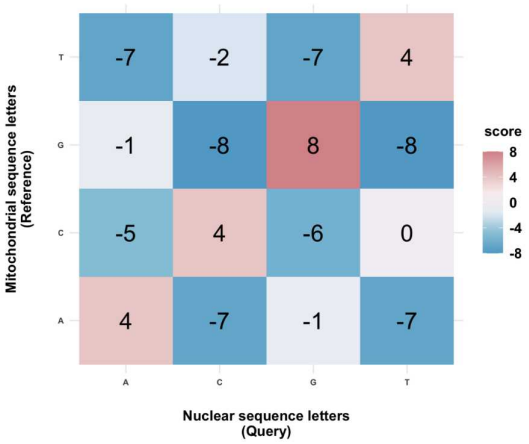
LAST commands in the nuclear (NU) genome-mitochondrial (MT) protein comparison are shown below:

```
lastdb -P8 -q -c mitoprodb $mitoproFASTA

last-train --codon --pid=70 --sample-number=0 -P8
mitoprodb $nuclearFASTA > nu2mitopro.train

lastal -H1 -K1 -m500 -P8 -p nu2mitopro.train
mitoprodb $nuclearFASTA > nu2mitopro.maf
```

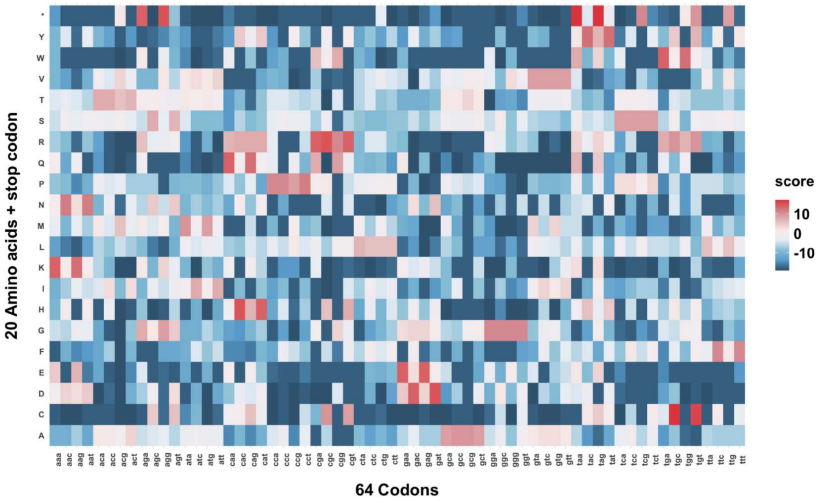
A sample scoring matrix trained on alignments between the human NU genome and MT genome is shown in Figure 1a. And the scoring matrix used in human NU genome:MT protein is shown in Figure 1b.

Several options were used in the LAST commands to improve sensitivity and ensure accuracy in detecting ancient (low-identity) NUMTs. For example, in DNA:DNA comparison, the `-c` option, combined with `-R00`, soft-masks repetitive elements from mitochondrial sequences while retaining them in nuclear sequences during the initial matches. It helps prevent false homologous predictions while still allowing the detection of NUMTs duplicates in the nuclear genome. The `-J1` option uses a new approach to better detect subtly related sequences by summing the probabilities of all alternative alignments between them [3]. `-P8` enables parallel processing with 8 threads to enhance computational speed. In protein:DNA comparison, the `-K1` option means that for each region of the nuclear genome, only the alignment with the highest score will be output. The `-m500` option makes the search slower but more sensitive. It is noteworthy that the human mitochondrial genome comprises the heavy (H)-strand and light (L)-strand and data from NCBI provides the L-strand sequence of the mitochondrial genome. Our comparisons and learned scoring matrix are based on the L-strand mitochondrial sequence.

Subsequently, we repeated the searches in the NU genome:MT genome comparison and NU genome:MT protein comparison using a reversed query sequence as a negative control. True NUMTs cannot be found in the reverse test, so all hits that appear in it are random. The highest score obtained from the reverse



(a) Score matrix used in human NU genome:MT genome comparison



(b) Score matrix used in human NU genome:MT protein comparison

Fig. 1: Score matrices used in human NUMTs detection

test was used as a threshold to filter out alignments with scores lower than it from the original search.

As an important novel step, alignments overlapping with nuclear ribosomal RNA (rRNA) were also excluded because LAST's high sensitivity can produce alignments between the 16S rRNA mitochondrial gene and its nuclear homologous sequences (Figure 2). Such alignments could lead to overestimating ancient NUMT counts, as some mitochondrial sequences might be mistakenly aligned with conserved nuclear rRNA regions.

```
Query = NR_003287.4 Homo sapiens 28S ribosomal N5 (RNA28SN5) Length: 5070bp

Subject = NC_012920.1 Homo sapiens mitochondrion Length: 16569bp

 Score = 53.8 bits (33), Expect = 1.1e-08
 Identities = 98/156 (62%), Gaps = 1/156 (0%)
 Strand = Plus / Plus

Query: 3774 AAGGTAGCCAAATGCCTCGTCATCTAATTAGTGACGCGCATGAATGGATGAACGAGATTC 3833
            ||||||||  |||  || ||    ||| ||| |||  | |||||||| |  |||||  |
Sbjct: 2590 AAGGTAGCATAATCACTTGTTCCTTAAATAGGGACCTGTATGAATGGCTCCACGAGGGTT 2649

Query: 3834 CCACTGTCCCTACCTACTATCCAGCGAAA-CCACAGCCAAGGGAACGGGCTTGGCGGAAT 3892
            |  ||||| ||  ||  || |||| ||||   ||   |  | |||  |||  |    |
Sbjct: 2650 CAGCTGTCTCTTACTTTTAACCAGTGAAATTGACCTGCCCGTGAAGAGGCGGGCATAACA 2709

Query: 3893 CAGCGGGGAAAGAAGACCCTGTTGAGCTTGACTCTA 3928
            ||||  |   |||||||||| | |||||| | | ||
Sbjct: 2710 CAGCAAGACGAGAAGACCCTATGGAGCTTTAATTTA 2745


Subject = NC_012920.1 Homo sapiens mitochondrion Length: 16569bp

 Score = 49 bits (30), Expect = 2.9e-07
 Identities = 104/178 (58%)
 Strand = Plus / Plus

Query: 4380 AAGTTACCACAGGGATAACTGGCTTGTGGCGGCCAAGCGTTCATAGCGACGTCGCTTTTT 4439
            ||||||||  ||||||||| |     |       | || || |||| | ||       |||
Sbjct: 2921 AAGTTACCCTAGGGATAACAGCGCAATCCTATTCTAGAGTCCATATCAACAATAGGGTTT 2980

Query: 4440 GATCCTTCGATGTCGGCTCTTCCTATCATTGTGAAGCAGAATTCACCAAGCGTTGGATTG 4499
                | ||||||| || ||     |||    ||  ||||     |  ||  ||| | |||
Sbjct: 2981 ACGACCTCGATGTTGGATCAGGACATCCCGATGGTGCAGCCGCTATTAAAGGTTCGTTTG 3040

Query: 4500 TTCACCCACTAATAGGGAACGTGAGCTGGGTTTAGACCGTCGTGAGACAGGTTAGTTT 4557
            |||| | | |||      |||||| ||| ||| ||||||  || |  |||||  ||||
Sbjct: 3041 TTCAACGATTAAAGTCCTACGTGATCTGAGTTCAGACCGGAGTAATCCAGGTCGGTTT 3098
```

Fig. 2: The alignments between human mitochondrion and human nuclear rRNA gene RNA28SN5

The strand information in NU genome:MT protein comparison is not consistent with that in genome comparison. Therefore, aligning the strand orientation between the two sets of results is convenient. To achieve this, it is necessary to determine which mitochondrial DNA strand each mitochondrial protein originates from, by the following commands. `fix-protein-strand` checks strand for the mitochondrial protein in the mitochondrial genome and changes the strand of alignments in the NU genome:MT protein comparison (see https://github.com/Koumokuyou/NUMTs/ for details. The link is a UCSC track hub

for our NUMTs data that displays genomic data tracks and details of NUMTs-searching pipeline mentioned in this study).

```
lastdb -P8 -q -c mitoprodb $mitoproFASTA

last-train -P8 --codon mitoprodb $mitogenoFASTA > mtgeno2pro.train

lastal -P8 -D1e10 -p mtgeno2pro.train
mitoprodb $mitogenoFASTA > mtgeno2pro.maf

fix-protein-strand mtgeno2pro.maf
nu2mitopro_movrrna.bed > nu2mitopro_movrrna_fix.bed
```

Finally, the filtered results from NU genome:MT genome and NU genome:MT protein were merged to get the final NUMTs dataset using bedtools v2.29.1. If the two alignments respectively from the two comparisons are in the same strand and either overlap by at least 1bp or are bookended, they are merged into a single NUMT. Additionally, alignments shorter than 30 bp in the merged result were removed. This length threshold was empirically determined, based on previous studies [14,17], to ensure sequences long enough to yield statistically significant alignments.

Assembly of NUMT segments To facilitate the subsequent analysis in building the NUMT orthology relationships network, adjacent NUMTs were assembled into larger contiguous regions, referred to as "blocks," using bedtools. A "block" is defined as a genomic region composed of one or more NUMT segments that are closely spaced in the nuclear genome. Specifically, adjacent NUMTs were combined into a single NUMT block when the distance between them in the nuclear genome was less than 2000 bp, regardless of their strand orientation. This approach consolidates fragmented NUMT sequences into more comprehensive units, thereby simplifying the analysis of their evolutionary relationships.

2.3 Identification of ancient NUMTs in the mammalian genomes

120 pairwise genome alignments were conducted by LAST, ensuring that only orthologous segments were identified while excluding non-homologous insertions between any two species within the 16 species [4]. A NUMT from one species was considered ancient only if it overlapped with at least two genome-to-genome alignments between that species and others. Using this threshold, we collected all ancient NUMT blocks from 16 species to prepare repositories of ancient NUMT blocks in each species for subsequent analysis.

2.4 Analyzing NUMT orthology relationships between species

The approach to identifying orthologous NUMTs is similar to that of previous research by Hazkani-Covo et al. [7]. We analyzed NUMT orthology relationships between two species based on the constructed ancient block repositories

mentioned above. Suppose we have two species, A and B. We first determined preliminary orthologous NUMTs in species A and B by assessing whether they overlap with inter-genome alignments between these species by at least half their length. Next, the subsequent analysis depended on the type of relationship between them. There are three types of orthologous NUMTs identification scenarios:

1. **Type 1:** A NUMT block from species A overlaps with orthologous segments between species A and B and with at least one species B NUMT block within those orthologous segments. In this case, the blocks from species A and B are considered orthologous NUMTs. The number of orthologous NUMTs between A and B includes the number of blocks from species B that overlap with the block from species A.
2. **Type 2:** A NUMT block from species A overlaps with orthologous segments between species A and B, but no NUMT block from species B is present in these segments. This absence suggests the potential existence of a highly degenerated NUMTs block in species B that could not be detected using NU genome-MT genome or NU genome-MT protein comparisons. To ensure a thorough detection of ancient NUMT, the aligned region of the species A NUMT block in species B is added to the species B ancient NUMT repository. The species A NUMT block with this potential NUMT block in species B, is counted in the number of orthologous NUMT between A and B.
3. **Type 3:** This scenario is the reverse of Type 2. A NUMT block from species B overlaps with orthologous segments between species A and B, but no NUMT block from species A is present in these segments. Here, the aligned region of the species B NUMT block in species A is added to the species A ancient NUMT repository as a potential ancient NUMT block. The NUMT block from species B, along with this potential NUMT block in species A, is counted in the number of orthologous NUMTs between A and B.

3 Results

3.1 Newly detected human NUMTs

Using our optimized NUMTs-search pipeline, we predicted a total of 1,013 NUMTs in the human reference genome. The number of NUMTs identified per chromosome ranged from approximately 20 to 100 (Figure 3). To assess the accuracy of our results, We compared our dataset with the reference UCSC human NUMTs database. The coordinates of UCSC NUMTs were converted to the hg38 genome assembly using the UCSC LiftOver tool, with 750 out of 766 NUMTs successfully mapped. Remarkably, our method identified 734 (98%) of these 750 UCSC NUMTs, with an overlap rate exceeding 50% of the sequence length. This high recovery rate demonstrates the reliability of our approach in accurately detecting previously annotated NUMTs.

In addition to recovering nearly all NUMTs from the UCSC database, our pipeline identified 398 novel NUMTs, accounting for 39% of the total predictions. As a result of applying a strict global E-value threshold and implementing

a reverse test to eliminate potential random matches, the additional NUMTs predicted by our pipeline are strongly supported as true alignment rather than random hits. These newly discovered NUMTs were not reported in the UCSC database, indicating the sensitivity of our pipeline in detecting NUMTs that earlier approaches may have missed.

3.2 Annotation of human NUMTs in hg38

Our hg38 NUMTs dataset formatted in BED has been uploaded to UCSC genome browser, and adopted as their hg38 human NUMTs database. It should be pointed out that the reverse test was **omitted** during NUMTs identification in this uploaded dataset to display a broader set of potential NUMTs, albeit at a relatively lower confidence level. To enable visualization, the dataset was converted to bigBed format and implemented as the NuMTs track using a series of UCSC utilities (Figure 4).

This NuMTs track provides essential information for each NUMT in the human reference genome, designated as *hg38_NUMTs_#*, including genomic coordinates, strand orientation, and confidence scores. The arrow direction of each item corresponds to the strand on which the NUMT is aligned. The confidence score for each alignment is calculated as:

$$\text{Score} = \text{-}10 \times \log_{10}(\text{E-value}) \tag{1}$$

These scores reflect alignment confidence and are visually represented using varying shades of gray, where darker shades indicate stronger confidence. Alignments with score >=100 (i.e. E-value <= $1e^{-10}$) are displayed in black. When a NUMT results from the merging of multiple alignments, the score is taken as the highest among all contributing alignments.

3.3 Ancient NUMT insertions across mammalian clades

The optimized NUMTs-searching pipeline was applied to identify NUMTs in 15 other mammalian genomes. Information, including total NUMT counts, total NUMT length, and the NUMT coverage ratio in the nuclear genome, is summarized in Figure 5a. Compared to previous studies, our results reported a higher number of NUMTs or greater total NUMT length in most species, demonstrating the sensitivity and effectiveness of our method [6,17].

A stacked bar chart was used to examine the correlation between NUMT counts and matching methods (Figure 5b). The merged counts detected by both DNA-level and protein-level matches constituted the largest proportion (40%–70%) of total NUMT counts. Together with NUMTs detected only in the NU genome: MT genome comparison, these merged results represented the majority of the detected NUMTs. Although NUMTs identified only through the NU genome: MT protein comparison were relatively few, accounting for about 5% to 15% of the total, their presence underscores the advantage of incorporating protein-based matching methods. Further analysis of the human NUMTs

identified in our study that overlap with the UCSC database (665 out of 1013) revealed that 233 (35%) were detected only through DNA:DNA comparison, while 430 (65%) were detected in both DNA:DNA and Protein:DNA comparisons. Notably, only 2 NUMTs (0.3% of 665) were detected exclusively through Protein:DNA comparison, compared to 54 NUMTs (5% of 1013) in our complete set of human NUMTs. Our approach enables the identification of a broader spectrum of NUMTs, particularly ancient mitochondrial protein fossils that are challenging to detect using traditional DNA-based methods.

By comparing ancient NUMTs and orthologous segments in 120 pair-wise alignments, we established the NUMTs orthologous network among 16 species (Figure 6). The insertion time of these ancient NUMTs was inferred based on their presence at nodes within the phylogenetic tree of selected mammalian genomes, as established by Timetree.org. Since we mainly focus on the ancient NUMTs in genomes, the NUMTs orthology network is based on ancient NUMTs inserted at least in the nodes of common ancestors in their order. Most of the ancient NUMTs in mammals were inserted within the last 100 million years, with only a few appearing to have been integrated near the root of the mammalian phylogenetic tree.

A total of 193 ancient orthologous NUMTs were identified in the human genome, with 93 of them estimated to have been inserted at least 100 million years ago (Figure 7). To our knowledge, previous studies have not reported ancient NUMTs that predate the common ancestor of Boreoeutheria. The identification of such deeply conserved NUMTs significantly extends the timeline of NUMT insertion events that people can detect. We selected the 11 most conserved NUMTs shared by humans and over 10 species as samples, grouping them into 5 groups based on their presence in inter-genome alignments (Table 1). Strikingly, the most ancient NUMTs (shown in group 1) could date back at least 160 million years ago. It is worthwhile mentioning that this therian human NUMT is a potential NUMT from the aligned region in the human genome of a conserved brushtail possum NUMT, overlapping with 11 mammalian genomes. It may represent a highly degenerated NUMT that could not be directly detected by applying our NUMTs-searching pipeline to the human genome. However, by leveraging the aligned region from the conserved brushtail possum NUMT, we can infer its existence and regard it as a potential ancient human NUMT. Additional orthologous NUMTs searches in the wallaby and Tasmanian devil confirmed the conservation of this brushtail possum NUMT, suggesting with high confidence that it was inserted into the genome of the common ancestor of therian mammals. This NUMT originates from a mitochondrial tRNA region, supporting previous findings of slower evolutionary rates in mitochondrial tRNA-lookalikes compared to the adjacent numtDNA sequences [8]. Future studies are warranted to investigate whether ancient tRNA-derived NUMTs could potentially serve as a source of novel regulatory elements or influence genomic stability through long-term conservation and integration into nuclear genomes.

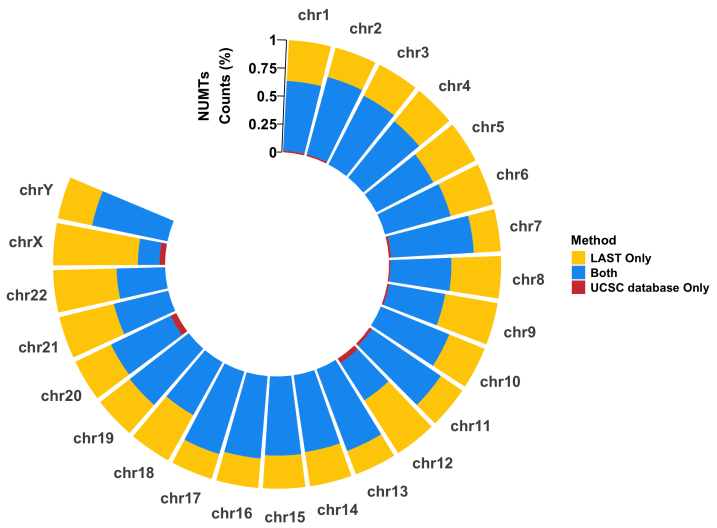


Fig. 3: Comparison of NUMTs Counts: Our pipeline vs. 2011 UCSC human reference NUMTs database

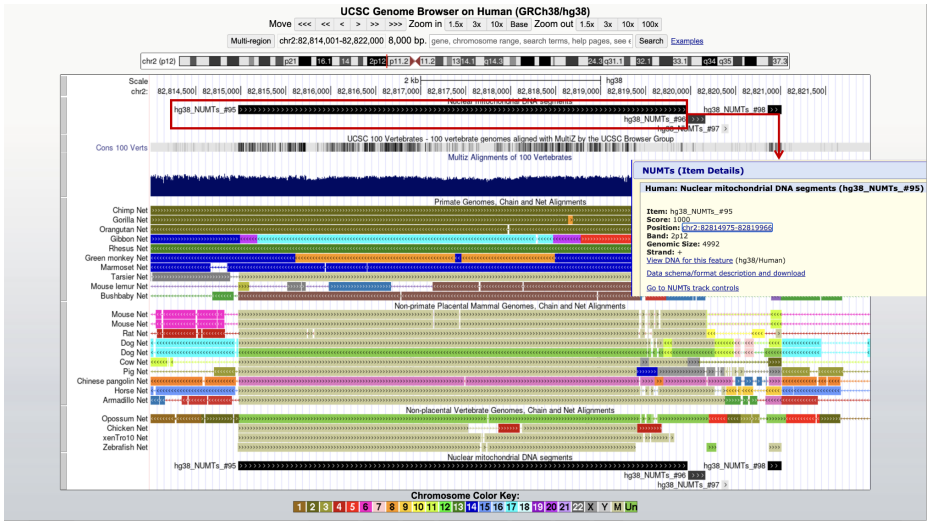


Fig. 4: Screenshot of hg38 human NUMTs track in the UCSC genome browser

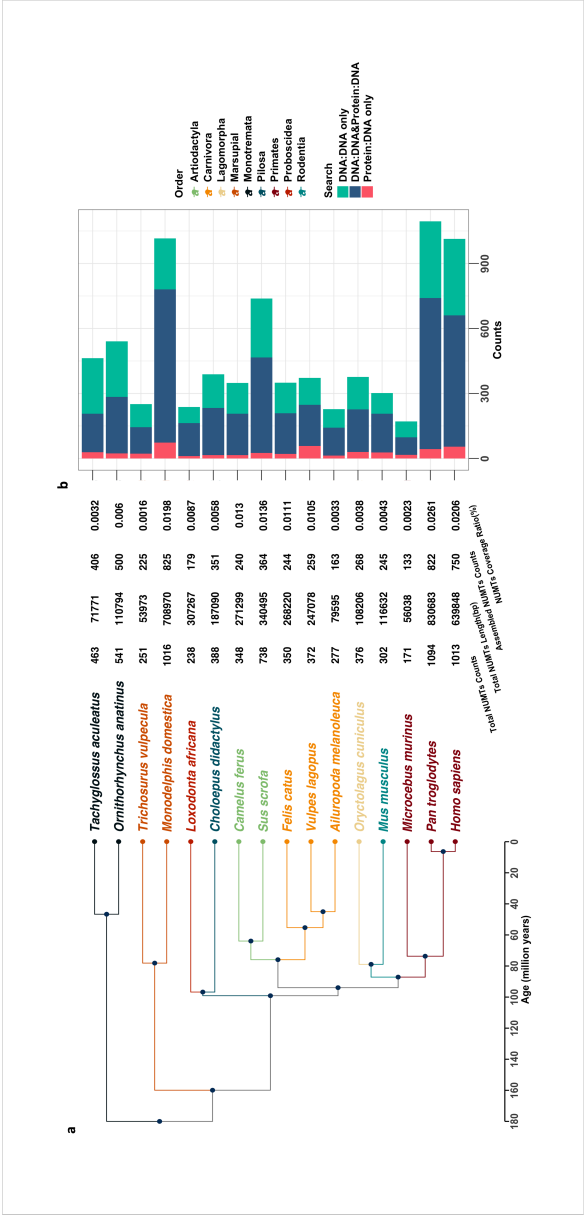


Fig. 5: Statistics of NUMTs in 16 mammalian genomes. a. From left to right, the columns show the total number of NUMTs, the total NUMTs length in base pairs, the number of assembled NUMTs, and the NUMTs length as a percentage of nuclear genome length. b. Results from different methods in total NUMTs counts.

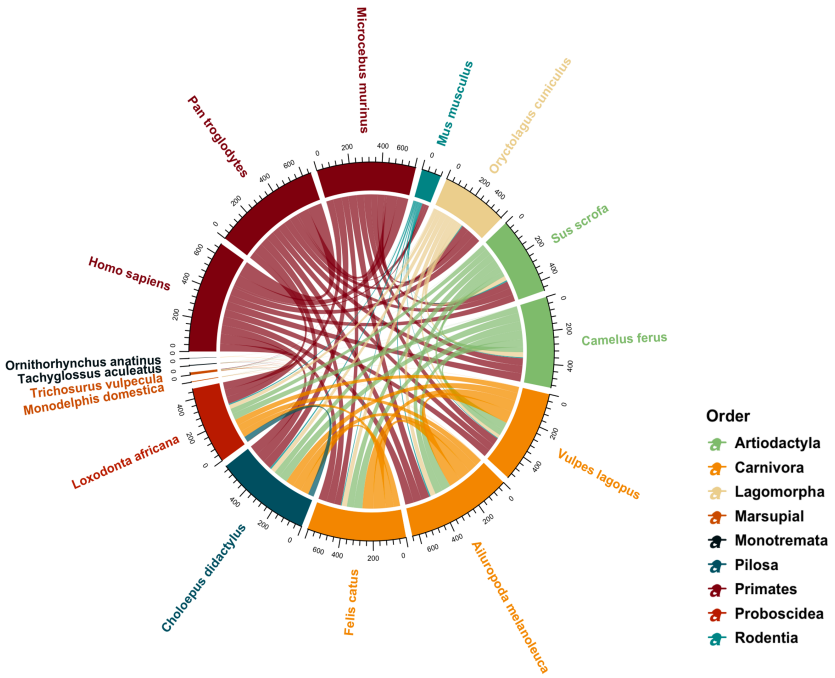


Fig. 6: Chord diagram illustrating the orthologous relationships of NUMTs among 16 species. Each chord represents the number of shared orthologous NUMTs between two species, with thicker chords indicating a higher number of shared NUMTs.

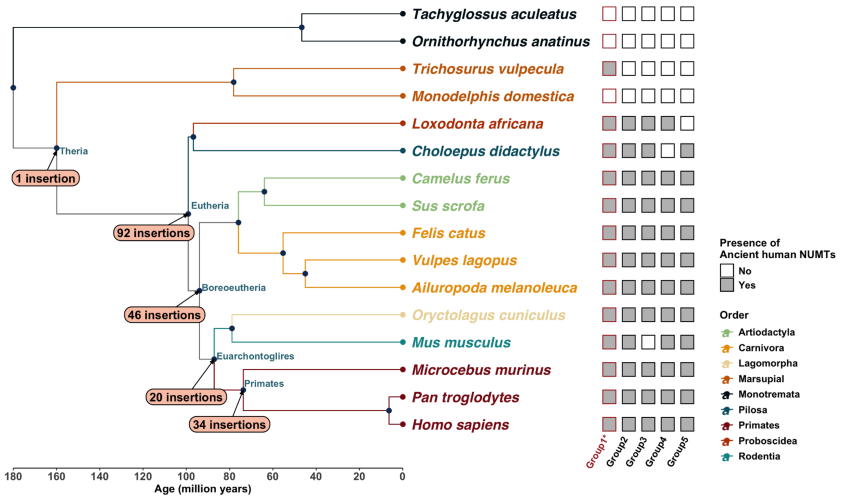


Fig. 7: The most ancient NUMTs in the human genome. The presence of the most ancient NUMT groups in human-other species alignments is shown on the right in the figure. A red border indicates potential NUMTs—highly degenerated NUMTs that could not be directly detected through NU genome:MT genome or NU genome:MT protein comparisons in the human genome but are inferred from alignments with other species.

Chromosome	ChromStart	ChromEnd	Description	Group
chr2	118055298	118055357	potential	1
chr2	40955198	40955359	human NUMTs assembled#60	2
chr3	71949580	71949689	human NUMTs assembled#141	2
chr7	116306306	116306513	human NUMTs assembled#335	2
chr16	17992090	17992363	human NUMTs assembled#608	2
chr4	123837642	123837724	human NUMTs assembled#215	3
chr5	37164410	37164493	human NUMTs assembled#237	3
chr7	37518741	37518852	human NUMTs assembled#317	3
chr11	37541525	37543000	human NUMTs assembled#492	3
chr9	7511591	7511719	human NUMTs assembled#406	4
chr5	160339468	160339652	human NUMTs assembled#271	5

Table 1: Description of the 11 most conserved human NUMTs across chromosomes

4 Conclusion and Discussion

Detecting highly degraded NUMTs remains challenging due to the limitations of conventional search strategies, which hinder our understanding of their potential influence on the genome and organism. This study proposes a novel detection pipeline using a refined sequence comparison method to overcome this limitation. A key advantage of this method is its parameter optimization, which is specifically designed to detect ancient NUMTs. This sensitive approach is expected to become a powerful tool for comprehensively identifying NUMTs.

Using this pipeline, we completed the annotation of NUMTs in the human genome assembly GRCh38, which has now been implemented in the UCSC Genome Browser. Analysis of NUMT orthology across 16 mammalian genomes revealed many ancient NUMTs that have not been previously reported. While the majority of ancient NUMTs identified in this study appear to have been inserted into the nuclear genome of the common ancestor of Eutherian mammals, even older NUMTs may be discovered with expanded analyses across genomes from more diverse orders.

In summary, this study not only deepens our knowledge of NUMTs but also serves as a testament to the power of innovative computational approaches in revealing hidden layers of genomic information. These findings are expected to inspire further research into the evolutionary and functional significance of NUMTs, as well as other relics of ancient genomic processes.

Acknowledgments. We are grateful to Silvia Rodriguez for discussions about detecting NUMTs in various species and for her assistance in refining the pipeline.

Disclosure of Interests. None declared.

References

1. Borensztajn, K., Chafa, O., Alhenc-Gelas, M., Salha, S., Reghis, A., Fischer, A.M., Tapon-Bretaudière, J.: Characterization of two novel splice site mutations in human factor vii gene causing severe plasma factor vii deficiency and bleeding diathesis. British Journal of Haematology **117** (2002). https://doi.org/10.1046/j.1365-2141.2002.03397.x
2. Calabrese, F.M., Balacco, D.L., Preste, R., Diroma, M.A., Forino, R., Ventura, M., Attimonelli, M.: Numts colonization in mammalian genomes. Scientific Reports **7** (2017). https://doi.org/10.1038/s41598-017-16750-2
3. Frith, M.C.: A simple method for finding related sequences by adding probabilities of alternative alignments. Genome Research p. gr.279464.124 (8 2024). https://doi.org/10.1101/gr.279464.124
4. Frith, M.C., Kawaguchi, R.: Split-alignment of genomes finds orthologies more accurately. Genome Biology **16** (2015). https://doi.org/10.1186/s13059-015-0670-9
5. Hamada, M., Ono, Y., Asai, K., Frith, M.C., Hancock, J.: Training alignment parameters for arbitrary sequencers with LAST-TRAIN. Bioinformatics **33** (2017). https://doi.org/10.1093/bioinformatics/btw742

6. Hazkani-Covo, E.: A burst of numt insertion in the dasyuridae family during marsupial evolution. Frontiers in Ecology and Evolution **10** (2022). https://doi.org/10.3389/fevo.2022.844443
7. Hazkani-Covo, E., Graur, D.: A comparative analysis of numt evolution in human and chimpanzee. Molecular Biology and Evolution **24** (2007). https://doi.org/10.1093/molbev/msl149
8. Hoser, S.M., Hoffmann, A., Meindl, A., Gamper, M., Fallmann, J., Bernhart, S.H., Müller, L., Ploner, M., Misslinger, M., Kremser, L., Lindner, H., Geley, S., Schaal, H., Stadler, P.F., Huettenhofer, A.: Intronic tRNAs of mitochondrial origin regulate constitutive and alternative splicing. Genome Biology **21** (2020). https://doi.org/10.1186/s13059-020-02199-6
9. Kiełbasa, S.M., Wan, R., Sato, K., Horton, P., Frith, M.C.: Adaptive seeds tame genomic sequence comparison. Genome Research **21** (2011). https://doi.org/10.1101/gr.113985.110
10. Liang, B., Wang, N., Li, N., Kimball, R.T., Braun, E.L.: Comparative genomics reveals a burst of homoplasy-free numt insertions. Molecular Biology and Evolution **35** (2018). https://doi.org/10.1093/molbev/msy112
11. Noutsos, C., Kleine, T., Armbruster, U., DalCorso, G., Leister, D.: Nuclear insertions of organellar DNA can create novel patches of functional exon sequences (2007). https://doi.org/10.1016/j.tig.2007.08.016
12. Pearson, W.R.: Identification of homologs. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics **1-3**, 980–984 (1 2019). https://doi.org/10.1016/B978-0-12-809633-8.20180-5
13. Simone, D., Calabrese, F.M., Lang, M., Gasparre, G., Attimonelli, M.: The reference human nuclear mitochondrial sequences compilation validated and implemented on the UCSC genome browser. BMC Genomics **12** (2011). https://doi.org/10.1186/1471-2164-12-517
14. Triant, D.A., Pearson, W.R.: Comparison of detection methods and genome quality when quantifying nuclear mitochondrial insertions in vertebrate genomes. Frontiers in Genetics **13** (11 2022). https://doi.org/10.3389/fgene.2022.984513
15. Tsuji, J., Frith, M.C., Tomii, K., Horton, P.: Mammalian numt insertion is nonrandom. Nucleic Acids Research **40** (2012). https://doi.org/10.1093/nar/gks424
16. Turner, C., Killoran, C., Thomas, N.S., Rosenberg, M., Chuzhanova, N.A., Johnston, J., Kemel, Y., Cooper, D.N., Biesecker, L.G.: Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer. Human Genetics **112** (2003). https://doi.org/10.1007/s00439-002-0892-2
17. Uvizl, M., Puechmaille, S.J., Power, S., Pippel, M., Carthy, S., Haerty, W., Myers, E.W., Teeling, E.C., Huang, Z.: Comparative genome microsynteny illuminates the fast evolution of nuclear mitochondrial segments (NUMTs) in mammals. Molecular Biology and Evolution **41** (2024). https://doi.org/10.1093/molbev/msad278
18. Xue, L., Moreira, J.D., Smith, K.K., Fetterman, J.L.: The mighty numt: Mitochondrial DNA flexing its code in the nuclear genome. Biomolecules **13**, 753 (4 2023). https://doi.org/10.3390/biom13050753
19. Yao, Y., Frith, M.: Improved DNA-versus-protein homology search for protein fossils. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2022). https://doi.org/10.1109/tcbb.2022.3177855
20. Zhang, G.J., Dong, R., Lan, L.N., Li, S.F., Gao, W.J., Niu, H.X.: Nuclear integrants of organellar DNA contribute to genome structure and evolution in plants (2 2020). https://doi.org/10.3390/ijms21030707