Molecular Antiquity Does Not Establish Mitochondrial Origin

Diogo Pratas[1,2,3,4]

[1]Department of Electronics, Telecommunications and Informatics, University of Aveiro, Portugal
[2]Institute of Electronics and Informatics Engineering of Aveiro, Portugal
[3]Intelligent Systems Associate Laboratory, Portugal
[4]Department of Virology, University of Helsinki, Finland

Ancient mitochondrial DNA is commonly authenticated using short fragment lengths and characteristic substitutions near read termini. Cytosine deamination typically produces C-to-T differences toward the 5′ end and the complementary G-to-A pattern toward the 3′ end. These signatures provide evidence that a molecule has undergone the chemical degradation expected in ancient material [1–3]. They do not, however, establish that the molecule originated from the mitochondrial genome.

This distinction is important because an ancient fragment from another genomic source can be genuinely old, carry genuine post-mortem damage and still align to the mitochondrial reference. Nuclear mitochondrial DNA segments (NUMTs) are a particularly relevant source of ambiguity. NUMTs originated through the integration of mitochondrial sequences into the nuclear genome and subsequently evolved as nuclear loci [4, 5]. An endogenous ancient NUMT fragment is authentic ancient nuclear DNA rather than modern contamination, although modern nuclear contamination may also contribute NUMT-derived reads.

The underlying concern is well established. Nuclear insertions were recognized early as difficult to distinguish from damaged or authentic organellar sequences in ancient DNA studies [6]. Analyses of museum and ancient specimens showed that nuclear copies can be recovered instead of the intended mitochondrial target, especially when DNA is scarce or degraded [7], while broader methodological work proposed experimental and computational strategies for preventing NUMT pollution of mitochondrial datasets [8]. In modern human sequencing, NUMTs are also known to bias mapping and mitochondrial heteroplasmy calls [5, 9].

More recently, de Flamingh et al. [10] introduced *Numt Parser*, which directly compares reads with cytoplasmic mitochondrial and characterized NUMT references and demonstrated improved mitogenome reconstruction in two ancient Cape lions. That study provides an important read-level precedent, but addresses known *Panthera* NUMTs and empirical historical specimens. In a related but biologically distinct setting, Feuerborn et al. [11] used competitive mapping to identify human contamination in ancient faunal genomic datasets. Dolenz et al. [12] further showed that reference bias and spurious ancient DNA mappings depend strongly on fragment length, alignment algorithm and mapping-quality threshold. The present note does not claim the first recognition of NUMT contamination or the first use of competitive mapping. Its contribution is an explicit human, reference-scale stress test using the complete T2T nuclear assembly, ultrashort reads down to 25 bp, controlled terminal damage and a quantitative comparison between mitochondrial-only and origin-aware mapping.

To characterize the available nuclear competitors, the human mitochondrial reference, namely the revised Cambridge Reference Sequence (rCRS) (NC_012920.1), was aligned against the nuclear component of the T2T-CHM13v2.0 assembly (GCF_009914755.1). The mitochondrial record was excluded from the nuclear reference. BLASTN was run with a 7-bp word size, with DUST filtering

and soft masking disabled, and an E-value threshold of 10^{-3}. Local alignments with at least 70% identity and a minimum aligned length of 30 bp were retained and merged when separated by no more than 20 bp. This procedure produced 691 candidate NUMT regions. T2T-CHM13v2.0 provides a more complete representation of mitochondrial-like nuclear regions than earlier human assemblies [13, 14]. The candidate regions varied substantially in length and mitochondrial identity (Figure 1); long, highly similar regions are expected to generate the greatest ambiguity because they can attract short reads across extended mitochondrial intervals.

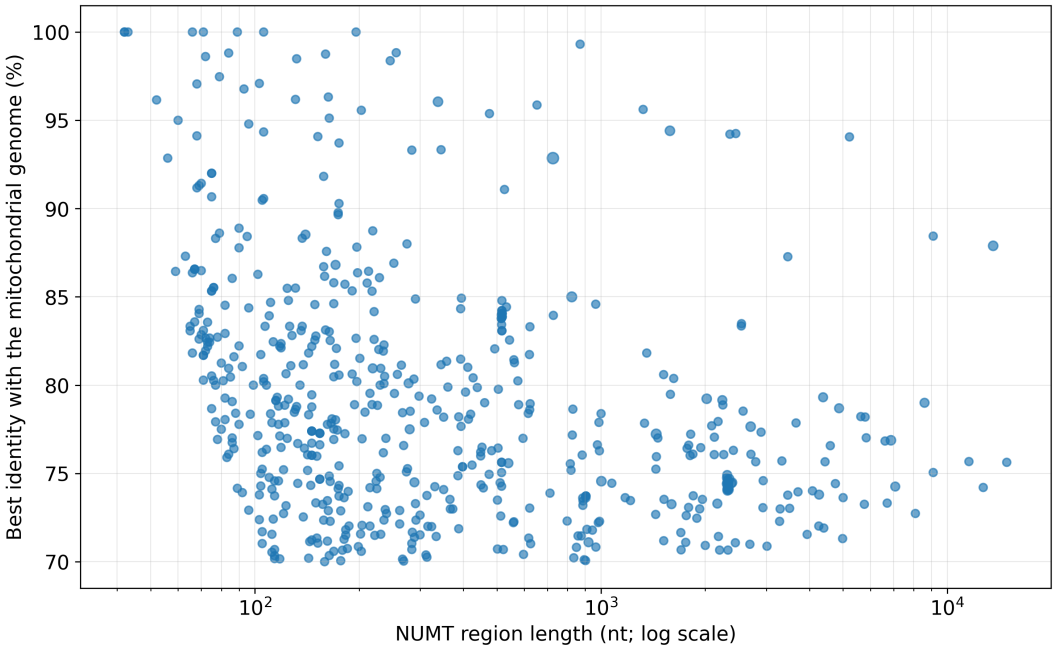

Figure 1: Length and mitochondrial identity of candidate NUMTs in T2T-CHM13v2.0. Long regions with high mitochondrial identity represent the most plausible competitors during short-read mapping.

The reference substitution spectrum between the mitochondrial reference and the candidate NUMTs is shown in Figure 2. Here, X-to-Y denotes base X in the mitochondrial reference aligned to base Y in the NUMT. Reference C-to-T and G-to-A differences are particularly relevant because NUMT-derived reads carrying these alleles generate the same mismatch classes commonly inspected during ancient DNA authentication. In this case, the mismatch reflects inherited NUMT divergence rather than post-mortem modification. Although reference substitutions are not expected to show the systematic terminal decay profile of biochemical damage, an ultrashort read may contain a reference-state difference near an end by chance. At low depth, a small number of such reads can distort an apparent terminal profile. Substitution class alone is therefore insufficient to infer either molecular antiquity or genomic origin.

To test whether this sequence-level ambiguity produces actual read-level misassignment, a controlled competitive-mapping benchmark was performed. High-risk NUMT intervals required at least 90% BLAST identity and an aligned length of at least 75 bp. Overlapping high-risk alignments were merged, yielding 41 simulation intervals. As a bacterial negative control, the strongest local alignment between rCRS and *Paracoccus denitrificans* PD1222 chromosome 1 (NC_008686.1) was selected. This alignment spanned 750 nucleotides at 64.667% identity.

Synthetic single-end reads of 25, 30, 35, 50 and 75 bp were sampled from rCRS, the 41 high-risk NUMT intervals and the bacterial region. For every combination of origin, length and damage condition, 50,000 reads were generated, giving 1.5 million reads in total. Damaged reads received terminal C-to-T substitutions at the $5'$ end and G-to-A substitutions at the $3'$ end according to

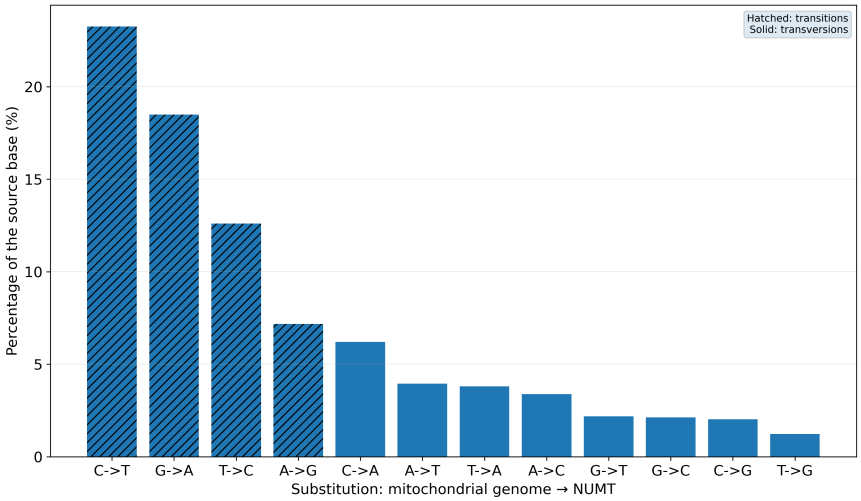

Figure 2: Substitution spectrum between the human mitochondrial reference and T2T-CHM13v2.0 NUMTs. Reference C-to-T and G-to-A differences generate damage-compatible mismatch classes when NUMT reads are aligned to the mitochondrial reference, although inherited NUMT divergence and post-mortem damage remain mechanistically distinct.

$p(d) = 0.3\exp(-d/2)$, where d is the distance from the corresponding read end. Damage was sampled across the first ten terminal positions. An independent sequencing-error probability of 0.001 per base was then applied. The random seed was 20260728.

Three Bowtie 2 reference configurations were constructed by concatenating the relevant reference sequences [15]: (i) mtDNA alone; (ii) mtDNA plus the complete T2T nuclear genome; and (iii) mtDNA plus the complete T2T nuclear genome and the bacterial chromosome. Reads were aligned in end-to-end very-sensitive mode, allowing one seed mismatch, with a 15-bp seed and `-score-min L,-0.6,-0.6`. A read was classified as ambiguous when the primary and second-best alignment scores differed by no more than three. MAPQ ≥ 30 was used as a secondary high-confidence criterion. The experiment is a targeted stress test of high-risk NUMTs and one selected bacterial region; it is not an estimate of their prevalence in archaeological libraries.

As depicted in Figure 3, mitochondrial-only mapping produced extensive false assignment of NUMT reads. Across fragment lengths, 81.698–89.002% of undamaged NUMT reads and 75.946–86.476% of damaged NUMT reads were assigned to mtDNA. At 25 bp, the corresponding rates were 81.698% without damage and 75.946% with damage. Even under the secondary MAPQ ≥ 30 criterion, 26.346–48.238% of undamaged NUMT reads and 18.888–40.052% of damaged NUMT reads remained high-confidence mitochondrial assignments. Terminal damage therefore did not resolve origin: damaged nuclear fragments were still readily accepted by a mitochondrial-only reference.

Adding the complete nuclear genome reduced false mitochondrial assignment to 0.008–0.036% for undamaged NUMT reads and 0.388–0.810% for damaged NUMT reads. Relative to mitochondrial-only mapping, this corresponds to reductions of 99.959–99.990% and 99.055–99.489%, respectively. Most affected reads were reassigned to the nuclear genome or labelled ambiguous because the mitochondrial and nuclear scores were similar. The residual damaged NUMT assignments show that origin-aware mapping greatly reduces, but does not eliminate ambiguity under all combinations of divergence, damage and sequencing error.

This gain in specificity incurred an expected sensitivity cost. Competitive mapping retained 67.894–92.076% of undamaged true mitochondrial reads and 64.332–88.622% of damaged true mitochondrial reads, with retention increasing consistently from 25 to 75 bp. At 25 bp, retention

was 67.894% without damage and 64.332% with damage; at 75 bp it increased to 92.076% and 88.622%, respectively. Most of the reduction resulted from ambiguity rather than confident nuclear assignment. At 25 bp, 32.066% of undamaged mitochondrial reads and 31.494% of damaged mitochondrial reads were classified as ambiguous. These reads should not automatically be interpreted as incorrectly rejected mitochondrial molecules: many ultrashort sequences genuinely lack sufficient information to discriminate between homologous mitochondrial and nuclear loci. The result therefore exposes a sensitivity–specificity trade-off that mitochondrial-only mapping conceals. Adding the bacterial chromosome to the mitochondrial-plus-nuclear reference had a negligible effect on mitochondrial and NUMT assignment rates, indicating that the nuclear competitor accounted for the dominant ambiguity in this targeted benchmark.

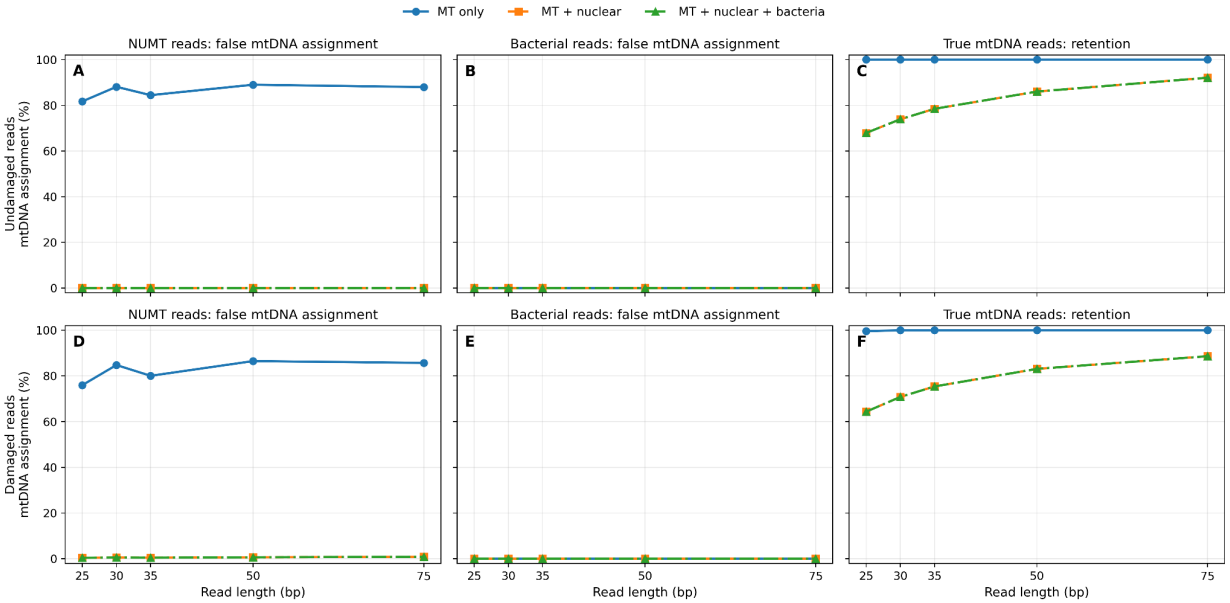

Figure 3: Read-level competitive-mapping benchmark for mitochondrial-origin assessment. For every origin, length and damage condition, 50,000 single-end reads of 25, 30, 35, 50 and 75 bp were simulated from the human mitochondrial reference, 41 high-risk T2T-CHM13v2.0 NUMT intervals and the strongest local human mtDNA–*Paracoccus denitrificans* alignment. Reads were generated without damage (top row) or with terminal C-to-T and G-to-A damage (bottom row), then mapped against mtDNA alone, mtDNA plus the complete T2T nuclear genome, or mtDNA plus the nuclear genome and bacterial chromosome. The panels report the percentage assigned to mtDNA. Similar primary and secondary scores ($\Delta \leq 3$) were classified as ambiguous. NUMT and bacterial simulations are targeted stress tests, not prevalence estimates.

The bacterial result provides an informative negative control. The strongest mtDNA–*P. denitrificans* local alignment contained C-to-T and G-to-A mismatch classes, but generated no appreciable mitochondrial misassignment under the tested conditions. In mitochondrial-only mapping, one of 50,000 undamaged bacterial reads of 25 bp (0.002%) and three of 50,000 damaged bacterial reads of 25 bp (0.006%) were assigned to mtDNA. No bacterial read of 30, 35, 50 or 75 bp was assigned to mtDNA. Thus, only four of 500,000 bacterial reads were assigned to mtDNA, and none met the MAPQ ≥ 30 high-confidence criterion. When the bacterial chromosome was included in the competitive reference, no bacterial read was assigned to mtDNA. The presence of damage-compatible substitution classes is therefore not sufficient to establish a practical mapping risk. A broader microbial screen selected using maximum short-window identity, rather than only the strongest long local alignment, would be required before drawing general conclusions about microbial competitors.

Taken together, these results separate two inferential questions that are often conflated. Terminal

deamination supports molecular antiquity, whereas mitochondrial origin requires an independent origin-aware comparison. Competitive mapping is one effective approach, but it is not the only possible source of origin evidence. Longer fragments, paired-end information, split alignments, nuclear flanking sequence, experimentally characterized NUMTs and direct read-comparison methods such as *Numt Parser* may also contribute. Damage estimates should be calculated only after origin filtering, and reads with near-equivalent mitochondrial and nuclear scores should be retained as ambiguous rather than forced into a mitochondrial assignment.

A robust workflow should therefore align reads against complete mitochondrial and nuclear references, preserve competing alignment scores and evaluate sensitivity to fragment-length, alignment and MAPQ thresholds. This recommendation complements, rather than replaces, earlier NUMT-filtering strategies [8, 10], competitive-contaminant mapping [11] and broader assessment of ancient DNA mapping bias [12]. After origin-aware filtering, the reconstructed mitogenome should also be evaluated for coverage stability, contamination and endogenous mitochondrial consensus quality [16], as well as for phylogenetic placement and temporal consistency. Ancient mitogenomes have been used as calibration points for the human mitochondrial clock [17]; such applications require particular caution because mapping artefacts can propagate into tree topology and age estimation.

Metagenomic classification may provide complementary evidence when environmental or laboratory DNA is abundant. FALCON2 and Kraken 2, for example, can highlight unexpected taxa before consensus reconstruction [18, 19]. Such classifications are not definitive for ultrashort reads and should guide, rather than replace, competitive alignment and manual validation. This multi-signal interpretation is consistent with ancient paleovirology, where molecular damage, taxonomic assignment and evolutionary context must be evaluated jointly [20].

The current study has several limitations. T2T-CHM13v2.0 is a largely haploid composite reference, comprising chromosomes 1–22 and X from T2T-CHM13v2.0 and chromosome Y from HG002 [21], and therefore does not capture the full population-level diversity of polymorphic NUMTs documented across human populations [5]. The simulations deliberately sampled high-risk NUMT intervals and therefore quantify vulnerability under an adverse scenario rather than real-library prevalence. They also use one aligner and one parameterization; mapping behaviour can change with algorithm, MAPQ threshold and fragment length [12]. The damage model is intentionally simplified and does not reproduce the full complexity of empirical ancient DNA libraries, including fragment-length distributions, base-quality profiles, strand asymmetry, library-preparation effects, contamination mixtures and genomic coverage heterogeneity. Finally, the single bacterial reference produced only four low-confidence mitochondrial assignments and should be treated as an exploratory negative control, not as evidence against a general microbial-confounding mechanism.

The central conclusion is that a molecule may be genuinely ancient and genuinely damaged while still being assigned to the wrong genomic compartment. Reliable ancient mitogenome reconstruction should distinguish (i) whether a molecule is compatible with ancient DNA, (ii) where that molecule originated and (iii) how much uncertainty accompanies each inference. Post-mortem damage supports molecular antiquity; mitochondrial origin must be evaluated independently through competitive or otherwise origin-aware analysis.

Code and data availability

The source code, complete benchmark pipeline, simulation parameters and analysis scripts are available in the AMITO repository at `https://github.com/cobilab/amito`.

References

[1] Susanna Sawyer, Johannes Krause, Katerina Guschanski, Vincent Savolainen, and Svante Pääbo. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE*, 7(3):e34131, 2012. doi: 10.1371/journal.pone.0034131. URL https://doi.org/10.1371/journal.pone.0034131.

[2] Jesse Dabney, Matthias Meyer, and Svante Pääbo. Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology*, 5(7):a012567, 2013. doi: 10.1101/cshperspect.a012567. URL https://doi.org/10.1101/cshperspect.a012567.

[3] Hákon Jónsson, Aurélien Ginolhac, Mikkel Schubert, Philip L. F. Johnson, and Ludovic Orlando. mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, 29(13):1682–1684, 2013. doi: 10.1093/bioinformatics/btt193. URL https://doi.org/10.1093/bioinformatics/btt193.

[4] Einat Hazkani-Covo, R. Martin Zeller, and William Martin. Molecular poltergeists: Mitochondrial DNA copies (NUMTs) in sequenced nuclear genomes. *PLoS Genetics*, 6(2):e1000834, 2010. doi: 10.1371/journal.pgen.1000834. URL https://doi.org/10.1371/journal.pgen.1000834.

[5] Wei Wei, Katherine R. Schon, Greg Elgar, Andrea Orioli, Melanie Tanguy, Adam Giess, Marc Tischkowitz, Mark J. Caulfield, Patrick F. Chinnery, et al. Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. *Nature*, 611:105–114, 2022. doi: 10.1038/s41586-022-05288-7. URL https://doi.org/10.1038/s41586-022-05288-7.

[6] M. Thomas P. Gilbert, Hans-Jürgen Bandelt, Michael Hofreiter, and Ian Barnes. Assessing ancient DNA studies. *Trends in Ecology & Evolution*, 20(10):541–544, 2005. doi: 10.1016/j.tree.2005.07.005. URL https://doi.org/10.1016/j.tree.2005.07.005.

[7] Robert-Jan den Tex, Jesus E. Maldonado, Richard Thorington, and Jennifer A. Leonard. Nuclear copies of mitochondrial genes: Another problem for ancient DNA. *Genetica*, 138 (9–10):979–984, 2010. doi: 10.1007/s10709-010-9481-9. URL https://doi.org/10.1007/s10709-010-9481-9.

[8] Sébastien Calvignac, Lara Konecny, Florian Malard, and Christophe J. Douady. Preventing the pollution of mitochondrial datasets with nuclear mitochondrial paralogs (NUMTs). *Mitochondrion*, 11(2):246–254, 2011. doi: 10.1016/j.mito.2010.10.004. URL https://doi.org/10.1016/j.mito.2010.10.004.

[9] Hannah Maude, Mira Davidson, Natalie Charitakis, Leo Diaz, William H. T. Bowers, Eva Gradovich, Toby Andrew, and Derek Huntley. NUMT confounding biases mitochondrial heteroplasmy calls in favor of the reference allele. *Frontiers in Cell and Developmental Biology*, 7:201, 2019. doi: 10.3389/fcell.2019.00201. URL https://doi.org/10.3389/fcell.2019.00201.

[10] Alida de Flamingh, Angel G. Rivera-Colón, Tom P. Gnoske, Julian C. Kerbis Peterhans, Julian Catchen, Ripan S. Malhi, and Alfred L. Roca. Numt Parser: Automated identification and removal of nuclear mitochondrial pseudogenes (NUMTs) for accurate mitochondrial genome reconstruction in Panthera. *Journal of Heredity*, 114(2):120–130, 2023. doi: 10.1093/jhered/esac065. URL https://doi.org/10.1093/jhered/esac065.

[11] Tatiana R. Feuerborn, Eleftheria Palkopoulou, Tom van der Valk, Johanna von Seth, Arielle R. Munters, Patrícia Pečnerová, Marianne Dehasque, Irene Ureña, Erik Ersmark, Vendela Kempe Lagerholm, Maja Krzewińska, Ricardo Rodríguez-Varela, Anders Götherström, Love Dalén, and David Díez-del Molino. Competitive mapping allows for the identification

and exclusion of human DNA contamination in ancient faunal genomic datasets. *BMC Genomics*, 21:844, 2020. doi: 10.1186/s12864-020-07229-y. URL https://doi.org/10.1186/s12864-020-07229-y.

[12] Stephanie Dolenz, Tom van der Valk, Love Dalén, and Peter D. Heintzman. Unravelling reference bias in ancient DNA datasets. *Bioinformatics*, 40(7):btae436, 2024. doi: 10.1093/bioinformatics/btae436. URL https://doi.org/10.1093/bioinformatics/btae436.

[13] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, et al. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022. doi: 10.1126/science.abj6987. URL https://doi.org/10.1126/science.abj6987.

[14] Yichen Tao, Chengpeng He, Deng Lin, Zhenglong Gu, and Weilin Pu. Comprehensive identification of mitochondrial pseudogenes (NUMTs) in the human telomere-to-telomere reference genome. *Genes*, 14(11):2092, 2023. doi: 10.3390/genes14112092. URL https://doi.org/10.3390/genes14112092.

[15] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012. doi: 10.1038/nmeth.1923. URL https://doi.org/10.1038/nmeth.1923.

[16] Gabriel Renaud, Viviane Slon, Ana T. Duggan, and Janet Kelso. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biology*, 16:224, 2015. doi: 10.1186/s13059-015-0776-0. URL https://doi.org/10.1186/s13059-015-0776-0.

[17] Qiaomei Fu, Alissa Mittnik, Philip L. F. Johnson, Kirsten Bos, Martina Lari, Ruth Bollongino, Chengkai Sun, Liane Giemsch, Ralf Schmitz, Joachim Burger, Anna Maria Ronchitelli, Fabio Martini, Renata G. Cremonesi, Jiří Svoboda, Peter Bauer, David Caramelli, Sergi Castellano, David Reich, Svante Pääbo, and Johannes Krause. A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology*, 23(7):553–559, 2013. doi: 10.1016/j.cub.2013.02.044. URL https://doi.org/10.1016/j.cub.2013.02.044.

[18] Luís L. Marques, Armando J. Pinho, and Diogo Pratas. FALCON2: Compression-based metagenomic classification of ancient viruses. *Bioinformatics*, 42(5):btag155, May 2026. doi: 10.1093/bioinformatics/btag155. URL https://doi.org/10.1093/bioinformatics/btag155.

[19] Derrick E. Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1):257, 2019. doi: 10.1186/s13059-019-1891-0. URL https://doi.org/10.1186/s13059-019-1891-0.

[20] Toni de Dios, Christiana L. Scheib, and Charlotte J. Houldcroft. An adagio for viruses, played out on ancient DNA. *Genome Biology and Evolution*, 15(3):evad047, 2023. doi: 10.1093/gbe/evad047. URL https://doi.org/10.1093/gbe/evad047.

[21] Arang Rhie, Sergey Nurk, Monika Cechova, Savannah J. Hoyt, Dylan J. Taylor, Nicolas Altemose, Paul W. Hook, Sergey Koren, Mikko Rautiainen, Ivan A. Alexandrov, et al. The complete sequence of a human Y chromosome. *Nature*, 621(7978):344–354, 2023. doi: 10.1038/s41586-023-06457-y. URL https://doi.org/10.1038/s41586-023-06457-y.