

A long-read human pangenome initiative for comprehensive interpretation of nuclear-embedded mitochondrial DNA

Received: 9 October 2025

Accepted: 20 March 2026

Published online: 15 May 2026

 Check for updates

Lianting Fu ^{1,12} , Jieyi Chen^{1,12}, Da Lian ^{1,12}, Siyuan Du², Dongya Wu ^{3,4,5}, Chentao Yang ^{6,7}, Ziyi Wang^{1,8}, Hongyi Ma^{1,8}, Zhengtong Li¹, Nicole J. Lake ^{9,10}, Xiangyu Yang ¹, Yongyong Shi ^{2,11}, Guojie Zhang ^{3,4,5} , Kaiyue Ma ¹  & Yafei Mao ^{1,3} 

Nuclear-embedded mitochondrial DNA segments (NUMT) preserve a record of ongoing mitochondrial-to-nuclear DNA transfer during evolution, with important implications for disease mechanisms and genome organization. Here, we develop a pangenome graph-based NUMT detection approach, achieving a 2.52-fold improvement in sensitivity and generating a high-resolution human NUMT map comprising 774 fixed and 280 polymorphic events, alongside 74 superpopulation-stratified loci. Notably, NUMTs derived from the 3'-end of mtDNA D-loop are less frequently fixed and exhibit *cis*-regulatory activity, revealing selective pressures shaping their genomic landscape. We also identify seven NUMTs associated with gene expression or splicing, suggesting their potential modulatory functions. Comparative analysis of complete primate genomes reveals lineage-specific NUMT dynamics, with particularly high rates in the *Pan* lineage. Furthermore, we uncover two NUMT-derived tandem repeats, establishing them as a novel source of complex variants. In summary, the integrated analysis enhances understanding of NUMT genomic architecture, population dynamics, and evolutionary implications, establishing them as dynamic genomic components of biomedical relevance.

Nuclear-embedded mitochondrial DNA segments (NUMTs) represent a unique class of genomic elements that chronicle the continuous transposition of mitochondrial DNA to the nuclear genome throughout evolutionary history^{1–3}. Since the endosymbiosis event that established mitochondria, making mitochondrial DNA (mtDNA) the most proximal “foreign DNA” to the nucleus, mtDNA has been continuously integrated to the nuclear genome, potentially through intracellular processes involving mitochondrial degradation and DNA repair mechanisms⁴. Through evolution, some of these inserted mtDNA sequences have become permanent components of the nuclear genome, expressing mitochondrial genes using nuclear and cytoplasmic machinery and reinforcing the mitochondrion-nucleus

codependency^{2,4}. NUMTs, representing the other inserted mtDNA sequences that had been widely recognized as non-functional, are associated with more and more implications for genetics and disease as indicated in recent studies^{5–7}.

A comprehensive inspection of NUMTs in large population datasets offers valuable resources for improving the understanding of their origin, evolution, genomic impact, and disease-related mechanisms. Previously, using next-generation sequencing data from the 100,000 Genomes Project in England, a study analyzed NUMTs in 66,083 individuals, revealing the ongoing NUMT formation⁸. Additionally, the study performed Oxford Nanopore sequencing for 39 individuals, showcasing the potential of long-read sequencing for NUMT characterization.

A full list of affiliations appears at the end of the paper.  e-mail: fulianting@sjtu.edu.cn; guojiezhang@zju.edu.cn; yukae@sjtu.edu.cn; yafmao@sjtu.edu.cn

However, the whole-genome landscape of NUMTs, particularly in complex genomic regions, has remained understudied^{8–11}. This incomplete characterization has left critical knowledge gaps in NUMT genomic location, sequence features, and population frequency patterns, as well as the functional consequences of polymorphic NUMTs that vary among individuals. Likewise, the relationship between NUMT methylation levels and allele frequencies in human populations has not been comprehensively characterized, hindering our understanding of the methylation dynamics following NUMT insertion.

Recent advances in long-read sequencing and pangenome construction provide unprecedented opportunities to bridge the gaps in NUMT biology^{12,13}. High-quality, haplotype-resolved genome assemblies from diverse populations enable precise identification of NUMTs in previously inaccessible genomic regions or those with complex concatenated structures, while pangenome graphs facilitate a comprehensive analysis of NUMT population genetics and evolutionary dynamics¹⁴. This technological revolution allows accurate distinction between fixed NUMTs present in all individuals and polymorphic variants exhibiting population-specific patterns, resolving the conflation inherent in single linear reference analyses. Furthermore, enhanced orthology alignments allow for precise NUMT characterization even within complex genomic regions^{15,16}. This high-resolution characterization facilitates downstream functional analyses, such as genome-wide association studies (GWAS) and expression quantitative trait locus (eQTL) studies^{17,18}. Additionally, by integrating pangenome-based NUMT characterization with comparative genomics across primate species, we can reconstruct the evolutionary history of NUMT insertion events and quantify lineage-specific insertion rates, thereby providing insights into cellular, reproductive, and developmental mechanisms driving NUMT insertion^{19–21}.

Here, we employed a pangenome graph-based approach for comprehensive NUMT identification and applied it to a Minigraph-Cactus (MC) pangenome graph constructed from 538 haplotype-resolved assemblies across three consortia [269 individuals, including 160 from the Asian Pan-Genome Project Phase 1 (APGp1), 44 from the Human Pangenome Reference Consortium Phase 1¹² (HPRCp1), and 65 from the Human Genome Structural Variation Consortium Phase 3²² (HGSVCp3)]. Additionally, we supplemented the pangenome analysis with short-read genotyping data from 2,504 unrelated individuals in the 1000 Genomes Project^{23,24} (IKGP). Our integrated analysis aims to: (1) generate a comprehensive map for human NUMTs, (2) characterize genomic features of fixed and polymorphic NUMTs, including mtDNA origin bias, localization preference, sequence divergence, and methylation dynamics, (3) explore functional impacts of polymorphic NUMTs on gene sequence, expression and splicing, and (4) delineate evolutionary dynamics of NUMTs across primates to understand the forces shaping NUMT insertion, retention, and expansion.

Results

NUMT characterization in 269 human genomes

Conventional NUMT detection relies on short-read alignments to identify split reads exhibiting nuclear-mitochondrial dual-alignment clipping, followed by breakpoint refinement through realignment^{8,25,26} (Fig. 1a). To overcome the inherent resolution and sensitivity limitations of the short-read approach, we developed a Pangenome Graph-based NUMT Detection (PG-NUMT) approach to comprehensively identify NUMTs across 200 +1 benchmark individuals, including 40 from HPRCp1, 160 from APGp1, and the reference genome T2T-CHM13 (Fig. 1b, Supplementary Data 1, and Supplementary Fig. 1, 2, Methods). Briefly, PG-NUMT aligns multiple primate mitochondrial genomes to the pangenome graph, enabling sensitive, comprehensive NUMT detection with precise breakpoint identification. Using PG-NUMT, we identified a total of 1,122 NUMTs, including 918 derived from T2T-CHM13 (“reference NUMTs”) and 204 from the other genomes (“non-reference NUMTs”). In comparison, only 83 non-reference NUMTs

were detected using the short-read approach, including two misidentified ones (Supplementary Data 2 and Supplementary Fig. 3). Benchmark results across the 200 individuals revealed a 2.52-fold improvement in NUMT detection sensitivity, with a particularly notable improvement for long NUMTs exceeding 1000 bp (Figs. 1c, d, from 5 to 19). Importantly, PG-NUMT can evaluate reference NUMT presence in any given individual and distinguish NUMTs between haplotypes, which is unattainable with short-read approaches (Figs. 1a, b). Our benchmark results demonstrate that PG-NUMT substantially enhances both the accuracy and sensitivity of NUMT detection (Supplementary Data 2 and Supplementary Fig. 4–6).

Given its superior performance in benchmark evaluations, PG-NUMT was applied to a Minigraph-Cactus (MC) pangenome graph based on T2T-CHM13, which incorporates 269 individuals representing diverse global superpopulations (538 haplotype-resolved assemblies; Supplementary Data 1). Within the MC pangenome graph, 1179 NUMTs were identified, including 918 reference and 261 non-reference NUMTs (Supplementary Data 3). Among the non-reference NUMTs, 7 were categorized as concatenated mega-NUMTs (NUMTs composed of multiple concatenated mtDNA-derived segments; Supplementary Data 3, Supplementary Fig. 4 and 7–9). Each individual possesses an average of 8.67 non-reference NUMTs (s.d. = 2.85, Supplementary Fig. 10). Among 1,056 NUMTs (excluding 123 in centromeric regions), 155 (14.7%) are singletons [allele count (AC) = 1], 61 (5.8%) are rare or ultra-rare NUMTs [AC > 1 and allele frequency (AF) < 1%], and 840 (79.5%) are common NUMTs (AF ≥ 1%), with all reference NUMTs classified as common. Notably, incorporating additional samples provides a marginal contribution to the pool of common NUMTs (Fig. 1e), indicating that the current MC pangenome graph already captures a substantial portion of the common NUMT landscape in humans.

Our large, diverse pangenome graph improves NUMT detection resolution, particularly for common NUMTs that are extensively represented. This enables accurate NUMT genotyping through short-read sequencing in even larger human cohorts, thereby enabling further inspection of the NUMT landscape. Here, we genotyped NUMTs in short-read WGS data from 2,504 unrelated individuals within 1KGP using PanGenie¹⁷ based on the MC pangenome graph, successfully characterizing 1043 loci (Fig. 1f and Supplementary Data 3, Methods). To evaluate the genotyping performance, allele frequencies from the MC pangenome graph were compared with those generated by PanGenie (Methods). The high correlation between approaches confirms PanGenie’s reliability for NUMT genotyping accuracy across the 2504 individuals (Pearson’s $r = 0.9998$, Fig. 1g and Supplementary Fig. 11, 12).

The fixed and population-stratified NUMTs in the human genome

Accurate and comprehensive NUMT characterization enhances the investigation of their population features. Here, NUMTs are classified into three categories based on allele frequencies and genome coordinates. Fixed NUMTs are stably integrated into the genomes with near-universal prevalence across populations (AF ≥ 95%), while polymorphic NUMTs exhibit individual-level variation (AF < 95%). Pericentromeric NUMTs, specifically located within centromeric regions, are classified as a separate category due to challenges preventing unambiguous classification as either fixed or polymorphic, such as sub-optimal assembly accuracy and divergent alignments across genomes. Using PG-NUMT and PanGenie, we characterized 774 fixed NUMTs, 280 polymorphic NUMTs, and 123 pericentromeric NUMTs across 269 haplotype-resolved individuals and 2504 unrelated short-read individuals (Fig. 2a and Supplementary Data 3). All 774 fixed NUMTs are present in the reference genome T2T-CHM13. Among polymorphic NUMTs, 71.1% exhibit low allele frequencies in human populations [47.8% ultra-rare (AF < 0.1%) and 23.3% rare (0.1% ≤ AF < 1%)]. Notably, among the remaining 28.9% of polymorphic NUMTs [common NUMTs (AF ≥ 1%)], some exhibit population-specific allele frequency variations

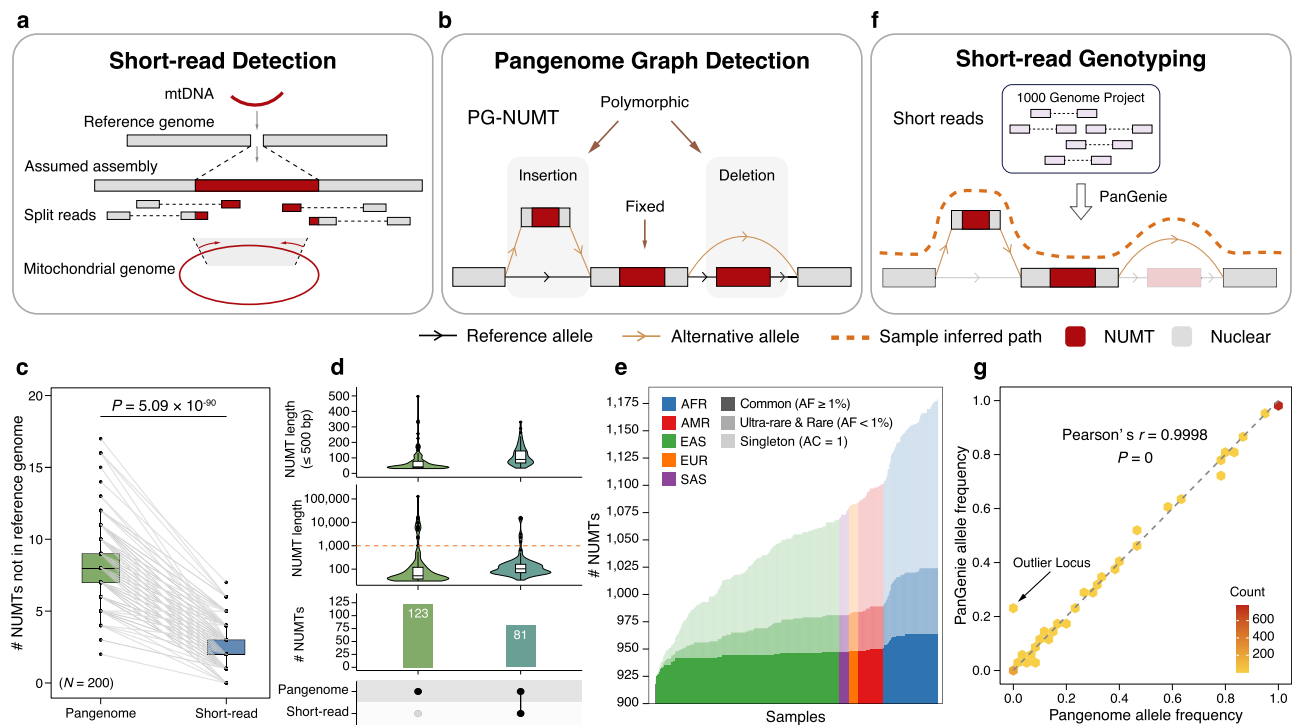


Fig. 1 | NUMT detection with pangenome graph. **a** NUMT detection using short-read sequencing data. **b** NUMT detection using the pangenome graph. PG-NUMT, Pangenome Graph-based NUMT Detection. **c** The paired comparison of detected NUMTs absent in the T2T-CHM13 reference genome shows a 2.52-fold increased sensitivity of PG-NUMT compared to the short-read approach (sample size $N = 200$, two-tailed Welch's t -test, $P = 5.09 \times 10^{-90}$). Connecting lines indicate the same samples. **d** Distribution comparison of length and counts of identified NUMTs between PG-NUMT and the short-read approach. Left, NUMTs uniquely identified by PG-NUMT ($n = 123$); Right, consensus NUMTs detected by both approaches, excluding two loci uniquely misclassified by short-read ($n = 81$; Supplementary Fig. 3). From bottom to top: UpSet plot comparing NUMT counts between approaches; violin plot of NUMT length distribution; violin plot of NUMT length

distribution (≤ 500 bp). **e** Cumulative NUMT growth curves depicting sequential assembly addition to the pangenome graph, where colors from light to deep indicate singleton ($AC = 1$), ultra-rare & rare ($AF < 1\%$), and common ($AF \geq 1\%$) NUMTs. **f** Short-read NUMT genotyping across 2,504 unrelated IKGp individuals by PanGenie within a pangenome graph. **g** Comparison of allele frequencies between PanGenie genotypes and the pangenome graph across 40 HPRCp1 and 160 APGp1 individuals shows high concordance (Pearson's $r = 0.9998$, two-tailed $P = 0$). The outlier locus (pannum_109, indicated by an arrow) exhibits reduced genotyping accuracy due to adjacent repetitive sequences and the inversion structure (Supplementary Fig. 12). Box plots denote median and interquartile range (IQR), with whiskers $1.5 \times IQR$. Source data are provided as a Source Data file.

(Supplementary Data 4), hinting at different demographic histories or genetic drift across human populations (Methods).

Fixed NUMTs exhibit a length range of 38 bp to 14,855 bp (median: 203 bp; mean: 637.8 bp, Supplementary Fig. 13), whereas polymorphic NUMTs range from 31 bp to an extraordinary 127,749 bp (median: 74.5 bp; mean: 1143.6 bp for all and 689.8 bp excluding the largest NUMT, Fig. 2b). Despite this broad range, the majority of both fixed (73.4%) and polymorphic (87.9%) NUMTs are shorter than 500 bp. Unexpectedly, no significant negative correlation was observed between NUMT length and allele frequency, challenging previously reported patterns⁸ (Pearson's $r = 0.04$, 95% CI: -0.08 - 0.15 , two-tailed $P = 0.56$, Fig. 2c).

To further explore the population stratification patterns of NUMTs, we compared their distribution across human superpopulations among 2504 unrelated IKGp individuals. A total of 74 superpopulation-stratified NUMTs were identified (Fig. 2d, Supplementary Data 4, and Supplementary Fig. 14, Methods). These population-stratified NUMTs likely represent independent genomic integration events during the evolutionary divergence of human populations.

Genomic landscape of the fixed and polymorphic NUMTs

NUMT polymorphisms and allele frequencies may reflect the complex interplay of genetic drift, population dynamics and potential functional mechanisms that govern the evolutionary fate of different NUMTs. Integrating this perspective in NUMT analyses by characterizing fixed

and polymorphic NUMTs separately helps reveal distinct genomic features, generating novel insights into the genomic effects of NUMTs. For instance, a remarkable bias was identified in the mtDNA origin distribution patterns. While polymorphic NUMTs exhibit uniform coverage across the mtDNA sequence (permutation test, $P = 0.54$, Methods), fixed NUMTs display a notably decreased coverage in the 3'-end region (chrM:72-573, mtDL3 hereafter, relative coverage < 0.30 , permutation test, $P = 0$) of the non-coding displacement loop (D-loop) (Fig. 3a and Supplementary Fig. 15). To exclude the alignment artifacts as a potential cause for this observation, rigorous realignment analyses were performed using multiple mtDNA queries ($N = 1000$)²⁷, or reference sequences with shifted start positions (Supplementary Fig. 16, Methods). Additionally, the divergent patterns observed between fixed and polymorphic NUMTs suggest that the reduced coverage in mtDL3 more likely arises from post-insertion events rather than bias during integration (Fig. 3a). To explore the mechanism underlying the selective removal, we assessed the regulatory potential of the mtDL3-derived sequence through a reporter assay (Figs. 3b-d and Supplementary Fig. 16, Methods). The mtDL3-derived sequence exhibits weak *cis*-regulatory activity, while neither the negative control nor the mtDL5-derived sequence (the 5'-end region of the D-loop, chrM:16,024-16,569) exhibits such activity (Figs. 3c, d). The potential modulatory effect of mtDL3-derived sequences provides a potential biological basis for their preferential elimination from fixed NUMT populations. However, the extent to which this activity drives their selective removal remains unclear and requires additional investigation.

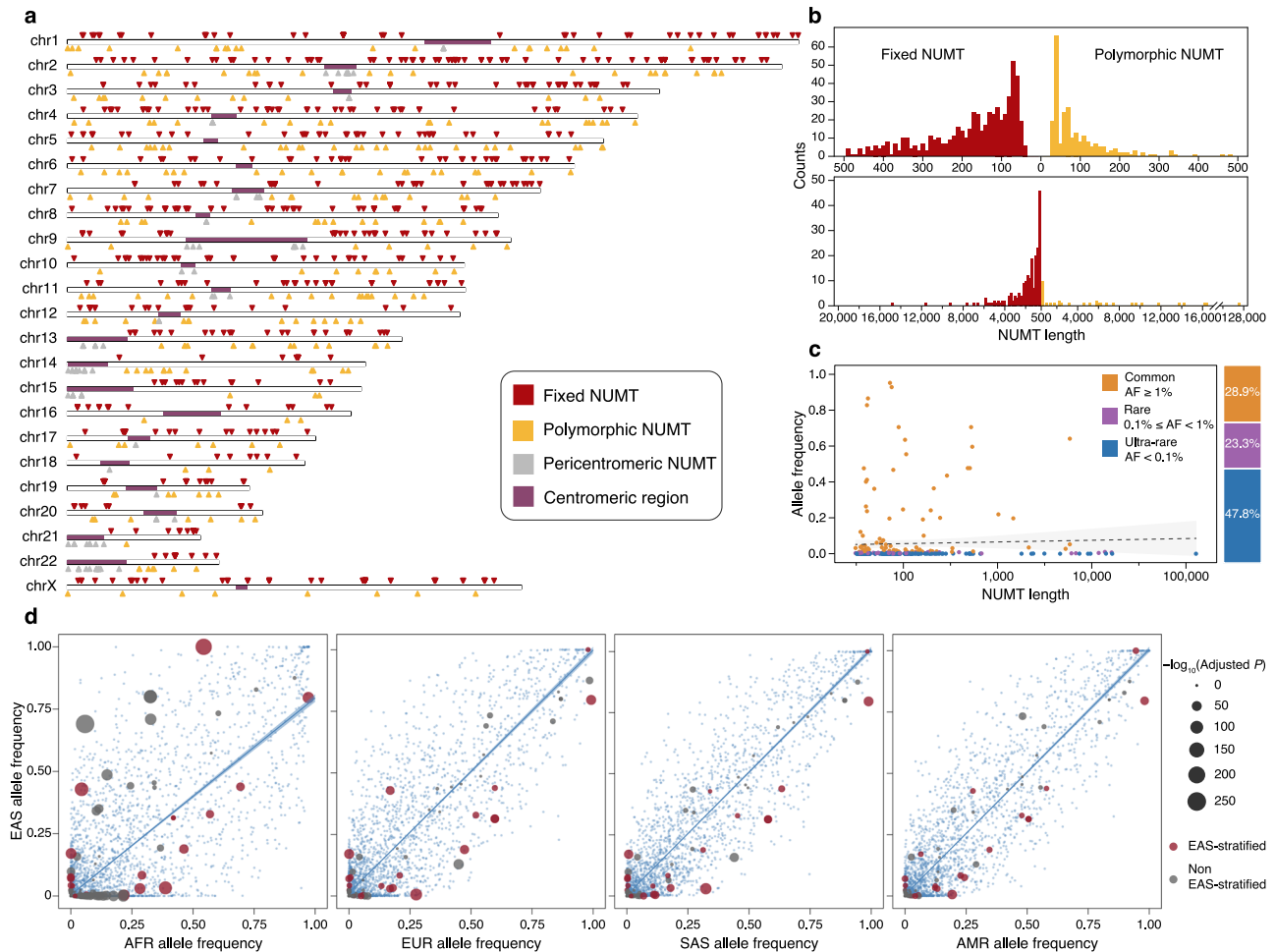


Fig. 2 | Overview of NUMTs in humans. **a** Ideogram of NUMTs from 539 human genome assemblies in this study. 774 fixed NUMTs are depicted above the ideogram track (red), while 280 polymorphic NUMTs (orange) and 123 pericentromeric NUMTs (gray) are shown below the ideogram track. Centromeric regions are labeled in purple. **b** Length distribution of fixed (left, red) and polymorphic (right, orange) NUMTs (top, 0–500 bp, 10-bp bin; bottom, > 500 bp, 200-bp bin). **c** The correlation between NUMT allele frequency and length is shown on the left. The proportions of NUMTs with different allele frequencies (ultra-rare (AF < 0.1%, blue),

rare (0.1% ≤ AF < 1%, purple), and common (AF ≥ 1%, orange)) are shown on the right. The shaded areas indicate 95% confidence intervals. **d** Allele frequency comparisons of NUMTs between East Asian (EAS) and other superpopulations. Circles represent NUMTs, with those colored red indicating EAS-stratified NUMTs (two-tailed Fisher’s exact test with Bonferroni correction, adjusted *P* value < 0.05 compared to other superpopulations). Blue dots denote SNV allele frequencies across superpopulations, serving as controls. The shaded areas indicate 95% confidence intervals. Source data are provided as a Source Data file.

Additionally, the mtDNA origin analysis also reveals pericentromeric NUMT enrichment in five specific mtDNA regions (relative coverage > 0.50, chrM:986–1,509; chrM:3,744–3,897; chrM:5,612–6,979; chrM:9,393–10,302; chrM:15,037–15,087, Fig. 3a and Supplementary Fig. 15). Since 81.43% of the length of these regions overlaps with segmental duplications, their elevated coverage is potentially attributable to genomic duplication events.

Next, we analyzed the localization preference of fixed and polymorphic NUMTs in the human genome. Enrichment analysis reveals that both fixed and polymorphic NUMTs are significantly enriched within or near segmental duplications (fixed NUMTs, permutation test, *P* = 0; polymorphic NUMTs, permutation test, *P* = 0.03; Fig. 3e, Methods). Fixed NUMTs are rarely found in regions near or within transposable elements, including DNA transposons, LINES, LTRs, and SINES (permutation test, *P* = 0, Supplementary Fig. 17). In contrast, polymorphic NUMTs display no significant deviation from the null distribution in association with transposable elements (permutation test, *P* = 0.40). These findings reveal distinct localization preferences between fixed and polymorphic NUMTs, implicating roles of duplications and transposable elements in shaping their genomic distribution (Fig. 3e and Supplementary Fig. 18, 19).

The integration of NUMTs into the nuclear genome may disrupt gene structure and regulation, potentially affecting protein function^{5,6}. To explore potential function-related effects, the distribution of fixed and polymorphic NUMT insertions was examined across different genomic regions. Fixed NUMTs showed significant enrichment in intergenic regions (permutation test, *P* = 0, Fig. 3f), representing the genomically “safe” locations where insertion is least likely to disrupt essential cellular functions. In contrast, polymorphic NUMTs showed no significant location bias (permutation test, *P* = 0.50), indicating that such location bias is established through post-integration selective mechanisms over evolutionary time rather than insertion location preference (Fig. 3f and Supplementary Fig. 20). Notably, one singleton 42-bp NUMT inserted into the coding sequence (CDS) of *GTF2I* was identified among 280 polymorphic NUMTs (Figs. 3f, g). *GTF2I* encodes a transcriptional factor and is located within a deletion dominantly associated with Williams-Beuren syndrome²⁸. The insertion occurs after the first nucleotide in exon 25 but is not predicted to affect splicing by SpliceAI²⁹ (delta scores for acceptor loss = 0.07; delta scores for acceptor gain, donor gain, and donor loss = 0.00). Nanopore long-read RNA-seq reads with the NUMT insertion, coupled with the absence of exon-skipped reads, confirm an in-frame insertion leading to

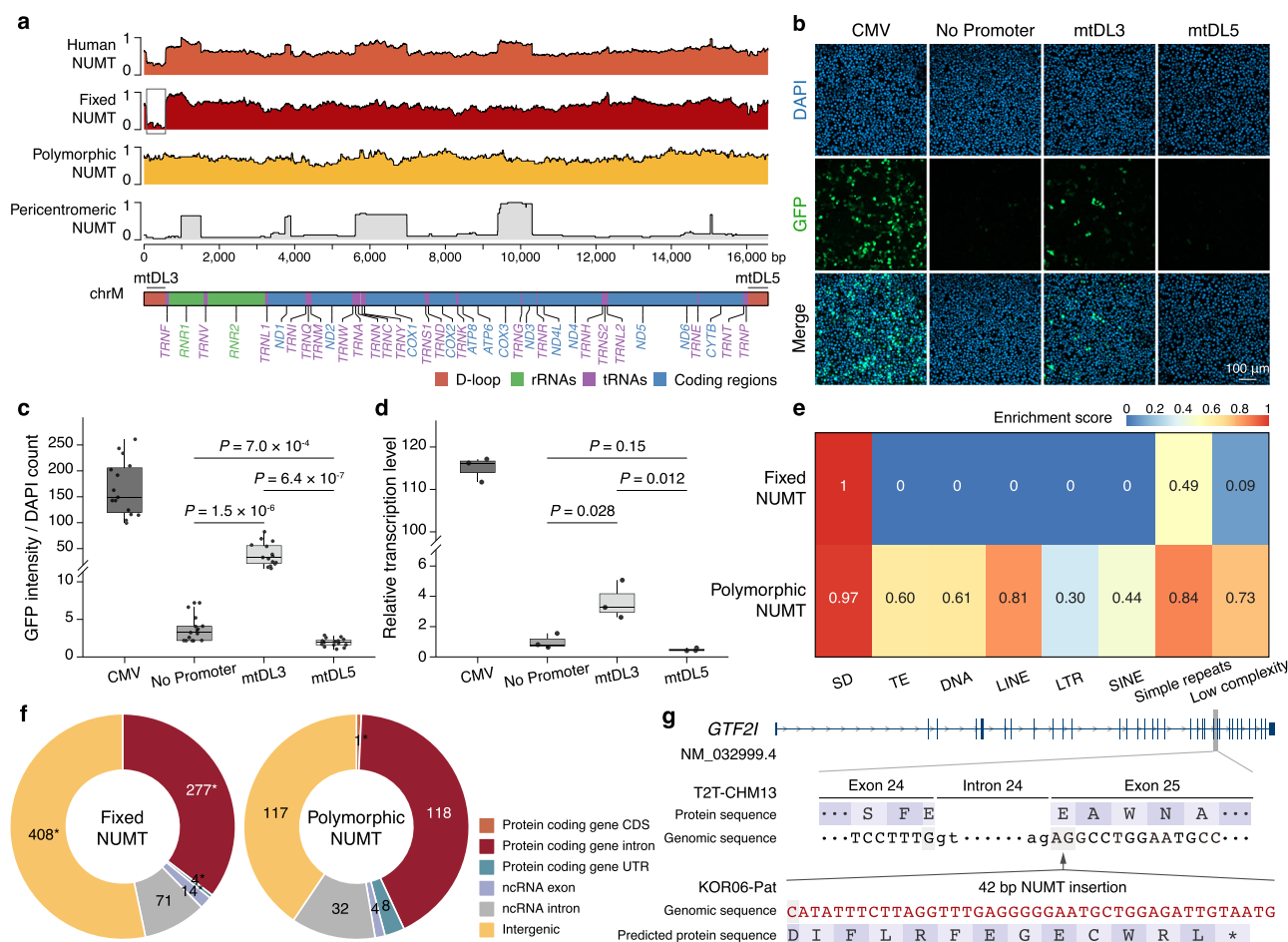


Fig. 3 | Distributions of fixed and polymorphic NUMTs. **a** NUMT coverage relative to the mtDNA coordinates. The x-axis represents the nucleotide positions of mtDNA, while the y-axis indicates the relative coverage distribution across distinct NUMT categories (normalized by their maximum coverages). The mtDL3 region (chrM:72-573) and the mtDL5 region (chrM:16,024-16,569) are marked with gray lines. Reduced coverage in the mtDL3 region of fixed NUMTs is highlighted by the gray box. **b** Representative fluorescence images of HEK293T cells transfected with recombinant plasmids, illustrating the in vitro promoter activity of mtDL3 and mtDL5. Images in the CMV group were acquired using substantially lower laser excitation intensity to avoid overexposure. DAPI, 4',6-diamidino-2-phenylindole; GFP, green fluorescent protein; CMV, cytomegalovirus enhancer and promoter. **c** Quantitative analysis of the fluorescence signals shown in Fig. 3b (two-tailed Welch's *t*-test). Five images were acquired for each of three independent transfection replicates. **d** Relative transcriptional activity of mtDL3 and mtDL5, measured by eGFP expression (two-tailed Welch's *t*-test). Three qPCR technical

replicates were performed for each of three independent transfection replicates. **e** Heatmap of enrichment analysis of fixed (top) and polymorphic (bottom) NUMT insertions across repetitive elements. Values represent enrichment scores derived from permutation tests relative to random expectations (Methods). SD segmental duplication, TE transposable element, DNA DNA transposon, LINE long interspersed nuclear element; LTR long terminal repeat, SINE short interspersed nuclear element. **f** Pie chart illustrating the proportion of fixed (left) and polymorphic (right) NUMT insertions across different genomic regions. CDS, coding sequence; UTR, untranslated region. The asterisk (*) indicates a significant deviation from the null distribution. **g** A NUMT insertion potentially alters the CDS of *GTF2I*. The 42-bp NUMT insertion sequence within the KOR06-Pat (NUMT insertion carrier) genome is highlighted in red, an in-frame insertion predicted to cause premature termination. The asterisk (*) indicates the predicted termination codon. Box plots denote median and IQR, with whiskers $1.5 \times \text{IQR}$. Source data are provided as a Source Data file.

premature termination and nonsense-mediated decay (Fig. 3g and Supplementary Fig. 21). However, considering the carrier's apparent lack of disease phenotype, further studies are required to investigate any potential physiological consequences or compensatory mechanisms.

Sequence features and methylation patterns of NUMTs

The nuclear environment imposes fundamentally different mutational constraints compared to the mitochondrial genome, leading to progressive sequence divergence that serves as a molecular clock for estimating NUMT integration timing⁸. We found that fixed NUMTs exhibit significantly lower sequence identity with mtDNA compared to polymorphic NUMTs (two-tailed Welch's *t*-test, $P = 7.67 \times 10^{-213}$, Fig. 4a and Supplementary Fig. 22), consistent with the expectation that fixed NUMTs have accumulated more mutations during their prolonged residence within the nuclear genome.

In addition, fixed NUMTs exhibit significantly reduced GC content compared to both polymorphic NUMTs (two-tailed Welch's *t*-test, $P = 2.83 \times 10^{-38}$, Fig. 4b) and the mitochondrial genome-wide average (permutation test, $P = 0$), with a GC content of 5.38% lower than the nuclear genome-wide average (from 0.408 to 0.386). This pattern aligns with mutation spectrum analyses, which reveal a pronounced preferential bias of C \rightarrow T mutation in both fixed and polymorphic NUMTs (Fig. 4c). However, the mutation spectra do not differ significantly between fixed and polymorphic NUMTs (Supplementary Data 5), suggesting shared post-integration mutational processes in both categories consistent with nuclear CpG transitions³⁰. These observations imply that the observed differences in sequence identity and GC content between NUMT categories primarily reflect the cumulative effects of time rather than distinct mutational mechanisms.

The human nuclear genome is typically hypermethylated in most inactive genomic regions^{31,32}, while the entire mitochondrial genome

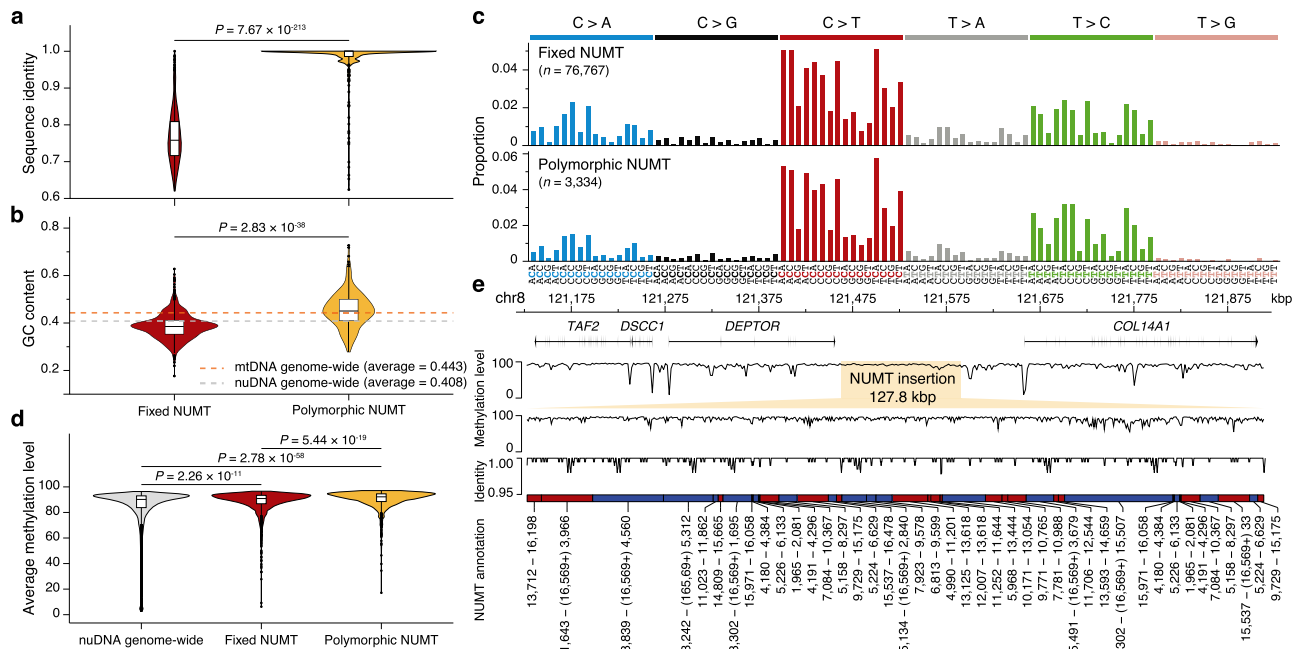


Fig. 4 | Sequence characteristics of fixed and polymorphic NUMTs.

a Comparison of sequence identity to mtDNA between fixed ($n = 774$) and polymorphic NUMTs ($n = 280$, two-tailed Welch's t -test). **b** Comparison of GC content between fixed ($n = 774$) and polymorphic NUMTs ($n = 280$, two-tailed Welch's t -test). The orange and gray lines represent the GC content observations of random sequences at the mitochondrial and nuclear genome-wide levels, respectively. **c** Trinucleotide mutation spectrum of fixed (top) and polymorphic (bottom) NUMTs compared with the revised Cambridge Reference Sequence (rCRS). **d** Average methylation level distribution of NUMT windows among fixed NUMTs (red), polymorphic NUMTs (yellow), and genome-wide nuclear DNA (gray) (two-tailed Mann–Whitney U -test). Each NUMT was divided into 200-bp windows, and

only NUMTs greater than 1 kbp were analyzed (Number of windows: fixed, $n = 1440$; polymorphic, $n = 1065$; genome-wide, $n = 10,000$). **e** Methylation pattern of a mega-NUMT on C060-CHA-N20-Mat. The top panel displays methylation levels across the mega-NUMT (chr8:121,463,091-121,590,920) and the flanking regions (chr8:121,126,410-121,914,383). The middle panel displays sequence identity to mtDNA. The bottom panel illustrates the concatenated NUMT structure with corresponding mitochondrial coordinates, where red and blue regions denote forward and reverse mtDNA strands, respectively. Coordinates formatted as “(16,569+) X” represent regions spanning the nucleotide position “1” of the mtDNA reference. Box plots denote median and IQR, with whiskers $1.5 \times \text{IQR}$. Source data are provided as a Source Data file.

exhibits low-to-none methylation levels^{33,34}. Considering this distinction, methylation analyses for NUMTs may reveal properties related to their insertion and post-insertion mechanisms. Contrasting sharply with mtDNA, both fixed and polymorphic NUMTs exhibit high methylation levels across most genomic windows (200-bp windows; fixed NUMTs: mean = 88.41%, median = 90.89%; polymorphic NUMTs: mean = 90.46%, median = 92.23%, Fig. 4d, Methods). Considering methylation levels are comparable between fixed and polymorphic NUMTs despite their presumably different evolutionary ages, the related epigenetic response is likely to be both rapid and stable, thereby effectively masking age-dependent differences. Moreover, uniform patterns of methylation level and sequence identity persist across all constituent segments of a concatenated mega-NUMT (Fig. 4e). These observations align with the proposed formation mechanism in which multiple mtDNA fragments are simultaneously integrated into the nuclear genome during NUMT insertion²⁵.

Assessment of modulatory effects of polymorphic NUMTs

The potential for polymorphic NUMTs to influence gene function represents a critical frontier in understanding how mtDNA insertions contribute to human phenotypic diversity and disease susceptibility. Unlike fixed NUMTs, which have undergone extensive evolutionary selection, polymorphic NUMTs may retain modulatory potentials due to their recent integration and variable presence across populations.

To investigate the potential gene function modulation roles of polymorphic NUMTs, we performed *cis*-eQTL and *cis*-sQTL analysis of polymorphic NUMTs, as well as SNVs in their flanking 500-kb regions using RNA-seq data from 731 individuals in the MAGE dataset³⁵. We identified 7 significant eNUMT-eGene and 16 sNUMT-sGene pairs, where

eNUMTs/sNUMTs and eGenes/sGenes represent NUMTs contributing to *cis*-eQTL/sQTL signals and their corresponding target genes (Supplementary Data 6, 7). Among those, three NUMTs exhibited dual associations impacting both gene expression and splicing, including pannumt_14 (associated with *GNL2*, Figs. 5a–c), numt_353 (associated with *RASGRP3*, Figs. 5d–f), and numt_310 (associated with *SMAD2*, Figs. 5g–i).

Specifically, pannumt_14 represents a 71-bp NUMT located in the 5' upstream region of G Protein Nucleolar 2 (*GNL2*, encoding a nucleolar GTP-binding protein required for nuclear export and maturation³⁶) and is seen in conjunction with decreased expression (Figs. 5a, b). Similarly, numt_310 also showed reduced expression association (Figs. 5g, h), which represents a 210-bp NUMT within the intron of SMAD Family Member 2 (*SMAD2*), a gene encoding the signaling transducer and transcriptional factor that regulates cell proliferation, apoptosis, and differentiation through TGF- β signaling pathway³⁷. By contrast, we identified an enhanced expression correlation between numt_353, a 245-bp NUMT insertion and its adjacent gene RAS Guanyl-Releasing Protein 3 (*RASGRP3*) (Fig. 5d). *RASGRP3* encodes a guanine nucleotide exchange factor that activates the oncogenes *HRAS* and *RAP1A*, reported to be associated with systemic lupus erythematosus and several cancers³⁸. Therefore, while none of these NUMTs constituted the lead QTL variants in their respective genomic loci (Figs. 5a, d, g, and Supplementary Fig. 23), the crucial importance of these three genes warrants further functional examination of whether the presence/absence of these NUMTs possesses gene modulatory potential.

Interestingly, these three NUMTs also exhibit significantly lower allele frequencies in EAS than in other superpopulations (Supplementary Data 4), suggesting that these variants may reflect population-specific demographic histories or represent founder effects during



Fig. 5 | Significant *cis*-eQTL effects of NUMTs. a, d, g LocusZoom plots of regional eQTL association analyses for three NUMTs exhibiting dual QTL effects, including *pannumt_14* and *GNL2* (**a**), *numt_353* and *RASGRP3* (**d**), *numt_310* and *SMAD2* (**g**). Arrows indicate the direction of gene transcription. **b, e, h** Covariate-adjusted normalized expression residuals comparisons across NUMT genotypes (two-tailed Mann–Whitney *U*-test) for *GNL2* (**b**), *RASGRP3* (**e**), and *SMAD2* (**h**). **c, f, i** Normalized

expression comparisons across superpopulations (two-tailed Mann–Whitney *U*-test) for *GNL2* (**c**), *RASGRP3* (**f**), and *SMAD2* (**i**). Sample sizes of each superpopulation: EAS ($N = 141$), AFR ($N = 196$), AMR ($N = 113$), EUR ($N = 142$), and SAS ($N = 139$). Box plots denote median and IQR, with whiskers $1.5 \times \text{IQR}$. Source data are provided as a Source Data file.

human population expansion. However, population-specific expression differences corresponding to NUMT frequency patterns were observed only for *GNL2* (Fig. 5c) and *RASGRP3* (Fig. 5f), but not for *SMAD2* (Fig. 5i), indicating that the interplay of NUMT presence, population stratification, and functional consequences is not straightforward and likely depends on additional genetic and environmental factors.

NUMT evolution in primates

Although previous studies revealed the continuous integration of NUMTs during primate evolution, the lack of high-quality genomic resources has hindered a deeper understanding of NUMT phylogenetic patterns and insertion rates across primate lineages^{11,20,39}. Here, we utilized 20 high-quality long-read non-human primate genomes (together with 270 human genomes) to systematically identify NUMTs, classify them as either shared or lineage-specific, and estimate the NUMT insertion rate in evolutionarily conserved syntenic regions for each lineage (Fig. 6a, Supplementary Data 8–9, and Supplementary Fig. 24). NUMT insertion events exhibit a significant positive correlation with the evolutionary genetic distances between species, as measured by the SNV rate (Pearson's $r = 0.83$, two-tailed $P = 2.71 \times 10^{-3}$, Supplementary Fig. 25). Furthermore, the observed NUMT accumulation pattern is consistent with lineage-specific insertion events¹⁹ (Supplementary Fig. 25). Notably, the *Pan* lineage exhibits higher NUMT insertion rates in syntenic genomic regions across primates (chimpanzees, 16.0 insertions per million years; bonobos, 18.7), whereas gorillas display a comparatively lower insertion rate (3.1 insertions per million years). In addition to these syntenic regions, 8 recurrent duplication clusters were identified in evolutionarily

dynamic regions across primates (Fig. 6b, Supplementary Data 10, and Supplementary Fig. 26), with most prominent example located at chr11:89,828,225–90,018,356 in T2T-CHM13 (Fig. 6b). Phylogenetic analysis of these NUMT segments across primates reveals that duplication events occurred independently in both the *Pan* and gorilla lineages (Fig. 6c). Another distinct expansion pattern was observed in macaques, where *DUX4*-NUMT joint tandem duplication generated two macaque-specific arrays (chr1 and chr6, Supplementary Fig. 27, 28). This complex structure disrupts the continuous high-methylation region typically observed in humans. These findings indicate that duplication serves as a driving force behind NUMT expansion during primate evolution, a pattern similarly observed in human populations (Fig. 3e and Supplementary Fig. 18, 19). This duplication-driven expansion model provides a mechanistic explanation for the clustered distribution of certain NUMTs observed in human and non-human primate genomes and suggests that NUMT abundance may be influenced by both post-integration duplications and initial rates of mitochondrial DNA transfer. Additionally, fixed NUMTs in non-human primates consistently exhibit decreased coverage in the mtDL3 region across the mtDNA sequence, mirroring the pattern observed in human fixed NUMTs (Supplementary Fig. 29). This phylogenetic conservation of sequence bias suggests that the selective pressures responsible for mtDL3 depletion have operated throughout primate evolution, indicating that certain mitochondrial sequences possess intrinsic properties that make them less suitable for long-term nuclear residence.

The origins of VNTR within NUMTs

Previous comparative genomic analyses reveal that homologous NUMT loci accumulate sequence variations throughout evolution,

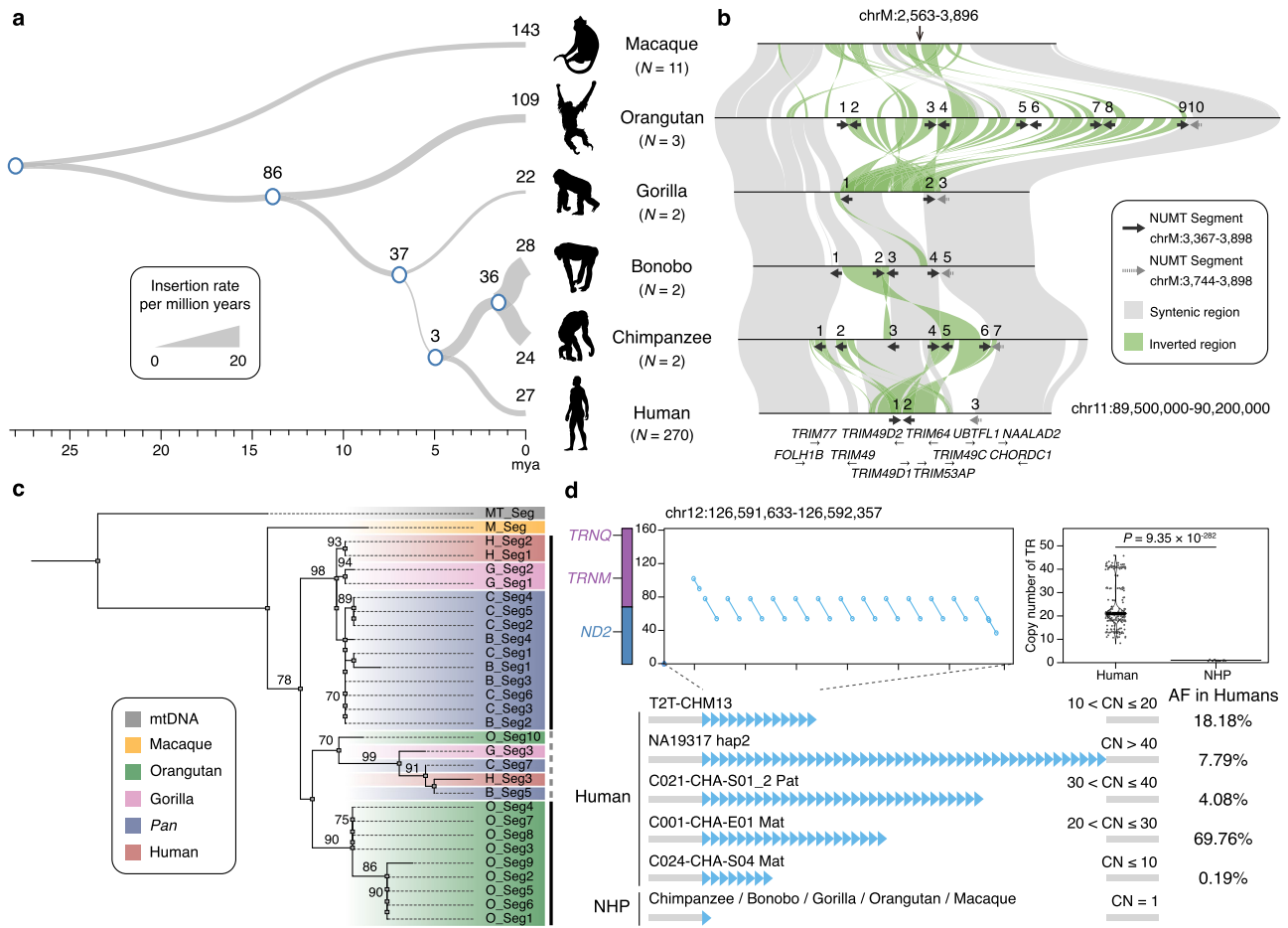


Fig. 6 | NUMT evolution in primates and VNTR origins within NUMTs. **a** The phylogenetic tree of fixed lineage-specific NUMTs in syntenic regions in primate evolution. Tree branch widths are scaled proportionally to the NUMT insertion rate per million years for each lineage (ranging from 1.5 to 18.7 insertions per million years). The number of NUMT insertions for each lineage is shown at each node or tip. Primate silhouettes are obtained from PhyloPic (<https://www.phylopic.org>). **b** Syntenic comparison of recurrent NUMTs across primates at chr11:89,500,000-90,200,000 in T2T-CHM13. Black and gray arrows indicate NUMT segments corresponding to human chrM:3,367-3,898 and chrM:3,744-3,898, respectively. Arrow orientation reflects the direction of the NUMT segments. The gene models are

present at the bottom. **c** The phylogenetic tree of NUMT segments from Fig. 6b illustrates NUMT duplication and lineage-specific expansion. Bootstrap values (bootstrap value ≥ 70) are indicated at branch nodes. **d** A human-specific NUMT-derived VNTR. The syntenic comparison of the NUMT sequence and mtDNA sequence (left) and tandem repeat motif copy number distribution (right; two-tailed Welch's *t*-test) are shown at the top. Schematic representations of the human-specific NUMT-derived VNTR and its allele frequency distribution across humans are displayed at the bottom. CN, copy number. Source data are provided as a Source Data file.

including SNVs and indels³⁹. Additionally, our comparative analysis reveals a previously unrecognized source of genomic variation through the identification of NUMT-derived variable number tandem repeats (VNTRs), establishing mitochondrial insertions as contributors to the diversity of complex genomic regions. We identified two distinct NUMT-derived VNTRs with different evolutionary histories and potential functional significance, which are a human-specific NUMT-derived VNTR (Fig. 6d) and a NUMT-derived VNTR expanded in humans (Supplementary Fig. 30). The human-specific NUMT-derived VNTR (chr12:126,591,633-126,592,357 in T2T-CHM13) resides within a “gene desert” region (68,542 bp from the nearest gene *LOC100996671*). This VNTR displays a copy number variation ranging from 8 to 46 in humans and has expanded in archaic humans (including both the *Neanderthal* and the *Denisovan*; copy number > 1 ; Supplementary Fig. 31) but remains a single-copy locus in non-human primates. Notably, 2-bp “AT” microhomologies are observed at the repeat boundaries, suggesting that replication slippage might be associated to this expansion (Supplementary Fig. 32). Similarly, the NUMT-derived VNTR at chr2:238,011,642-238,015,272 in T2T-CHM13 exhibits a copy number range of 9 to 18 in humans, representing an

expansion compared to the fewer than 7 copies typically observed in non-human primates (Supplementary Fig. 30). Interestingly, this VNTR resides within an intron of *MLPH* (encoding melanophilin), a gene potentially associated with human pigmentation⁴⁰ and Griscelli Syndrome 3⁴¹. The discovery that mitochondrial sequences can serve as substrates for VNTR formation adds a new dimension to our understanding of how repetitive elements contribute to genome evolution and suggests that the functional impact of NUMT insertions may extend beyond their initial integration sites through subsequent structural rearrangements.

Discussion

This comprehensive investigation of nuclear-embedded mitochondrial DNA sequences represents the most extensive characterization of NUMT landscapes in human and primate genomes to date, fundamentally advancing our understanding of mitochondrial-nuclear genome interactions and their evolutionary consequences. By integrating population genomics, comparative phylogenetics, and functional genomics approaches, we have revealed that NUMTs constitute a dynamic and functionally relevant component of primate genome

evolution, challenging views of these sequences as merely genomic fossils and establishing them as potential active contributors to genetic diversity, gene modulation, and evolutionary adaptation.

Long-read sequencing and pangenome graphs enable the precise identification of NUMT insertion sites and sequence characteristics. Moreover, this approach enhances comparative analyses of NUMT diversity by effectively distinguishing fixed NUMTs from polymorphic NUMTs in human populations. Fixed and polymorphic NUMTs exhibit both shared and divergent patterns in mtDNA origin bias, genomic insertion preference, sequence divergence, and methylation levels, which aid in studying post-integration modification processes and selective constraints on NUMT insertion. Notably, contrary to what was previously reported⁸, no significant negative correlation was observed between NUMT length and allele frequency in this study (Fig. 2c). Therefore, it is likely that the selective removal of certain NUMTs is more associated with factors other than NUMT length, such as insertion position and insertion sequences. Additionally, high-quality genomes unlock NUMT detection within previously uncharacterized regions, including centromeric regions. In this study, it is revealed that pericentromeric NUMTs are enriched in human acrocentric chromosomes (chr13, chr14, chr15, chr21, and chr22, permutation test, $P=0$), suggesting the enrichment in specific mtDNA regions likely originates from dynamic recombination events among both homologous and heterologous acrocentric chromosomes⁴². Additionally, high-quality genomes enable precise identification of 7 mega-NUMTs (maximum length 127.7 kbp), revealing their complex structures.

The conservation of sequence-specific selective pressures across primate evolution, particularly the consistent depletion of mtDL3-derived sequences in fixed NUMTs, reveals universal constraints on mitochondrial-nuclear genome compatibility that may have operated for millions of years (Fig. 3a and Supplementary Fig. 29). This phylogenetic conservation implies that certain molecular mechanisms responsible for NUMT sequence evolution might be fundamental to eukaryotic biology rather than species-specific adaptations. Understanding these constraints could provide additional insights into the cellular processes that govern foreign DNA integration and retention, with implications for both the understanding of natural genome evolution and the design of therapeutic genetic interventions. Interestingly, mtDL3 also exhibits low constraint within the mitochondrial genome⁴³ (Supplementary Fig. 33). This raises the question of whether the observed decreased mtDL3 coverage in fixed NUMTs may be an analytical artifact. Considering (1) mtDL3 coverage is not decreased in polymorphic NUMTs and (2) realignment using multiple mtDNA²⁷ confirms the decrease in fixed NUMTs, we hypothesize that the selected removal of mtDL3 in fixed NUMTs should have a biological cause (Supplementary Fig. 16). However, whether the overlap of low coverage in fixed NUMTs and reduced constraint in the mitochondrial genome is coincidental or biological requires additional mechanistic studies. Moreover, the compromised stability of genomic DNA may also be associated with selective removal, warranting follow-up in future studies.

The precise characterization of NUMTs provides novel insights into the mechanisms underlying their insertion. Previous studies have highlighted the critical role of double-strand break (DSB) repair mechanisms, particularly microhomology-mediated end joining (MMEJ) in *de novo* NUMT insertion⁸. In addition to that, our analyses provided substantial direct evidence revealing that segmental duplication of pre-integrated NUMTs serves as a key mechanism driving NUMT expansion across human populations and throughout primate evolution. These findings propose that NUMT dynamics may involve initial *de novo* insertions mediated by mitochondrial DNA transfer to the nucleus, followed by duplication events within nuclear genomes. This process model better elucidates the mechanisms governing NUMT formation and explains the polymorphic variations observed across populations and evolutionary lineages.

Considering the potential genomic impacts NUMT insertion may generate, it is of significant interest to identify the insertion locations and associated gene-related effects. Fixed NUMTs are stably integrated into genomes, predominantly localized in non-functional genomic regions. In contrast, polymorphic NUMTs show greater potential to exert functional consequences. Both prior research and our current analyses indicate that while polymorphic NUMTs may disrupt protein-coding genes and cause diseases^{5,6}, these variants occur infrequently in human populations as a result of negative selection. Common polymorphic NUMTs display heterogeneous allele frequency distributions across populations. Notably, certain NUMT variants exhibit *cis*-eQTL and/or *cis*-sQTL effects and colocalize with lead SNPs in shared linkage disequilibrium (LD) blocks. Importantly, while fine-mapping analysis cannot directly indicate the function of NUMTs, it does not exclude their modulatory effects within these LD blocks. These potential modulatory roles require validation through precise experimental evidence.

The combination of high-quality long-read genomes^{44–46} with advanced whole-genome alignment tools⁴⁷ has enhanced cross-species comparisons among primates, generating higher-resolution lineage-specific NUMT datasets than those reported in prior studies^{11,20,39}. The observed divergence in lineage-specific NUMT insertion rates across species likely stems from different frequencies of mtDNA transfer events and lineage-specific selective pressures, which may be associated with interspecies variations in mitochondrial architecture and reproductive strategies⁴⁸. Moreover, comparative analyses also revealed two NUMT-derived VNTRs, with divergent nuclear-mitochondrial genome constraints likely underlying their formation in these mtDNA-derived segments. Finally, considering cross-species NUMT comparisons offer critical insights into characteristics and mechanisms of NUMT integration, high-quality long-read genomes across the broader primate phylogeny are needed to achieve more comprehensive evolutionary analyses.

Methods

NUMT detection based on pangenome graph

The Minigraph-Cactus (MC) pipeline¹⁵ was used to construct the human draft pangenome using T2T-CHM13v2.0 as the reference genome, followed by post-processing using *vcfbub* (v0.1.0) and *vcfwave*⁴⁹ (v1.0.10). All tools were run with default parameters as detailed in the APG flagship paper.

Non-reference sequences were extracted from the pangenome graph VCF file using *BCFtools*⁵⁰ (v1.21). For comprehensive NUMT detection, both the reference genome T2T-CHM13v2.0 and non-reference pangenome sequences were queried using *BLASTN*^{51,52} (v2.14.0) against 12 mitochondrial genomes from humans and five non-human primates (Supplementary Data 8 and Supplementary Fig. 1). To maintain methodological consistency, the parameter set established in prior studies^{9,27,45} was utilized: `-reward 2 -penalty -3 -gapopen 5 -gapextend 2 -task blastn -max_target_seqs 10000`. BLAST alignments were filtered using an E-value threshold of $<1 \times 10^{-3}$ and sequence identity $\geq 60\%$. Mitochondrial DNA coordinates for all identified NUMTs were standardized to the revised Cambridge Reference Sequence⁵³ (rCRS, NC_012920.1) using *LiftOver*⁵⁴ (v1.2.2) with the corresponding conversion chain files (Supplementary Fig. 1). No minimum sequence length threshold was applied during detection. Spatially proximal (gap <50 bp) polymorphic segments validated on the same haplotype were merged into single complex events and classified as “mega-NUMTs”. Realignment analyses employed two validation strategies: (a) using 1000 distinct human mtDNA queries²⁷ covering variants in the D-loop region across populations, and (b) shifting the start position of each reference sequence (from 0 to 8000 in rCRS) to avoid potential edge effects.

To benchmark the reliability of our pangenome-based approach, we employed an established short-read NUMT detection pipeline⁸.

Candidate insertions were identified based on clusters supported by at least two pairs of discordant reads, with all other parameters left at their default values. We selected a subset encompassing 200 individuals for benchmarking based on the availability of matched, high-coverage Illumina short-read data and used T2T-CHM13 as the reference genome. Since the short-read approach is designed to detect non-reference insertions, comparisons between approaches were restricted to non-reference NUMT. We defined shared detections as those where insertion sites fell within 200 bp flanking regions (upstream 100 bp and downstream 100 bp) and inserted mtDNA sequences exhibited > 90% overlap. Finally, approach-specific calls were cross-validated using long-read sequencing data and genome assemblies. NUMTs detected exclusively by the short-read approach lacked long-read or assembly support and were therefore classified as false positives.

NUMT genotyping and population stratification analysis

PanGenie (v3.0.1) was employed with default parameters to genotype NUMT-associated loci based on the MC pangenome graph with short-read sequencing data from 2504 unrelated individuals in the 1000 Genomes Project. A total of 1043 loci were successfully characterized, while 123 pericentromeric NUMTs were excluded due to short-read limitations, and 13 loci were filtered during PanGenie preprocessing. For each NUMT, allele frequency differences between superpopulations were analyzed using two-tailed Fisher's exact test, followed by Bonferroni correction for multiple testing. Superpopulation-stratified NUMTs were identified through comparisons between a specific superpopulation and other superpopulations, with all adjusted *P* values < 0.05. A genome-wide selection scan was performed using a sliding window approach with a 20-kbp window size and a 5-kbp step size. Nucleotide diversity (π) and fixation index (F_{ST}) were calculated using VCFtools (v0.1.17). Candidate selective sweep regions were defined based on the top 5% of F_{ST} and the bottom 5% of π diversity. Although two superpopulation-stratified NUMTs co-localized with these candidate regions, they were considered unlikely targets of selection due to their low allele frequencies ($AF < 1\%$).

Cell culture

HEK293T cells were maintained in Dulbecco's Modified Eagle's Medium (DMEM; Servicebio, #G4515) supplemented with 10% fetal bovine serum (FBS; Servicebio, #G8003), 1× Penicillin-Streptomycin (Servicebio, #G4003). Cells were cultured at 37 °C with 5% CO₂. CNI iPSCs were derived from the blood of a Chinese donor⁵⁵ and maintained in eTeSR (STEMCELL, #100-1215).

Plasmid construction

We constructed a pMPRA1-nopro-eGFP plasmid using the backbone of pMPRA1 (MiaoLingBio, P31645) (nopro: no promoter). To construct the pMPRA1-CMV-eGFP, pMPRA1-mtDL3-eGFP, and pMPRA1-mtDL5-eGFP plasmids, PCR products were cloned into the pMPRA1-nopro-eGFP backbone linearized with *NcoI* (NEB, #R0193V) using the NEB-uiler HiFi DNA Assembly Master Mix (NEB, #E2621S). To amplify the mtDL3 and mtDL5 sequences, DNA was isolated from CNI iPSCs using FastPure® Cell/Tissue DNA Isolation Mini Kit (Vazyme, #DC102). PCR was performed using Q5 High-Fidelity DNA Polymerase (NEB, #M0491L) (Supplementary Data 11). DH5 α chemically competent cells (Vazyme, #C502) were used for the transformation. The plasmids were sequenced and confirmed with Sanger sequencing (BioSune) and whole plasmid sequencing (mtDL3 & mtDL5, nanopore & NGS, GenScript). EndoFree Maxi Plasmid Kit was used to prepare plasmids for downstream experiments (TIANGEN, #DP117).

Transfection for RT-qPCR and fluorescence imaging

Transfection was performed in HEK293T using Lipomaster 3000 Transfection Reagent (Vazyme, #TL301) and Opti-MEM I Reduced

Serum Medium (Gibco, #31985062) following the Lipomaster 3000 manual. Plasmids were quantified using the Qubit dsDNA Quantification HS Assay Kit (Invitrogen, #Q32851) and the same amount of substance was used for different plasmids. For RT-qPCR, cells were harvested on Day 3 post-transfection. For fluorescence imaging, cells were replated onto coverslips coated with Matrigel (Corning, #354277) on Day 2 post-transfection and imaged on Day 3 post-transfection.

RT-qPCR

Total RNA was extracted using FastPure® Complex Tissue/Cell Total RNA Isolation Kit (Vazyme, #RC113) following the manual. First-strand cDNA was synthesized immediately after RNA extraction using the HiScript® III 1st Strand cDNA Synthesis Kit (+gDNA wiper) (Vazyme, #R312). qPCR was performed using the ChamQ Universal SYBR qPCR Master Mix (Vazyme, #Q711). eGFP primers were used to evaluate the transcriptional activity. HPRT1 was used as the internal reference gene. Three technical replicates were performed for each of three independent transfection replicates. (Supplementary Data 11)

DAPI staining and confocal imaging

Cells were fixed for 30 minutes at 4 °C with Paraformaldehyde Fixative (Servicebio, #G1101), and then washed with PBS (Servicebio, #G4207). Coverslips were mounted on glass slides using Anti-fade Mounting Medium (with DAPI) (Servicebio, #G1407) and stored at 4 °C for the mounting medium to dry. Coverslips were imaged with the DAPI and the GFP channels of a Zeiss LSM 900. For the GFP channel, different laser settings were used for CMV (500 V, 0.2%) and other groups (650 V, 3.0%), because the CMV group, serving as a positive reference, exhibited a much stronger signal. Five pictures were taken for each of three independent transfection replicates (10× objective).

NUMT insertions across repetitive elements and different genomic regions

For the analysis of mtDNA origin distribution patterns, NUMT coverage relative to mtDNA coordinates was calculated as the sum of corresponding mtDNA sequences for each NUMT category and normalized by the maximum coverage across the mitochondrial genome. To further assess the statistical significance of the coverage patterns observed in the mtDL3 region, the average relative coverage of mtDL3 was calculated as the observed value. Empirical *P*-values were calculated using a permutation test by resampling 10,000 sets of random mtDNA intervals matched to mtDL3 for both fixed and polymorphic NUMTs.

For the enrichment analysis of NUMT insertions across repetitive elements, repetitive elements in 200-bp flanking regions of each NUMT were quantified using BEDtools (v2.31.1). The average number of intersected repetitive elements in NUMT flanking regions was designated as the observed value. Enrichment scores were derived from a permutation test by resampling 10,000 sets of random genomic intervals matched in length to the observed NUMTs. In this context, an enrichment score approaching 1 indicates a significant enrichment of repetitive elements relative to the null distribution, whereas a score approaching 0 indicates depletion. The repetitive elements of T2T-CHM13v2.0 were annotated by RepeatMasker⁴⁴ (download from <https://github.com/marbl/CHM13>).

Genomic regions were hierarchically classified to resolve overlapping annotations based on functional priority: protein-coding genes (CDS > UTRs > introns) > non-coding RNA (exons > introns). Any remaining regions were classified as intergenic. The genomic category for each NUMT was annotated using BEDtools (v2.31.1), and the count of NUMTs in each category was quantified as the observed value. Empirical *P*-values were calculated using a permutation test by resampling 10,000 sets of random genomic intervals matched in length to the observed NUMTs. BEDtools (v2.31.1) was used to annotate the genomic region category for each NUMT. Then, NUMT counts

were quantified across genomic region categories and the permutation test was performed for each category. The potential impact of NUMT insertions on protein sequences was predicted using the standard nuclear genetic code. The gene annotations of T2T-CHM13v2.0 are sourced from the NCBI RefSeq database⁴⁴ (downloaded from https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_009914755.1/).

Sequence characteristics of NUMTs

For sequence identity and mutation spectrum analysis, NUMT sequences were aligned with their corresponding mitochondrial DNA sequences (rCRS) using MAFFT^{56,57} (v7.525). Subsequently, sequence identity was calculated and mutation sites were characterized through Biopython⁵⁸. The low-identity tail of polymorphic NUMTs largely comprises sequences embedded within large-scale duplications, insertions, or deletions (defined as NUMTs constituting <50% of the total variant length), representing ancient, pre-existing NUMTs subject to post-insertion genomic events (Supplementary Fig. 22).

Methylation analysis was restricted to APGp1 individuals, as their HiFi reads (sequenced from blood samples, performed on the PacBio Revio platform) preserve the native base-modification signals essential for methylation calling. HiFi reads were aligned to their diploid genome using pbmm2⁵⁹ (v1.16.0). CpG methylation frequencies within 200-bp windows were estimated using pb-CpG-tools⁶⁰ (v3.0.0) and BEDtools (v2.31.1). Only NUMTs longer than 1 kbp were analyzed, as longer sequences provide sufficient CpG sites for reliable statistical analysis. The nuclear genome-wide null distribution was established by dividing the whole genome into 200-bp windows and randomly selecting 10,000 control windows for methylation calculation.

cis-QTL effect analysis of NUMTs

For association analysis between NUMTs and gene transcription, the normalized expression and splicing data from the MAGE project³⁵ (Multi-ancestry Analysis of Gene Expression) were obtained for QTL analysis. NUMT genotypes of 731 MAGE project individuals from the 1000 Genomes Project were used for downstream analysis.

Both *cis*-eQTLs and *cis*-sQTLs were identified using QTLtools^{61–63} (v1.3.1), following the methods described in the MAGE manuscript. For *cis*-eQTL mapping, the inverse-normal-transformed TMM values were regressed onto NUMT genotypes for all variants within 1 Mb upstream and downstream from the NUMT start site with minor allele frequency (MAF) > 0.01. The top five genotype principal components (PCs) calculated from whole-genome SNPs, sample sex, and 60 Probabilistic Estimation of Expression Residuals (PEER) factors were included as covariates³⁵. For *cis*-sQTL mapping, the normalized intron excision ratios were regressed onto the same variant genotypes as in *cis*-eQTL analysis, but with 15 PEER factors used as covariates. Among all NUMT-gene pairs (including 2,817 eNUMT-eGene pairs and 21,973 sNUMT-sGene pairs), significant *cis*-eQTLs and *cis*-sQTLs were defined by the Bonferroni correction threshold of $P < 2 \times 10^{-6}$ (calculated as 0.05/24790). For fine-mapping in each locus with significant *cis*-eQTLs or *cis*-sQTLs, SuSiE⁶⁴ was used to identify SNVs or NUMTs in 95% credible sets.

Comparative genomic analysis of NUMTs across primates

NUMTs were identified in non-human primate genomes using BLASTN (v2.14.0, as detailed in ‘NUMT detection based on pangenome graph’). For non-human primates, NUMTs present in all haplotypes within a species were classified as fixed for downstream comparative analyses. Whole-genome alignment between primate and human genomes was conducted using minimap2⁴⁷ (v2.28) with the following parameters: ‘-x asm20 --eqx -c --cs -K 500 M -k 15 -m 10 -A 1 -B 2 -O2,12 -n2 -g 100 -r 200,100000 --secondary=no -s 1000’. To ensure ortholog specificity and mapping quality, only primary alignments were retained. Next,

interspecies chain files were generated from the alignment BAM files using transanno (v0.4.5) to capture evolutionarily conserved syntenic regions. Comparative analysis was then performed using LiftOver (v1.2.2) to convert the genomic coordinates of NUMTs. The presence or absence of each NUMT insertion was assessed at orthologous loci within these conserved syntenic regions, enabling the characterization of cross-species orthologous and lineage-specific NUMTs. To assess potential TE-mediated expansions, NUMTs in T2T-CHM13 were intersected with a high-confidence dataset of TE transduction events⁶⁵ using BEDtools (v2.31.1), revealing no overlapping loci. The insertion rate was then calculated as the number of lineage-specific NUMTs divided by the estimated divergence time¹⁹.

For recurrent duplication NUMT analysis, LiftOver (v1.2.2) was used with the ‘multiple’ parameter to reprocess duplication NUMT events. Genomic sequences of recurrent duplication NUMT regions were extracted using SAMtools⁵⁰ (v1.16). Cross-species syntenic comparisons across primates were performed using minimap2 (v2.28), with subsequent visualization of synteny relationships using Saffire⁶⁶ and SVbyEye⁶⁷ (v0.99.0).

For phylogenetic analysis of duplication NUMTs, SAMtools (v1.16) was employed to extract the genomic sequences of duplication NUMT segments in primates. Multiple-sequence alignment was then conducted with MAFFT (v7.525), followed by maximum-likelihood phylogenetic tree reconstruction using IQ-TREE⁶⁸ (v2.2.2.6).

NUMT-derived VNTR analysis

Intersection analysis between NUMT coordinates and simple tandem repeat annotations was performed using BEDtools (v2.31.1) to identify NUMT-derived VNTRs. Comparative analysis of tandem repeat copy numbers was performed on genome assemblies from diverse human populations and non-human primates using mummer⁶⁹ (v3.23) and BLASTN (v2.14.0). The simple tandem repeats of T2T-CHM13v2.0 were annotated by TRF⁴⁴ (download from <https://genome.ucsc.edu>).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Long-read human genome assemblies were acquired from 269 individuals across three initiatives: the Asian Pan-Genome Project Phase 1 (APGp1, <https://github.com/Asian-Pan-Genome/APGp1>), the Human Pangenome Reference Consortium Phase 1¹² (HPRCp1, <https://data.humanpangenome.org/assemblies>), and the Human Genome Structural Variation Consortium Phase 3²² (HGSCp3, <https://www.internationalgenome.org/data-portal/data-collection/hgscv3>). Long-read non-human primate genomes were retrieved from NCBI and GitHub^{19,45,46} (Supplementary Data 8). High-coverage short-read Illumina sequencing data from 2,504 unrelated individuals were obtained from the 1000 Genomes Project^{23,24} (IKGP, https://www.internationalgenome.org/data-portal/data-collection/1000genomes_30x). Source data are provided with this paper.

Code availability

Custom scripts used in this study are available at GitHub (https://github.com/LiantingFu/NUMT_Analysis)⁷⁰.

References

- Zhou, W. et al. Somatic nuclear mitochondrial DNA insertions are prevalent in the human brain and accumulate over time in fibroblasts. *PLoS Biol.* **22**, e3002723 (2024).
- Lopez, J. V. et al. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *J. Mol. Evolution* **39**, 174–190 (1994).

3. Ricchetti, M., Tekaia, F. & Dujon, B. Continued colonization of the human genome by mitochondrial DNA. *PLOS Biol.* **2**, e273 (2004).
4. Hazkani-Covo, E., Zeller, R. M. & Martin, W. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLOS Genet.* **6**, e1000834 (2010).
5. Turner, C. et al. Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer. *Hum. Genet.* **112**, 303–309 (2003).
6. Borensztajn, K. et al. Characterization of two novel splice site mutations in human factor VII gene causing severe plasma factor VII deficiency and bleeding diathesis. *Br. J. Haematol.* **117**, 168–171 (2002).
7. Dayama, G., Emery, S. B., Kidd, J. M. & Mills, R. E. The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Res.* **42**, 12640–12649 (2014).
8. Wei, W. et al. Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. *Nature* **611**, 105–114 (2022).
9. Simone, D. et al. The reference human nuclear mitochondrial sequences compilation validated and implemented on the UCSC genome browser. *BMC Genom.* **12**, 517 (2011).
10. Tsuji, J., Frith, M. C., Tomii, K. & Horton, P. Mammalian NUMT insertion is non-random. *Nucleic Acids Res.* **40**, 9073–9088 (2012).
11. Jensen-Seaman, M. I. et al. A comparative approach shows differences in patterns of numt insertion during hominoid evolution. *J. Mol. Evol.* **68**, 688–699 (2009).
12. Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
13. Gao, Y. et al. A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–121 (2023).
14. Taylor, D. J. et al. Beyond the human genome project: the age of complete human genome sequences and pangenome references. *Annu. Rev. Genom. Hum. Genet.* **25**, 77–104 (2024).
15. Hickey, G. et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2024).
16. Garrison, E. et al. Building pangenome graphs. *Nat. Methods* **21**, 2008–2012 (2024).
17. Ebler, J. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
18. Wang, T. et al. The human pangenome project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
19. Mao, Y. et al. Structurally divergent and recurrently mutated regions of primate genomes. *Cell* **187**, 1547–1562.e13 (2024).
20. Uvizl, M. et al. Comparative Genome Microsynteny Illuminates the Fast Evolution of Nuclear Mitochondrial Segments (NUMTs) in Mammals. *Mole. Biol. Evol.* **41**, msad278 (2024).
21. 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 Genom. Bioinform.* **2**, lqaa089 (2020).
22. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
23. Sudmant, P. H. et al. An integrated map of structural variation in 2504 human genomes. *Nature* **526**, 75–81 (2015).
24. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell* **185**, 3426–3440 (2022).
25. Wei, W. et al. Nuclear-mitochondrial DNA segments resemble paternally inherited mitochondrial DNA in humans. *Nat. Commun.* **11**, 1740 (2020).
26. 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).
27. Tao, Y. et al. Comprehensive identification of mitochondrial pseudogenes (NUMTs) in the human telomere-to-telomere reference genome. *Genes* **14**, 2092 (2023).
28. Barak, B. et al. Neuronal deletion of Gtf2i, associated with Williams syndrome, causes behavioral and myelin alterations rescuable by a remyelinating drug. *Nat. Neurosci.* **22**, 1197 (2019).
29. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548 (2019).
30. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
31. Fraser, H. B. et al. Population-specificity of human DNA methylation. *Genome Biol.* **13**, R8 (2012).
32. Loyfer, N. et al. A DNA methylation atlas of normal human cell types. *Nature* **613**, 355–364 (2023).
33. F. C. Lopes, A. Mitochondrial metabolism and DNA methylation: a review of the interaction between two genomes. *Clin. Epigenetics* **12**, 182 (2020).
34. Guitton, R., Nido, G. S. & Tzoulis, C. No evidence of extensive non-CpG methylation in mtDNA. *Nucleic Acids Res.* **50**, 9190–9194 (2022).
35. Taylor, D. J. et al. Sources of gene expression variation in a globally diverse human cohort. *Nature* **632**, 122–130 (2024).
36. Liang, X. et al. Structural snapshots of human pre-60S ribosomal particles before and after nuclear export. *Nat. Commun.* **11**, 3542 (2020).
37. Akhurst, R. J. & Hata, A. Targeting the TGF β signalling pathway in disease. *Nat. Rev. Drug Discov.* **11**, 790–811 (2012).
38. Nagy, Z. et al. Function of RasGRP3 in the formation and progression of human breast cancer. *Mol. Cancer* **13**, 96 (2014).
39. Hazkani-Covo, E. & Graur, D. A comparative analysis of Numt evolution in human and chimpanzee. *Mol. Biol. Evol.* **24**, 13–18 (2007).
40. Pickrell, J. K. et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19**, 826–837 (2009).
41. Huang, Q. et al. Identification of a novel MLPB missense mutation in a chinese griscelli syndrome 3 patient. *Front. Med.* **9**, 896943 (2022).
42. Guarracino, A. et al. Recombination between heterologous human acrocentric chromosomes. *Nature* **617**, 335–343 (2023).
43. Lake, N. J. et al. Quantifying constraint in the human mitochondrial genome. *Nature* **635**, 390–397 (2024).
44. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
45. Yoo, D. et al. Complete sequencing of ape genomes. *Nature* **641**, 401–418 (2025).
46. Zhang, S. et al. Integrated analysis of the complete sequence of a macaque genome. *Nature* **640**, 714–721 (2025).
47. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
48. Anderson, M. J. et al. Functional evidence for differences in sperm competition in humans and chimpanzees. *Am. J. Phys. Anthropol.* **134**, 274–280 (2007).
49. Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S. & Prins, P. A spectrum of free software tools for processing the VCF variant call format: vcfliib, bio-vcf, cyvcf2, hts-nim and slivar. *PLOS Comput. Biol.* **18**, e1009123 (2022).
50. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Giga-Science* **10**, giab008 (2021).
51. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
52. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinforma.* **10**, 421 (2009).
53. Andrews, R. M. et al. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23**, 147 (1999).

54. Hinrichs, A. S. et al. The UCSC genome browser database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
55. Yang, C. et al. The complete and fully-phased diploid genome of a male Han Chinese. *Cell Res.* **33**, 745–761 (2023).
56. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
57. Katoh, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
58. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
59. Pacific Biosciences. *pbmm2*, <https://github.com/PacificBiosciences/pbmm2> (2020).
60. Pacific Biosciences. *pb-CpG-tools*, <https://github.com/PacificBiosciences/pb-CpG-tools> (2022).
61. Delaneau, O. et al. A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).
62. Ongen, H. et al. Estimating the causal tissues for complex traits and diseases. *Nat. Genet.* **49**, 1676–1683 (2017).
63. Fort, A. et al. MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets. *Bioinformatics* **33**, 1895–1897 (2017).
64. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **82**, 1273–1300 (2020).
65. Hoyt, S. J. et al. From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).
66. Vollger, M. R. et al. rustybam: a composable toolkit for alignment analysis and visualization with Saffire. *bioRxiv* <https://doi.org/10.64898/2026.02.16.706142> (2026).
67. Porubsky, D. et al. SVbyEye: a visual tool to characterize structural variation among whole-genome assemblies. *Bioinformatics* **41**, btaf332 (2025).
68. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
69. Marçais, G. et al. MUMmer4: A fast and versatile genome alignment system. *PLOS Comput. Biol.* **14**, e1005944 (2018).
70. Fu, L. A long-read human pangenome initiative for comprehensive interpretation of nuclear-embedded mitochondrial DNA. LiantingFu/NUMT_Analysis. *Zenodo* <https://doi.org/10.5281/zenodo.19091229> (2026).
- Jiao Tong University (SJTU) 2030 Initiative (grant no. WH510363003/016), SJTU Global Initiative Fund (Type-B), the Computational Biology Program of Science and Technology Commission of Shanghai Municipality (grant no. 24JS2840300), Yongxin Youth Award Fund, and Zhongying Young Scholars Program to Y.M.; by the Shanghai Post-doctoral Excellence Program (grant no. 2024338), the Shanghai Rising-Star Program (grant no. 24YF2721800), and the Shanghai Magnolia Talent Plan Pujiang Project (grant no. 24PJA049) to K.M.; and by Shanghai Jiao Tong University Medical-Engineering Interdisciplinary Research Fund (grant no. YG2025QNA47) to X.Y.

Acknowledgements

We thank Nahyun Kong and Dr. Sheng Chih Jin for their suggestions for this study. We thank the HPRC and Primate T2T Consortium for providing the long-read human and great ape genome assemblies. The computations in this study were run on the Siyuan-1 supported by the Center for High Performance Computing at Shanghai Jiao Tong University. This work was supported, in part, by National Key Research and Development Program of China (grant no. 2025YFC3410300), National Natural Science Foundation of China grants (grant no. 32370658), the Scientific Research Innovation Capability Support Project for Young Faculty (grant no. SRICSPYF-ZY2025101), “Shuguang Program” supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission (grant no. 25SG17), Natural Science Foundation of Chongqing, China (grant no. CSTB2024NSCQ-JQX0004), the New Cornerstone Science Foundation through the XPLORER-PRIZE, Shanghai

Author contributions

L.F., K.M. and Y.M. conceived the project; L.F. performed NUMT detection in humans and NUMT evolutionary analysis across primates; L.F., J.C., D.L., S.D., Z.W., H.M., Z.L., N.L., X.Y., Y.S. and K.M. contributed to NUMT genomic feature analysis; J.C., S.D. and Y.S. contributed to QTL analysis of polymorphic NUMTs; D.L., Z.W., H.M., Z.L. and K.M. contributed to the analysis of NUMT selective removal mechanisms; and D.W., C.Y. and G.Z. generated the genome assemblies of APG individuals. L.F., J.C., G.Z., K.M., and Y.M. drafted the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-026-71348-5>.

Correspondence and requests for materials should be addressed to Lianting Fu, Guojie Zhang, Kaiyue Ma or Yafei Mao.

Peer review information *Nature Communications* thanks Zixia Huang, who co-reviewed with Marek Uvizl; Weichen Zhou and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

¹Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Ministry of Education, Shanghai Key Laboratory of Embryo Original Diseases, International Peace Maternity and Child Health Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China. ²Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Ministry of Education, Shanghai Jiao Tong University, Shanghai, China. ³International Institutes of Medicine, Fourth Affiliated Hospital, Zhejiang University School of Medicine, Yiwu, China. ⁴Center of Evolutionary & Organismal Biology, and Women's Hospital at Zhejiang University School of Medicine, Zhejiang University, Hangzhou, China. ⁵School of Medicine, Zhejiang University, Hangzhou, China. ⁶State Key Laboratory of Genome and Multi-omics Technologies, BGI Research, Shenzhen, China. ⁷Guangdong Provincial Key Laboratory of Genome Read and Write, BGI Research, Shenzhen, China. ⁸Zhiyuan College, Shanghai Jiao Tong University, Shanghai, China. ⁹Department of Genetics, Yale School of Medicine, New Haven, CT, USA. ¹⁰Center for Genomic Health, Yale School of Medicine, New Haven, CT, USA. ¹¹Institute of Neuroscience, International Center for Primate Brain Research, Center for Excellence in Brain Science and Intelligence Technology (CEBSIT), Chinese Academy of Sciences, Shanghai, China. ¹²These authors contributed equally: Lianting Fu, Jieyi Chen, Da Lian. ✉ e-mail: fulianting@sjtu.edu.cn; guojiezhang@zju.edu.cn; yukae@sjtu.edu.cn; yafmao@sjtu.edu.cn