NUMT PARSER: automated identification and removal of nuclear mitochondrial pseudogenes (numts) for accurate mitochondrial genome reconstruction in *Panthera*

Alida de Flamingh*[1,2], Angel G. Rivera-Colón*[3], Tom P. Gnoske[4], Julian C. Kerbis Peterhans[4,5,6], Julian Catchen [2,3], Ripan S. Malhi[1,2,7] & Alfred L. Roca[1,2,8]

* Equal contributing first authors

Corresponding author/s: Alida de Flamingh & Angel G. Rivera-Colón

[1] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign (UIUC), Urbana IL, 61801, USA

[2] Program in Ecology, Evolution and Conservation Biology, UIUC, Urbana IL, 61801, USA

[3] Department of Ecology, Evolution, and Behavior, UIUC, Urbana IL, 61801, USA

[4] Field Museum of Natural History (FMNH), Chicago, IL, 60605, USA

[5] Science & Education, Field Museum of Natural History, Chicago, IL, 60605, USA

[6] College of Arts & Science, Roosevelt University, Chicago, IL, 60605, U.S.A.

[7] Department of Anthropology, UIUC, Urbana IL, 61801, USA

[8] Department of Animal Sciences, UIUC, Urbana IL, 61801, USA

Abstract:

Nuclear mitochondrial pseudogenes (numts) may hinder the reconstruction of mtDNA genomes and affect the reliability of mtDNA datasets for phylogenetic and population genetic comparisons. Here, we present the program NUMT PARSER, which allows for the identification of DNA sequences that likely originate from numt pseudogene DNA. Sequencing reads are classified as originating from either numt or true cytoplasmic mitochondrial (cymt) DNA by direct comparison against cymt and numt reference sequences. Classified reads can then be parsed into cymt or numt datasets. We tested this program using whole genome shotgun-sequenced data from two ancient Cape lions (*Panthera leo*), because mtDNA is often the marker of choice for ancient DNA studies and the genus *Panthera* is known to have numt pseudogenes. NUMT PARSER decreased sequence disagreements that were likely due to numt pseudogene contamination and equalized read coverage across the mitogenome by removing reads that likely originated from numts. We compared the efficacy of NUMT PARSER to two other bioinformatic approaches that can be used to account for numt contamination. We find that NUMT PARSER outperformed approaches that rely only on read alignment or Basic Local Alignment Search Tool (BLAST) properties, and was effective at identifying sequences that likely originated from numts while having minimal impacts on the recovery of cymt reads. NUMT PARSER therefore improves the reconstruction of true mitogenomes, allowing for more accurate and robust biological inferences.

Keywords: Cape lion; mitogenome assembly; read filter; cymt; *Panthera leo melanochaitus*

Introduction:

Mitochondrial DNA is often used for phylogenetic studies that seek to investigate matrilineal inheritance patterns (Chaitanya *et al.*, 2014), inter- and intra-specific divergences (Cronin *et al.*, 1991; Gill *et al.*, 1993; Bowers *et al.*, 1994), and for studies that use samples with low DNA copy numbers (Hofreiter, Serre, *et al.*, 2001; Merheb *et al.*, 2019). However, the presence of nuclear mitochondrial (numt) pseudogenes (designated as *Numt* by Lopez et al (Lopez *et al.*, 1994)) may hinder the identification of true cytoplasmic mitochondrial (cymt) DNA sequences and the reliability of mtDNA for phylogenetic and population genetic comparisons (Bensasson *et al.*, 2001; Smart *et al.*, 2019). Numts arise when mitochondrial DNA is incorporated into the nuclear genome during chromosomal double-strand break repair by nonhomologous recombination (Blanchard and Schmidt, 1995, 1996; Bensasson *et al.*, 2001). Organellar DNA fragments are found in the nuclear genomes of many eukaryotes (Bensasson *et al.*, 2001; Gaziev and Shaikhaev, 2010), mostly in non-coding intergenic regions and introns (Bensasson *et al.*, 2001; Gaziev and Shaikhaev, 2010; Smart *et al.*, 2019). Numts may vary in size and sequence depending on the mitochondrial DNA that is integrated into the nuclear genome during double stranded break repair. The nuclear genome integration sites also vary among taxa, with most numts occurring as single copies at dispersed genomic locations (Blanchard and Schmidt, 1995; Zischler, 2000). Some taxa have tandemly repeated numt sequences (e.g., domestic cats (Lopez *et al.*, 1994)), while in other taxa numts may be present in telomeric, centromeric and/or other regions of their nuclear genomes (e.g., grasshoppers (Vaughan *et al.*, 1999)).

Primers used to amplify targeted regions of the mitogenome through PCR may also amplify numt pseudogenes, and in some cases may even preferentially amplify numts compared to cymts (Collura and Stewart, 1995). We, and others (e.g., (Bensasson *et al.*, 2001; Goios *et al.*, 2008; Song *et al.*, 2008)), use the word *contaminate* to refer to mtDNA datasets where the reads stem from both cymt and numt DNA templates, and where mtDNA datasets are thus contaminated by reads that originate outside of the mitochondrion. The numt contaminants may show up as “ghost” bands or additional bands in electrophoresis gels post-PCR-amplification (Den Tex *et al.*, 2010), or as sequence ambiguities and nonsense mutations in sequenced DNA, and may result in unexpected phylogenetic placements (Zhang and Hewitt, 1996; Triant and Dewoody, 2009). When amplifying whole genome DNA using high-throughput sequencing technologies that do not target specific genomic regions, mitogenome regions with paralogous numt pseudogenes might be over-represented in a sequence read pool. Numt dataset contamination may therefore be evident when aligning data obtained from whole genome DNA, combining nuclear and mitochondrial reads, to a reference mitogenome. Assuming that DNA amplification is uniform across the genome, then mitogenome regions with corresponding numt pseudogenes will appear to have higher read coverage compared to the

mitogenome-wide average, since those regions will be represented in the contaminated dataset by both cymt and numt DNA reads.

Contamination by numts is especially problematic for taxa that have many copies of one or more numt pseudogenes (e.g., cats and other felids have a tandemly repeated 7.9 kilobase pair (kb) numt pseudogene (Lopez *et al.*, 1994)). The higher the numt copy number, or number of tandem repeats, the more the pseudogene will be represented in the shotgun sequence read pool. Taxa that have many numt repeats are therefore especially prone to dataset contamination and incorrect mitogenome characterization.

The unknown presence of numt pseudogenes can lead to incorrect but seemingly convincing phylogenetic results. When numt sequences from one taxon are compared to cymt sequences from other taxa, the resulting phylogenetic inference may not be accurate since the sequences that are being compared are not orthologous. Numt contamination has been a challenge for systematics and phylogenetic inferences of contemporary samples (Sorenson and Quinn, 1998; Allende *et al.*, 2001; Pereira and Baker, 2004; Antunes and Ramos, 2005; Podnar *et al.*, 2007), and has also been especially problematic for studies that use ancient DNA (Woodward *et al.*, 1994; van der Kuyl *et al.*, 1995; Zischler, Hoss, *et al.*, 1995) prior to the implementation of ancient DNA validation procedures (Cooper and Poinar, 2000; Gilbert *et al.*, 2005).

Here, we developed and tested a program, NUMT PARSER, which allows for the identification and filtering of DNA sequences that likely originate from numts in short-read sequencing datasets. Sequencing reads are classified as putatively originating from either cymt or numt DNA by direct comparison against cymt and numt reference sequences. Classified reads can then be parsed into separate datasets that contain putative cymt or numt derived reads. We tested NUMT PARSER using whole genome shotgun-sequenced data from two ancient Cape lions (*Panthera leo*). We focused our analysis on a species within the genus *Panthera* as many taxa within this genus are known to have numt pseudogenes (see section "Numt contamination in *Panthera*"). We also specifically tested the efficacy of NUMT PARSER on ancient DNA samples that likely contain degraded DNA, as mtDNA is often the marker of choice for ancient DNA studies, and because past studies have found numt pseudogene contamination to be common among ancient DNA samples (Den Tex *et al.*, 2010). We compared NUMT PARSER to two alternative bioinformatic approaches that may be used to account for numt contamination. We show that NUMT PARSER outperforms other bioinformatic approaches and is effective at identifying sequences that likely originate from numt pseudogene DNA while having minimal impacts over the recovery of true cymt reads.

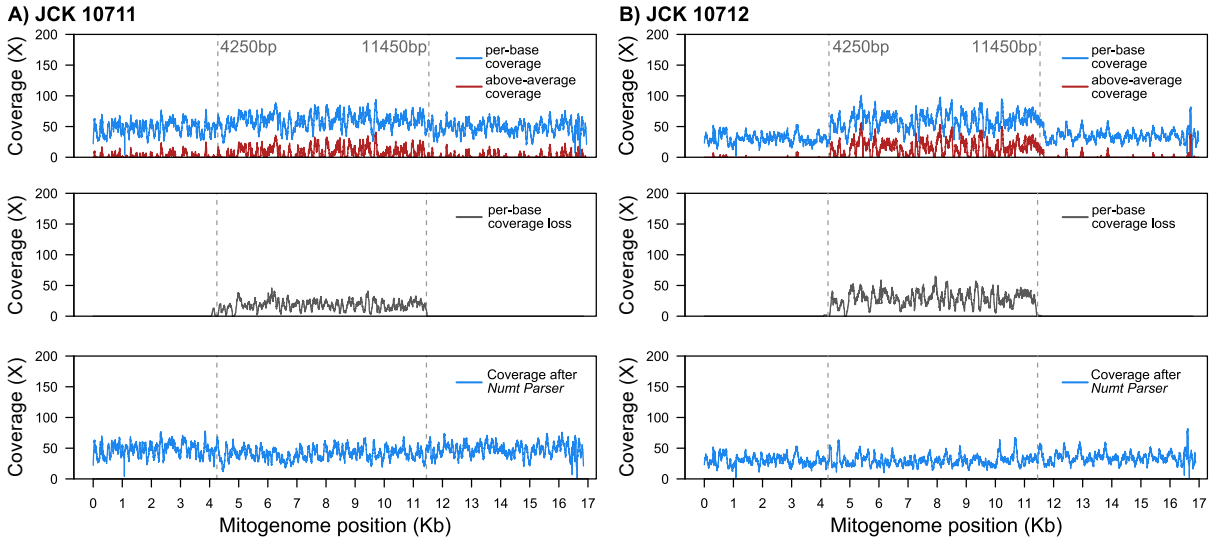

Figure 1. Read coverage across the Cape lion mitogenomes for JCK 10711 (A - left panels) and JCK 10712 (B – right panels). Top panels in A and B show per base coverage (blue) and above-average coverage (red) before NUMT PARSER filtering. Vertical grey dashed lines flank the 7,200 bp region for which a numt paralog exists. Middle panels show coverage loss across the mitogenome due to NUMT PARSER filtering. Lower panels show read coverage across the mitogenome after NUMT PARSER filtering. The decrease in read coverage in the region corresponding to the 7,200bp numt is most evident for JCK 10712.

Methods:

Ancient DNA extraction

We collected and analyzed bone and tooth samples from two Cape lion specimens currently housed at the Field Museum of Natural History (FMNH) in Chicago, USA. Cape lions are thought to have gone extinct in ca. 1850 (Van Bree, 1998). The two skulls studied in this paper were from the colloquially named "Cape Flats" region of South Africa, near the current city of Cape Town on the south-western tip of southern Africa (also see co-publication, de Flamingh et al XXX).

In keeping with their current FMNH designation, the Cape lion samples are listed as JCK 10711 and JCK 10712. Samples were collected at the FMNH following a protocol that minimizes contamination by non-target DNA sources. Samples were collected in a dedicated workspace and with equipment that were decontaminated with Takara Bio DNA-off prior to collection of each sample. Disposable personal protective equipment (lab coat, hair net, gloves, cover sleeves) were changed and decontaminated with Takara Bio DNA-off between collection events. The surface of the collection location on each lion specimen skull or tooth was decontaminated with 6% sodium hypochlorite (full strength Clorox bleach) and rinsed with DNA-free ddH_2O. We collected approximately 0.2 g of bone or tooth powder using a Dremel hand drill that was decontaminated with Takara Bio DNA-Off prior to collection, using a new with a 1mm bit for each collection event.

DNA was extracted in the Malhi ancient DNA laboratory at the Carl. R. Woese Institute for Genomic Biology at the University of Illinois at Urbana-Champaign. The Malhi ancient DNA laboratory is an isolated, air-filtered facility that is dedicated to molecular analysis of ancient and low-template DNA samples. Each tooth and/or bone sample was incubated under rotation in 4 ml of digestion buffer (0.5 M EDTA, 33.3 mg/ml proteinase K, 10% N-lauryl sarcosine) for 12–24 hours at 37°C. Using Amicon K4 centrifugal units, we then concentrated the sample to approximately 250 µl, and used this concentrate as starting template for the DNA extraction using a Qiagen PCR Purification Kit with a final DNA elution volume of 60 µl. Whole genomic libraries were constructed using the NEBNext® Ultra II™ DNA Library Prep kit and NEBNext® Multiplex Oligos (Unique Dual Indexes) for Illumina®. The extracted DNA was pre-treated with USER (Uracil-Specific Excision Reagent) enzyme to remove cytosine to uracil nucleotide base changes that are common in ancient DNA (Supplementary file 1 – Step #7A; (Hofreiter, Serre, *et al.*, 2001)).

All samples and negative libraries were pooled, and shotgun sequenced on an Illumina NovaSeq 6000 S1 1x100bp flow cell at the Roy J. Carver Biotechnology Center at the University of Illinois at Urbana-Champaign (see Supplementary file 1 for a step-by-step description of ancient DNA extraction and library preparation protocols).

Bioinformatic preparation of NUMT PARSER input files

Samples were de-multiplexed and were merged into a single datafile per lion. Sequence reads were trimmed to have a minimum length of 25 bp using `FastP` v.0.19.6 (Chen *et al.*, 2018). We assessed whether DNA showed damage patterns that are characteristic of ancient DNA by aligning trimmed reads to the complete lion genome (Armstrong *et al.*, 2020) using the `mem` module from BWA (Li and Durbin, 2010) and quantifying damage in `mapDamage2` (Jónsson *et al.*, 2013) using a fragment size of 70 bp.

Reads were aligned (Li and Durbin, 2010) to a previously published reference lion mitogenome that was reconstructed using a dataset reported to show no evidence of numt contamination (hereafter referred to as the cymt genome; GenBank accession KP202262; (Li *et al.*, 2016); see Li et al 2016 for details on phylogenetic validation of mitogenome authenticity), and to a numt reference sequence (see below) (Figure 2A). Alignments were converted to BAM format and filtered to remove unmapped reads and alignments with a mapping quality less than 30 using SAMTOOLS `view` v. 1.1 (Li *et al.*, 2009). Filtered alignments were then sorted and indexed, with duplicates marked and removed with the PICARD TOOLKIT v. 2.10.1 ("Picard Toolkit." 2019. Broad Institute). Filtered BAM files were converted to SAM format for compatibility with NUMT PARSER (Figure 2A).

Numt contamination in Panthera

We tested NUMT PARSER on Cape lion samples, because many taxa within the genus *Panthera* are known to have numt pseudogenes. For example, Li et al. (2016) showed that

about half of the published big cat mitogenomes in the genus *Panthera* contained long stretches of high sequence divergence consistent with numt contamination. Lopez et al. (1994) showed that some taxa within the family Felidae have a ~7.9kb numt pseudogene that is tandemly repeated 38-76 times and comprises a macrosatellite with repeats of multiple lengths. Using a 12S mtDNA fragment, Lopez et al. (1994) showed that this numt originated before the divergence of modern cat species within the genus *Felis*. Therefore, the development and testing of NUMT PARSER using specifically lion data may be especially pertinent since taxa within Felidae*,* and specifically within the genus *Panthera*, have been reported to have numt dataset contamination and incorrect mitogenome characterization Li et al. (2016).

Defining the numt reference sequence

Li et al. (2016) previously identified a set of nine published felid mitogenomes in GenBank in which the sequences are composites of both cymt and numt sequences. Specifically, they identified a published mitochondrial genome sequence for an African lion (GenBank accession number KF907306) that contains a region with elevated nucleotide divergence when compared to a reference mitogenome used in their study. The region of divergence likely represents a large numt pseudogene spanning a ~7,200 basepair (bp) region of above-average coverage that was also observed across our Cape lion mitogenomes before NUMT PARSER filtering (Figure 3). Mitogenome regions that have numt pseudogene paralogs will have a marked increase in read coverage compared to mitogenome-wide average since those regions will be represented by reads from both cymt and numt sequences. In agreement with the region identified by Li et al. (2016), we therefore defined the numt pseudogene for use in NUMT PARSER as the region of above-average coverage that corresponds to base pair positions 4,250 to 11,450 of the cymt mitogenome (GenBank accession KP202262;(Li *et al.*, 2016)) and to genes *ND1* partial, *ND2*, *COX1*, *COX2*, *ATP8*, *ATP6*, *COX3*, *ND3*, *ND4L*, *ND4* partial (Figure 3A). Our characterized numt pseudogene reference for use in NUMT PARSER is consistent with numt pseudogene regions reported previously for other taxa in the genus *Panthera* (Li *et al.*, 2016).

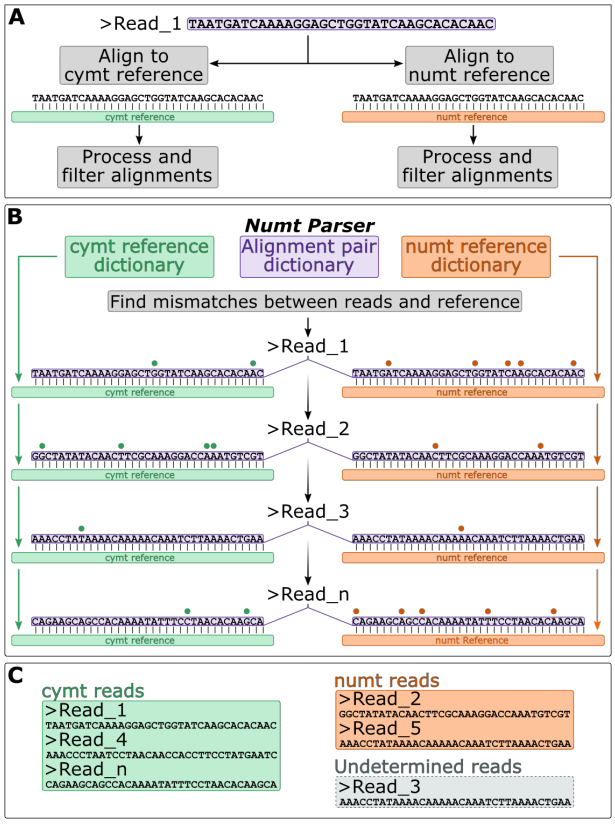

Figure 2. The NUMT PARSER pipeline for the identification and removal of numts. (A) After demultiplexing and trimming, sequencing reads are independently aligned to a mitogenome reference sequence, and to a reference sequence for lion numt. The resulting alignments are subsequently filtered and processed by NUMT PARSER. (B) After alignment, reads are processed by NUMT PARSER. First, the program reads the reference sequences for both cymt (green) and numt (orange) references and corresponding alignments (purple). The program then iterates over the alignment dictionary, processing each read independently. For each read query, NUMT PARSER compares the read sequence against the two references and identifies mismatches between reads and each of the references (shown as dots above read alignments). The program outputs a tally of the number of mismatches observed for each read against both the cymt and numt reference sequences. (C) Using the NUMT PARSER output, the total read pool can then be parsed into datasets containing sequences likely originating from either cymt or numt DNA. Read_n indicates the *n*-th read in the read dataset.

Identifying numt reads with NUMT PARSER

The NUMT PARSER analysis (Figure 2B) starts by loading the reference sequences of both the cymt and numt sequences into a reference sequence dictionary that respectively stores the IDs and sequences as key-value pairs. Read alignments in SAM format against both the cymt and numt reference sequences are also loaded and then processed into a separate alignment pair dictionary. Taking each unique read ID as a key, the alignment pair dictionary stores the mapping information (chromosome, base pair, alignment flags, and CIGAR (Li *et al.*, 2009)) for alignments against both the cymt and numt references. The software then iterates over all reads, comparing them directly to both cymt and numt sequences present in the reference sequence dictionary. Using the corresponding positional information from the alignment, every nucleotide in the read is directly compared against the corresponding site in the reference and a tally is taken of all mismatches between both sequences. NUMT PARSER accounts for the underlying alignment information and thus handles indels and clipped bases, as determined by the CIGAR string, which are also tallied as mismatches. The number of matches (aligned portion minus tallied mismatches) is then divided by the total length of the aligned portion of the read to obtain a percent identity of each read against each reference. The NUMT PARSER output is a table which contains the percent identity, alignment length, and number of mismatches of each

unique read ID against both the cymt and numt references. The output table also contains a classifier that tags each read as putatively originating from cymt DNA (tagged as "cymt") or from numt DNA (tagged as "numt") based on the highest percent identity, or tagged as "Undetermined" when reads have equal percent identity to both the cymt mitogenome and numt pseudogene references (Table 1). Provided that the reference sequences and alignments are available, NUMT PARSER works as a standalone Python program with no requirement of outside software or library dependencies. All required code for running NUMT PARSER is available at https://github.com/adeflamingh/NuMt_parser.

Table 1. Example table produced by NUMT PARSER. This table contains the percent identity, alignment length, and number of mismatches of each unique read ID against both the cymt and numt reference sequences. The output table also contains a classifier that tags each read as putatively originating from cymt or numt based on the highest percent identity or tagged as "undetermined" when reads have equal percentage identity to both the true mitogenome and numt pseudogene references. Read_n indicates the *n*-th read in the dataset.

Read_ID[1]	**mt_ aln_bp[2]**	**mt_ mismatch[3]**	**mt_ identity[4]**	**numt_ aln_bp[5]**	**numt_ mismatch[6]**	**numt_ identity[7]**	**Candidate[8]**
Read_1	100	2	0.980000	97	5	0.948454	cymt
Read_2	74	2	0.972973	None	None	None	cymt
Read_3	100	2	0.980000	100	1	0.990000	numt
Read_4	93	1	0.989247	None	None	None	cymt
Read_5	100	0	1.0000	None	None	None	cymt
Read_n	100	0	1.0000	100	0	1.0000	undetermined

[1]Read identifier; [2]Lenght of alignment to cymt (bp); [3]Number of bp mismatches between read and cymt reference; [4]Percentage identity between read and cymt reference; [5] Length of alignment to numt (bp); [6]Number of bp mismatches between read and numt reference; [7] Percentage identity between read and numt reference; [8]Putative DNA from which the read originated.

Post NUMT PARSER processing

NUMT PARSER produces an output table that lists per-read statistics for alignments to the cymt and numt reference (Table 1). This output table can be used to identify a subset of reads to retain for further analysis, allowing the user to subset the sequenced reads into a dataset most relevant for their study, and also allows for the compilation of datasets that include only reads originating from numts. For example, in this study we used the output table to compile a list of read IDs that were classified as "cymt" or "Undetermined" candidates. We then used PICARD TOOLKIT v. 2.10.1 to select and retain only reads included in the read ID list and exported the output in BAM format. Mitogenome alignment statistics (e.g., loss of coverage across genome pre- and post-filtering with NUMT PARSER) and graphs were created using the statistical software R (R Core Team, 2019).

Numt filtering using alternative bioinformatic approaches

In addition to the use of primers that prevent numt amplification (Curry *et al.*, 2019), previous studies (Curry *et al.*, 2021) have also used bioinformatic approaches that relied purely

on read alignment properties to account for numt DNA contamination in lion datasets. We compared NUMT PARSER to this published approach, and to a second approach that used Basic Local Alignment Search Tool (BLAST) to categorize reads as putatively cymt or numt in origin.

Alternative bioinformatic approach 1: BWA and SAMTOOLS

We filtered reads based on the alignment properties to a concatenated reference file that contained both cymt and numt reference sequence (hereafter referred to as SAMTOOLS filtering). This approach has previously been used to account for numt contamination when reconstructing lion mitogenomes (Curry *et al.*, 2021). Using the `mem` module of BWA, we aligned reads to a single FASTA file containing both cymt and numt sequences. Alignments were then processed with SAMTOOLS `view` to remove unmapped reads, alignments with a mapping quality less than 30, secondary, and supplementary alignments. This filtered dataset was composed of only high-quality alignments that primarily mapped either to the cymt or numt references. The processed and indexed alignments were further filtered using SAMTOOLS `view` to exclude reads mapped to the numt reference, effectively keeping only reads of true cymt origin. Coverage for this filtered dataset was then calculated using the SAMTOOLS `depth` function.

Alternative bioinformatic approach 2: Basic Local Alignment Search Tool (BLAST)

NUMT PARSER was also compared against an approach that filters the dataset using BLAST (hereafter referred to as BLAST filtering). A local nucleotide BLAST database was constructed from a single reference FASTA file composed of both cymt and numt reference sequences using `makeblastdb` v2.4.0+ (Camacho *et al.*, 2009). The reads of JCK 10711 and JCK 10712 were separately matched against this database using `blastn` v2.4.0+ and hits with an e-value of at least 0.001 and a maximum of 10 hits per query were saved in tabular format (`outfmt 6`). The resulting output table was then filtered to obtain the top BLAST hit per read using a custom Python script which compared e-values from all hits for a given read query (see Data Availability). If a top hit was the result of a comparison to the cymt reference sequence, the read query was returned as being putatively of cymt origin. From the resulting list of read IDs, the original alignment file for each sample was filtered using PICARD TOOLKIT to generate a cymt-only BAM file. Coverage for this filtered dataset was then calculated using the SAMTOOLS `depth` function.

Patterns of sequence disagreements

For each filtering method, mitochondrial consensus sequences were generated using the aligned reads for JCK 10711 and JCK 10712. Variant sites (those differing from the reference cymt genome) were identified and filtered, using a minimum per-base quality of 20 and minimum coverage of 3X, in BCFTOOLS (Danecek *et al.*, 2021) and BEDTOOLS (Quinlan and Hall, 2010). The coverage cut-off was implemented to increase accuracy of variant calls considering DNA damage and low DNA quality associated with ancient and historic samples (Parks and

Lambert, 2015). The resulting consensus sequences for each sample, alongside the consensus built from the numt-contaminated dataset (KF907306), were individually aligned to the cymt reference (KP202262) using the MAFFT aligner v7.310 (Katoh and Standley, 2013). We used these resulting alignments to visualize the sequence disagreements (i.e., differences due to a combination of biological polymorphism and technical error) between a given consensus sequence and the reference (Figure 3B). In addition, we aligned the unfiltered and NUMT PARSER-filtered consensus sequences of JCK 10711 and JCK 10712 against each other in order to observe the retention of variant sites with putative biological information between the two individuals (Figure 3C).

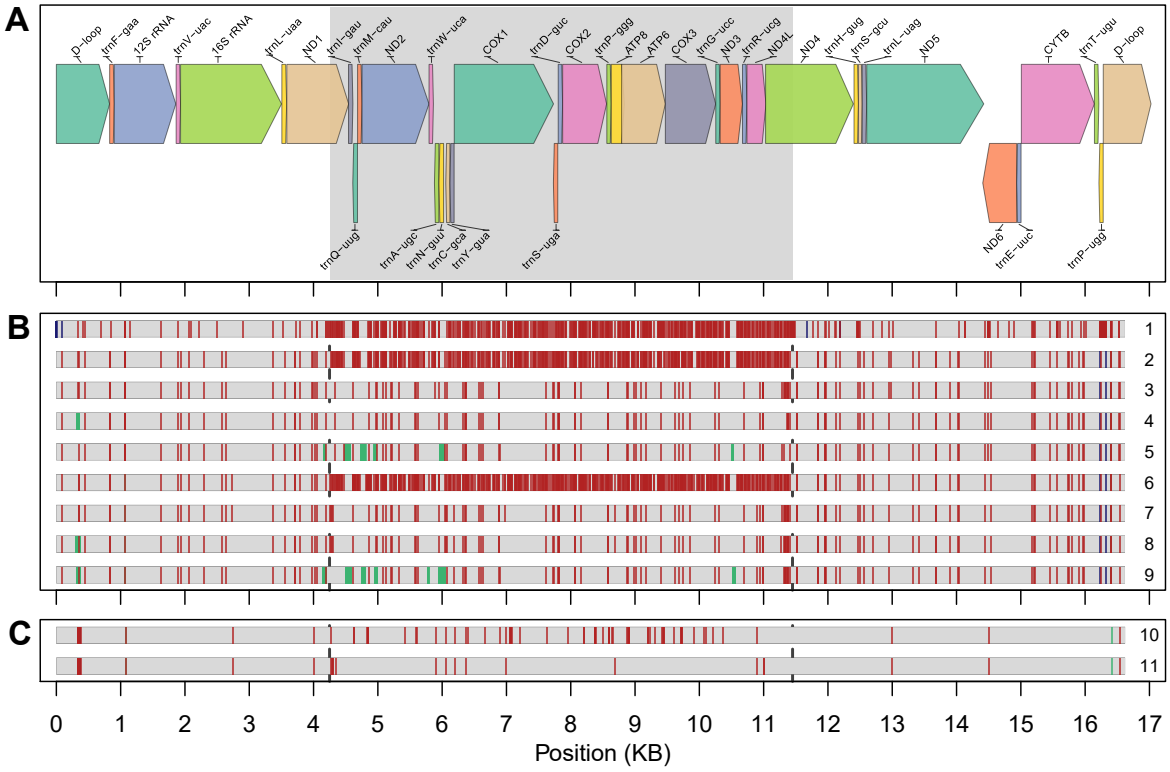

Figure 3. (A) The numt reference for use in NUMT PARSER was defined as the region of the Cape lion mitogenome that had above average coverage (see text). This region is highlighted by the grey box and corresponds to base-pair positions 4,250 to 11,450 of the reference cymt mitogenome (GenBank accession KP202262;(Li *et al.*, 2016)) including genes *ND1* (partial), *ND2*, *COX1*, *COX2*, *ATP8*, *ATP6*, *COX3*, *ND3*, *ND4L*, and *ND4* (partial). (B) Alignments of the unfiltered and filtered consensus sequences against the cymt reference (GenBank accession KP202262.1) for each Cape lion. Horizontal grey bar shows the span of the alignment, with the colored vertical bars displaying the location of sequence disagreements. Red bars represent single-base differences. Green bars show disagreements due to uncalled bases (Ns) in the consensus sequence. Blue indicates a deletion in the consensus with respect to the cymt reference. Dark grey dashed lines show the boundaries of the numt pseudogene. Numbers on the right are used to specify which consensus sequence was aligned to the cymt reference: (1) published sequence with known numt incorporation (GenBank accession KF907306; numt reference), (2) unfiltered JCK 10711, (3) NUMT PARSER-filtered JCK 10711, (4) BLAST-filtered JCK 10711, (5) SAMTOOLS-filtered JCK 10711, (6) unfiltered JCK 10712, (7) NUMT PARSER-filtered JCK 10712, (8) BLAST-filtered JCK 10712, (9) SAMTOOLS-filtered JCK 10712. In the unfiltered consensus (2 and 6), majority of

sequence disagreements are localized within the span of the numt sequence. These disagreements are largely removed by the filtering of the numt-contaminant reads from the consensus. (C) Alignments between the JCK 10711 and JCK 10712 consensus sequences for unfiltered datasets (10) and filtered datasets (11). Alignment of the unfiltered consensus (10) leads to an increased presence of disagreements between the two consensus sequences for these lions. The disagreements between these consensus sequences overlap with the numt pseudogene region which is consistent with our hypothesis that JCK 10712 had higher numt contamination than JCK 10711. Filtering numt contamination with NUMT PARSER leads to a reduction in the number of disagreements, but still allows for the retention of biologically-informative differences between the two samples (red bars in 11).

Results:

The Cape lion DNA damage patterns showed nucleotide base-pair changes characteristic of ancient DNA (e.g., deamination of cytosine to uracil; (Hofreiter, Jaenicke, *et al.*, 2001)). The read coverage distribution across the reconstructed mitogenomes differed between samples. Sample JCK 10711 showed relatively uniform coverage across the mitogenome, while JCK 10712 exhibited a pronounced increase in coverage in the region corresponding to the putative numt pseudogene (Figure 1 - blue). We see interspersed peaks of above-average coverage across the entire mitogenome of JCK 10711, although there is a slight increase in above-average coverage at the location of the numt pseudogene (Figure 1A - red). Above-average coverage for JCK 10712 is mostly limited to the numt pseudogene region (Figure 1B – red), suggesting that JCK 10712 has a higher degree of numt dataset contamination. Despite the differences in coverage between samples, we observe that disagreements increase for both samples in the region corresponding to the location of the 7,200 bp numt pseudogene in *Panthera* (Figure 3B, rows 2 and 6). The pattern of disagreements is similar to that of the numt contaminated sequence (Figure 3B, row 1), identified by Li et al. (2016) and previously accessioned in GenBank (KF907306), providing additional support for the presence of numt contamination in these ancient lion samples.

NUMT PARSER filters out reads of numtDNA origin without compromising biological signal

After filtering out reads of putative numt origin using NUMT PARSER, the loss of read coverage was mainly limited to the 7,200 bp region that correspond to the previously identified numt of *Panthera*. We observed a greater loss of coverage across the numt pseudogene region in JCK 10712 than in JCK 10711, consistent with their different levels of numt contamination (Table 2, Figure 1). NUMT PARSER therefore allows the user to equalize the average coverage across the genome by selectively removing reads of putative numt origin, without a substantial loss of read coverage from reads of cymt origin.

When comparing the consensus sequences against the cymt reference (Figure 3B), numt dataset contamination is evident as an overall increase in sequence disagreements (due to polymorphism and technical error), similar to that of a dataset with known contamination. When analyzing our Cape Lion sequences before and after NUMT PARSER processing, we found

that NUMT PARSER filtering led to a substantial reduction in the number of disagreements to the reference across both samples (Figure 3B, rows 3 and 7). This reduction demonstrates the removal of numt reads from the sequence pool, resulting in an uncontaminated mitogenome dataset.

Following NUMT PARSER filtering, the alignment between the filtered JCK 10711 and JCK 10712 consensus sequences still showed evidence of disagreements (Figure 3C), indicating that informative nucleotide base differences have been retained throughout the filtering process. Moreover, the filtered consensus sequences for JCK 10711 and JCK 10712 are different, highlighting that the two Cape lions had distinct mitogenome haplotypes. NUMT PARSER therefore allows for the removal of disagreements that are the consequence of numt pseudogene contamination but concomitantly retains informative nucleotide base differences, thus not sacrificing the true biological signal in the data.

Table 2. Alignment statistics for Cape lion mitochondrial genomes before and after filtering with NUMT PARSER.

	Before NUMT PARSER filtering			**After NUMT PARSER filtering**		
Sample	Reads (n)	Avg. cov (X)	Breadth of cov (%)	Reads (n)	Avg. cov (X)	Breadth of cov (%)
JKC 10711	13 729	52.99	100	11 879	45.1329	100
JKC 10712	10 891	44.6264	100	8 070	31.9798	100

NUMT PARSER outperforms available filtering alternatives

Similar to the results of NUMT PARSER, filtering reads of putative numt origin resulted in a greater loss of coverage for JCK 10712 (Figure 4B) than for JCK 10711 (Figure 4A) when using either the SAMTOOLS or BLAST filtering approaches. This is consistent with our hypothesis that JCK 10712 had higher numt contamination than JCK 10711. Post-filtering coverages differed for all three of the bioinformatic approaches tested. The mitochondrial-wide average coverage for JKC 10711 and JKC 10712, respectively, were 52.99x and 44.63x for the unfiltered datasets, 45.13x and 31.98x after NUMT PARSER filtering, 40.80x and 28.51x after SAMTOOLS filtering, and 42.32x and 30.33x after BLAST filtering (see Supplementary file 2). These trends are repeated in the resulting uncalled sites (i.e., with less than 3x coverage) pre- and postfiltering. Respectively for JKC 10711 and JKC 10712, both unfiltered and NUMT PARSER filtered datasets contained 1 and 2 uncalled sites, the SAMTOOLS filtering resulted in 245 and 340 uncalled sites, while BLAST filtering resulted in 37 and 38 uncalled sites (green bars in Figure 3B, Supplementary file 2).

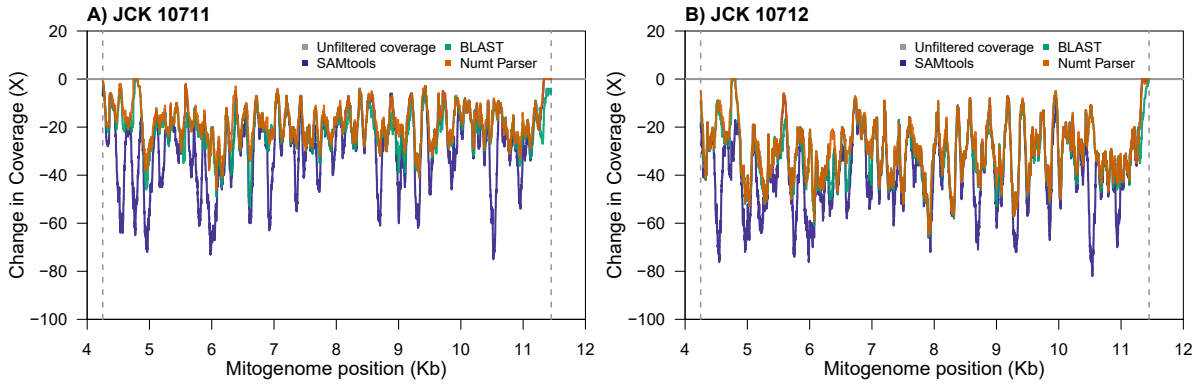

Figure 4. The three different bioinformatic approaches (NUMT PARSER, SAMTOOLS filtering, BLAST filtering) yielded consensus sequences with different coverages for JCK 10711 (A, left) and JCK 10712 (B, right). Coverage is only shown for positions within the defined 7,200bp numt region (x-axis). The y-axis shows the difference in coverage of the filtered datasets against the original, unfiltered dataset (horizontal grey line). Lines show that the coverage loss is greater for the SAMTOOLS filtered (purple) dataset than for BLAST filtered (green), and NUMT PARSER filtered (orange) datasets.

While the mitogenome-wide average loss of coverage was slightly higher for SAMTOOLS and BLAST filtering than for NUMT PARSER filtering, when we focus on the numt pseudogene region we see a much greater differential effect (Figure 4; Supplementary file 2). Comparing pre- and post-filtering coverages for the 7,200 bp numt region, we found that the SAMTOOLS filtered dataset (Figure 4, purple) had the largest decrease in coverage, likely because of the secondary alignments of reads to both the cymt and numt references (see Discussion). This marked decrease in coverage was evident even for JKC 10711, which has lower levels of numt contamination, for which SAMTOOLS filtering resulted in a coverage reduction of 27.52x (Figure 4A, purple).

The pre- and post-filter coverages for datasets filtered with BLAST (Figure 4, green) and NUMT PARSER (Figure 4, orange) yielded similar changes in depth of coverage in both JKC 10711 and JKC 10712. Within the numt region, the loss in coverage between pre- and post-filter datasets for NUMT PARSER and BLAST was substantially less than what was observed for the SAMTOOLS filtering approach (average loss of 32.16x, versus reductions of 25.96x and 23.58x for BLAST and NUMT PARSER respectively). The subtle differences in coverage between the BLAST and NUMT PARSER result appear to be the product of the sensitivity of BLAST when aligning reads of small length (see Discussion). NUMT PARSER and BLAST filtering both resulted in consensus sequences that had fewer disagreements with the global alignment compared to the unfiltered datasets while still retaining putatively informative sites (Figure 3C), however NUMT PARSER had the highest coverage post filtering for both JKC 10711 and JKC 10712.

Discussion:

Our software NUMT PARSER can help eliminate the effects of numt sequences in contaminated mitogenome datasets by identifying DNA sequencing reads that likely originate

from numt pseudogenes. These contaminating numt sequences can then be removed so that the reads retained in the dataset more accurately represent the true sequence of the mitogenome. Here we demonstrate the effectiveness of NUMT PARSER for removing numts when reconstructing the mitogenomes of two extinct Cape lions and show that NUMT PARSER outperforms other bioinformatic approaches that can account for numt contamination.

NUMT PARSER achieves these results by retaining more cymt reads to yield higher average genome coverages post-filtering, while minimizing numt dataset contamination. Retaining as many reads as possible without compromising the accuracy of the consensus sequence is especially pertinent for studies on ancient DNA where endogenous DNA-copy numbers are often low and where low-depth genomic datasets are standard.

Comparing the efficacy of NUMT PARSER filtering to that of SAMTOOLS, we observe a substantial reduction in coverage between pre- and post-SAMTOOLS filtering (Figure 4). In addition, we see that the number of uncalled sites in the consensus alignments increased by two orders of magnitude after filtering (File S2), due to a greater loss in coverage. This suggests that filtering numt contamination using SAMTOOLS may result in the discarding of cymt-derived reads and, subsequently, in the reduction of true biological signal via the loss of putatively informative variant sites. A reason for the higher coverage loss in the SAMTOOLS-based method is the processing of alignments in a single reference containing both cymt and numt sequences. The presence of two simultaneous references on which to map the reads might obfuscate the identification of their true origin. Reads of cymt origin can still align to the numt reference with high mapping qualities (MAPQ), and *vice versa*, making the identification of primary alignments more difficult. Similarly, when reads map with equal mapping quality to both references, true cymt reads can be discarded when filtering reads that have multiple alignments, reducing the overall coverage and the generation of consensus sequences. Using NUMT PARSER's strategy of independently mapping to cymt and numt sequences was key in identifying the origin of a read.

Filtering efficacy also explains the success of NUMT PARSER over the BLAST-based approach. While the differences in coverage are not as substantial as when comparing unfiltered or NUMT PARSER filtered alignments against those filtered with SAMTOOLS (Figure 4), BLAST filtering did result in an order of magnitude increase in uncalled sites (File S2). Interestingly, we did observe that the majority of lost reads when filtering with BLAST were not the product of alignment filtering (i.e., reads discarded due to poor values of sequence similarly, alignment length, or e-value). Instead, it appears that a portion of reads are unable to be aligned by BLAST prior to filtering. By default, the BLAST+ suite uses the `megablast` algorithm when performing nucleotide-to-nucleotide alignments (Camacho et al. 2009, Madden 2020), which requires an exact match of at least 28bp to align two sequences (controlled by the `-word_size` parameter). This alignment requirement is too stringent for short reads – for which BLAST has never been the primary alignment tool – particularly those originating from

aDNA. For example, when aligning with `blastn` using default settings, we observe that only 93.86% and 95.66% of reads are aligned for JKC 10711 and JKC 10712, respectively. By manually decreasing the `word_size` parameter to 11, matching that of `-task blastn`, the proportion of alignments can be further increased to 99.54% for JKC 10711 and 99.59% for JKC 10712, at the cost of obtaining more small, spurious alignments. While we observe that increasing the performance of the BLAST filtering is possible, it does require the use of non-default parameters. That, in addition to the requirement of an external script for filtering resulting alignments, further limits the BLAST filtering approach when compared to NUMT PARSER.

NUMT PARSER relies on well-defined numt reference sequences (such as those available for *Panthera*) for both alignment and sequence identity comparisons and assumes that the cymt reference is authentic and was not reconstructed using a numt contaminated dataset. This requirement also applies to any alignment-based method for numt filtering, including both SAMTOOLS and BLAST filtering methods presented in this paper. The accurate characterization of numts can be cumbersome for non-model organisms. Numt pseudogene reconstruction and characterization may be approached through molecular laboratory and bioinformatic methods (Bensasson *et al.*, 2001; Smart *et al.*, 2019). For example, DNA extractions can be enriched for nDNA instead of cymt before PCR amplification (e.g., by using nDNA rich tissues (Zischler, Geisert, *et al.*, 1995; Zhang and Hewitt, 1996)). Alternatively, genomic library construction and sequencing technologies that rely on single molecule, long read sequencing (e.g., Nanopore and PacBio) may be used to identify true cymt mitogenomes and concomitantly also numt pseudogenes. These technologies produce longer sequence read-lengths that may allow for the sequencing of complete mitogenomes using sequences with read-lengths that surpass the length of numt pseudogenes (Vossen and Buermans, 2017). For numt pseudogenes in particular, longer reads may be especially beneficial as they can be used to identify insertion locations of the pseudogenes into the nuclear genome, thereby allowing researchers to both differentiate numt pseudogenes from the true cymt mitogenome and also include numt pseudogenes within the nuclear reference genomes (Sohn and Nam, 2018). Bioinformatic approaches have also been developed to identify and characterize numt pseudogenes. For example, Smart et al., 2019 characterize numt pseudogenes through the comparison of differential topological properties of haplotype networks that are based on cymt and numt pseudogene sequences (Smart *et al.*, 2019).

While our implementation of NUMT PARSER focuses on identifying and removing numt pseudogene sequences in *Panthera*, it can be similarly applied to identify, retain, and compile a numt pseudogene dataset. NUMT PARSER may therefore allow for the use of numt pseudogenes as genetic markers for evolutionary inferences. For example, numt pseudogene sequences may be important genetic markers for the discovery and characterization of ancient mitochondrial haplotypes (Lammers *et al.*, 2017). Numt pseudogenes have been applied to reconstruct

ancestral mitochondrial sequences in the polar bear (*Ursus maritimus*), where historical mitogenome sequences are thought to be lost in extant populations, but putatively preserved as numt pseudogene paralogs (Lammers *et al.*, 2017). Numt pseudogene sequences may also be informative in taxa where they represent DNA transfer events that predate speciation. In such taxa they may provide a model of evolutionary change that could potentially be used to resolve phylogenetic inconsistencies, e.g., in taxa where incomplete lineage sorting or hybridization occurred (Wang *et al.*, 2015).

We tested NUMT PARSER using shotgun sequenced data from two ancient Cape lions, and we showed that NUMT PARSER is effective at removing numt contamination for this species. NUMT PARSER may likely also be effective at removing numt contamination for other taxa within the genus *Panthera* that have been reported to show similar patterns of pseudogene transposition (Li et al. 2016).

Future studies that seek to reconstruct cytoplasmic mitochondrial genomes should consider the possibility of amplifying DNA sequences that originate from numt pseudogene DNA, and their effects in the resulting data and analysis. Such studies may consider the use of our program, NUMT PARSER, which allows for the identification and removal of numt pseudogene sequences from high-throughput sequencing data for cytoplasmic mitochondrial genome reconstruction. NUMT PARSER equalizes read coverage across the mitogenome by removing reads that likely originate from numt pseudogene DNA, decreasing sequence disagreements that likely arise from numt pseudogene contamination. The assessment and removal of reads originating from numt pseudogene DNA through the use of NUMT PARSER may improve the reconstruction of true mitogenomes, and would allow for more accurate and robust biological inferences.

Acknowledgements:

For funding, we thank the USAID Wildlife TRAPS Project and the UIUC ACES Office of International Programs. AdeF was supported by the Program in Ecology, Evolution and Conservation Biology Research Award, UIUC, and by the Cooperative State Research, Education, and Extension Service, US Department of Agriculture, under project number ILLU 875–952. AGR-C was supported by NSF grant 1645087. We thank the UIUC High-Throughput Sequencing and Genotyping Unit. We thank Dr. Tolulope Perrin-Stowe for her help in collecting the ancient DNA samples at the FMNH. We thank Hazel Singer and the rest of the family of the late Dr. R Singer (University of Chicago) for access to and use of these historical specimens.

Data availability:

The Cape lion mitogenome sequences are available on GenBank under accession numbers XXXXXXXX-XXXXXXXX.

GitHub repository: https://github.com/adeflamingh/NuMt_parser

References:

Allende LM, Rubio I, Ruíz-del-Valle V, Guillén J, Martínez-Laso J, Lowy E, *et al.* (2001). The Old World sparrows (genus Passer) phylogeography and their relative abundance of nuclear mtDNA pseudogenes. *J Mol Evol* **53**: 144–154.

Antunes A, Ramos MJ (2005). Discovery of a large number of previously unrecognized mitochondrial pseudogenes in fish genomes. *Genomics* **86**: 708–717.

Armstrong EE, Taylor RW, Miller DE, Kaelin CB, Barsh GS, Hadly EA, *et al.* (2020). Long live the king: chromosome-level assembly of the lion (*Panthera leo*) using linked-read, Hi-C, and long-read data. *BMC Biol* **18**: 1–14.

Bensasson D, Zhang D-X, Hartl DL, Hewitt GM (2001). Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol Evol* **16**: 314–321.

Blanchard JL, Schmidt GW (1995). Pervasive migration of organellar DNA to the nucleus in plants. *J Mol Evol* **41**: 397–406.

Blanchard JL, Schmidt GW (1996). Mitochondrial DNA migration events in yeast and humans: integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns. *Mol Biol Evol* **13**: 537–548.

Bowers N, Stauffer JR, Kocher TD (1994). Intra-and interspecific mitochondrial DNA sequence variation within two species of rock-dwelling cichlids (Teleostei: Cichlidae) from Lake Malawi, Africa. *Mol Phylogenet Evol* **3**: 75–82.

Van Bree PJH (1998). On a mounted skeleton of apparently the extinct Cape lion, *Panthera leo melanochaita* (Ch. H. Smith, 1842). *Contrib to Zool* **68**: 67–71.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, *et al.* (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Chaitanya L, van Oven M, Weiler N, Harteveld J, Wirken L, Sijen T, *et al.* (2014). Developmental validation of mitochondrial DNA genotyping assays for adept matrilineal inference of biogeographic ancestry at a continental level. *Forensic Sci Int Genet* **11**: 39–51.

Chen S, Zhou Y, Chen Y, Gu J (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: i884–i890.

Collura R V, Stewart C-B (1995). Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids. *Nature* **378**: 485–489.

Cooper A, Poinar HN (2000). Ancient DNA: do it right or not at all. *Science (80-)* **289**: 1139.

Cronin MA, Amstrup SC, Garner GW, Vyse ER (1991). Interspecific and intraspecific mitochondrial DNA variation in North American bears (*Ursus*). *Can J Zool* **69**: 2985–2992.

Curry CJ, Davis BW, Bertola LD, White PA, Murphy WJ, Derr JN (2021). Spatiotemporal Genetic Diversity of Lions Reveals the Influence of Habitat Fragmentation across Africa. *Mol Biol Evol* **38**: 48–57.

Curry CJ, White PA, Derr JN (2019). Genetic analysis of African lions (Panthera leo) in Zambia support movement across anthropogenic and geographical barriers. *PLoS One* **14**: e0217179.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, *et al.* (2021). Twelve years of SAMtools and BCFtools. *Gigascience* **10**: giab008.

Gaziev AI, Shaikhaev GO (2010). Nuclear mitochondrial pseudogenes. *Mol Biol* **44**: 358–368.

Gilbert MTP, Bandelt H-J, Hofreiter M, Barnes I (2005). Assessing ancient DNA studies. *Trends Ecol Evol* **20**: 541–544.

Gill FB, Mostrom AM, Mack AL (1993). Speciation in North American chickadees: I. Patterns of mtDNA genetic divergence. *Evolution (N Y)* **47**: 195–212.

Goios A, Prieto L, Amorim A, Pereira L (2008). Specificity of mtDNA-directed PCR—influence of NUclear MTDNA insertion (NUMT) contamination in routine samples and techniques. *Int J Legal Med* **122**: 341–345.

Hofreiter M, Jaenicke V, Serre D, Haeseler A von, Pääbo S (2001). DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res* **29**: 4793–4799.

Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S (2001). ancient DNA. *Nat Rev Genet* **2**: 353–359.

Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L (2013). mapDamage2. 0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**: 1682–1684.

Katoh K, Standley DM (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

van der Kuyl AC, Kuiken CL, Dekker JT, Perizonius WRK, Goudsmit J (1995). Nuclear counterparts of the cytoplasmic mitochondrial 12S rRNA gene: a problem of ancient DNA and molecular phylogenies. *J Mol Evol* **40**: 652–657.

Lammers F, Janke A, Rücklé C, Zizka V, Nilsson MA (2017). Screening for the ancient polar bear mitochondrial genome reveals low integration of mitochondrial pseudogenes (numts) in bears. *Mitochondrial DNA Part B* **2**: 251–254.

Li G, Davis BW, Eizirik E, Murphy WJ (2016). Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). *Genome Res* **26**: 1–11.

Li H, Durbin R (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**: 589–595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lopez J V, Yuhki N, Masuda R, Modi W, O'Brien SJ (1994). Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *J Mol Evol* **39**: 174–190.

Madden TL *BLAST+ features*. National Center for Biotechnology Information (US).

Merheb M, Matar R, Hodeify R, Siddiqui SS, Vazhappilly CG, Marton J, *et al.* (2019). Mitochondrial DNA, a powerful tool to decipher ancient human civilization from domestication to music, and to uncover historical murder cases. *Cells* **8**: 433.
Parks M, Lambert D (2015). Impacts of low coverage depths and post-mortem DNA damage on variant calling: a simulation study. *BMC Genomics* **16**: 1–12.
Pereira SL, Baker AJ (2004). Low number of mitochondrial pseudogenes in the chicken (Gallus gallus) nuclear genome: implications for molecular inference of population history and phylogenetics. *BMC Evol Biol* **4**: 17.
Podnar M, Haring E, Pinsker W, Mayer W (2007). Unusual origin of a nuclear pseudogene in the Italian wall lizard: intergenomic and interspecific transfer of a large section of the mitochondrial genome in the genus Podarcis (Lacertidae). *J Mol Evol* **64**: 308–320.
Quinlan AR, Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.
R Core Team (2019). R Core Team (2019). R: A language and environment for statistical computing. *R Found Stat Comput Vienna, Austria URL http//www R-project org/, page R Found Stat Comput*.
Smart U, Budowle B, Ambers A, Moura-Neto RS, Silva R, Woerner AE (2019). A novel phylogenetic approach for de novo discovery of putative nuclear mitochondrial (pNumt) haplotypes. *Forensic Sci Int Genet* **43**: 102146.
Sohn J, Nam J-W (2018). The present and future of de novo whole-genome assembly. *Brief Bioinform* **19**: 23–40.
Song H, Buhay JE, Whiting MF, Crandall KA (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proc Natl Acad Sci* **105**: 13486–13491.
Sorenson MD, Quinn TW (1998). Numts: a challenge for avian systematics and population biology. *Auk* **115**: 214–221.
Den Tex R-J, Maldonado JE, Thorington R, Leonard JA (2010). Nuclear copies of mitochondrial genes: another problem for ancient DNA. *Genetica* **138**: 979–984.
Triant DA, Dewoody JA (2009). Integrating numt pseudogenes into mitochondrial phylogenies: comment on 'Mitochondrial phylogeny of Arvicolinae using comprehensive taxonomic sampling yields new insights'. *Biol J Linn Soc* **97**: 223–224.
Vaughan HE, Heslop-Harrison JS, Hewitt GM (1999). The localization of mitochondrial sequences to chromosomal DNA in orthopterans. *Genome* **42**: 874–880.
Vossen RHAM, Buermans HPJ (2017). Full-length mitochondrial-DNA sequencing on the PacBio RSII. In: *Genotyping*, Springer, pp 179–184.
Wang B, Zhou X, Shi F, Liu Z, Roos C, Garber PA, *et al.* (2015). Full-length Numt analysis provides evidence for hybridization between the Asian colobine genera Trachypithecus and Semnopithecus. *Am J Primatol* **77**: 901–910.
Woodward SR, Weyand NJ, Bunnell M (1994). DNA sequence from Cretaceous period bone fragments. *Science (80-)* **266**: 1229–1232.
Zhang D-X, Hewitt GM (1996). Nuclear integrations: challenges for mitochondrial DNA markers. *Trends Ecol Evol* **11**: 247–251.
Zischler H (2000). Nuclear integrations of mitochondrial DNA in primates: inference of associated mutational events. *Electrophor An Int J* **21**: 531–536.
Zischler H, Geisert H, von Haeseler A, Pääbo S (1995). A nuclear'fossil'of the mitochondrial D-loop and the origin of modern humans. *Nature* **378**: 489–492.
Zischler H, Hoss M, Handt O, Von Haeseler A, Van der Kuyl AC, Goudsmit J (1995). Detecting dinosaur DNA. *Science (80-)* **268**: 1192–1193.

Supplements:

Supplementary Table 1. Sample names and complete mitochondrial genome coverage statistics for unfiltered, SAMTOOLS filtered, BLAST filtered, and NUMT PARSER filtered datasets. For each method, we report the number of nucleotide base-pairs (num.sites), the average depth of coverage as X-fold (avg.cov), the average change in consensus sequence coverage for different filtering approaches compared to unfiltered datasets (avg. change), the median and standard deviation (median.cov and stdv.cov), the minimum and maximum X-fold read coverage at each position in the mitogenome (min.cov and max.cov) and the number of sites in the genome that have less than 3 X-fold read coverage (low.cov.sites). *for JKC 10711 and JKC 10712 combined.

sample.id	method	num. sites	avg. cov	avg. change	median. cov	stdv. cov	min. cov	max. cov	low.cov. sites
JKC 10711	unfiltered	16,620	52.99	0.00	53.0	12.43	1.0	94.0	1.0
JKC 10711	samtools	16,620	40.80	-12.19	43.0	14.91	0.0	78.0	245.0
JKC 10711	blast	16,620	42.32	-10.67	42.0	10.92	0.0	76.0	37.0
JKC 10711	numt_parser	16,620	45.13	-7.86	45.0	10.67	1.0	78.0	1.0
JKC 10712	unfiltered	16,620	44.63	0.00	40.0	17.16	2.0	101.0	2.0
JKC 10712	samtools	16,620	28.51	-16.12	29.0	10.82	0.0	80.0	340.0
JKC 10712	blast	16,620	30.33	-14.29	30.0	8.73	0.0	75.0	38.0
JKC 10712	numt_parser	16,620	31.98	-12.65	31.0	8.76	2.0	82.0	2.0
Average*	unfiltered	16,620	48.81	0.00	46.5	14.80	1.5	97.5	1.5
Average*	samtools	16,620	34.65	-14.16	36.0	12.87	0.0	79.0	292.5
Average*	blast	16,620	36.33	-12.48	36.0	9.82	0.0	75.5	37.5
Average*	numt_parser	16,620	38.56	-10.25	38.0	9.72	1.5	80.0	1.5

Supplementary Table 2. Sample names and numt pseudogene coverage statistics for unfiltered, SAMTOOLS filtered, BLAST filtered, and NUMT PARSER filtered datasets. For each method, we report the number of nucleotide base-pairs (num.sites), the average depth of coverage as X-fold (avg.cov), the average change in consensus sequence coverage for different filtering approaches compared to unfiltered datasets (avg. change), the median and standard deviation (median.cov and stdv.cov), and the minimum and maximum X-fold read coverage at each position in the mitogenome (min.cov and max.cov). *for JKC 10711 and JKC 10712 combined.

sample.id	method	num.sites	avg.cov	avg.change	median.cov	stdv.cov	min.cov	max.cov
JKC 10711	unfiltered	7,201	59.375	0.000	60.0	12.0	23.0	94.0
JKC 10711	samtools	7,201	31.851	-27.525	33.0	14.6	0.0	69.0
JKC 10711	blast	7,201	38.897	-20.479	39.0	9.9	10.0	67.0
JKC 10711	numt_parser	7,201	41.374	-18.001	41.0	9.7	12.0	69.0
JKC 10712	unfiltered	7,201	59.760	0.000	60.0	13.5	21.0	101.0
JKC 10712	samtools	7,201	22.958	-36.803	23.0	10.6	0.0	62.0
JKC 10712	blast	7,201	28.860	-30.901	28.0	8.6	9.0	67.0
JKC 10712	numt_parser	7,201	30.609	-29.152	29.0	8.9	10.0	67.0
Average*	unfiltered	7,201	59.568	0.000	60.0	12.7	22.0	97.5
Average*	samtools	7,201	27.404	-32.164	28.0	12.6	0.0	65.5
Average*	blast	7,201	33.878	-25.690	33.5	9.2	9.5	67.0
Average*	numt_parser	7,201	35.991	-23.576	35.0	9.3	11.0	68.0