NUMT PARSER manuscript supplementary data

Supplementary information for:

> de Flamingh, A., Rivera-Colon, A.G., Gnoske, T.P., et. al. (*In Press.*) ***Numt Parser*: automated identification and removal of nuclear mitochondrial pseudogenes (numts) for accurate mitochondrial genome reconstruction in *Panthera*.** *Journal of Heredity*. (Link to *bioRxiv* preprint).

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 found 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.

General Methods

Sequencing reads and alignments generated from ancient DNA of two Cape lion (*Panthera leo melanochaitus*) samples. Raw reads were aligned to the *P. leo* mitochondrial reference (NCBI Accession KP202262.1) to obtain mitochondrial-specific reads. These mitochondrial reads were then processed using different methods (`BLAST`, `SAMtools`, `Numt Parser`) to identify and filter Numt-contaminant reads. See de Flamingh, et al. (2022) for additional information on the specific bioinformatic pipeline used and a description of the `Numt Parser` software.

Usage Notes

Files in BAM format (`.bam`) are stored in binary and require the use of `SAMtools` for conversion. SAM (`.sam`) and FASTA (`.fa`) files are in text format and can be accessed using any text editor software (in either the command line or an graphical application).

Sample Names

The two sampled Cape lion specimens (JCK10711 and JCK10712) are labelled according to their accession ID at the Field Museum of Natural History.

Mitochondrial Dataset

For each sample, the mitochondrial read dataset (in BAM format) is comprised of all ancient DNA sequencing reads that successfully aligned to the lion cymt reference (KP202262.1). The trimmed paired-end reads are aligned to the reference using `BWA mem` and filtered for alignment quality using `SAMtools view`. Unaligned reads are discarded in order to remove exogenous sequences and reads of nuclear origin. See link for the specification of the SAM/BAM format.

File Name	Description
`JCK10711_unfiltered.bam`	Unfiltered mitochondrial dataset, comprised of aligned reads in BAM format, for sample JCK10711
`JCK10712_unfiltered.bam`	Unfiltered mitochondrial dataset, comprised of aligned reads in BAM format, for sample JCK10712

Consensus FASTA

FASTA file containing the mitochondrial consensus sequence generated from the unfiltered mitochondrial dataset in each Cape lion sample. Since the unfiltered dataset contains both reads of cymt origin as well as numt contaminants, this consensus will contain high disagreements from the published lion mitochondrial reference.

File Name	Description
`JCK10711_unfiltered_consensus.fa`	Mitochondrial consensus sequence for sample JCK10711
`JCK10712_unfiltered_consensus.fa`	Mitochondrial consensus sequence for sample JCK10712

SAM Input

SAM files containing the `Numt Parser` input alignments. Each sample shows two SAM files, each showing the alignment of the unfiltered mitochondrial reads to 1) the lion cymt reference (KP202262.1), and 2) lion numt reference (from Li et al. (2016), see de Flamingh et al. (2022) for the specification of the numt reference). See link for the specification of the SAM/BAM format.

File Name	Description
`JCK10711_unfiltered_cymt.sam`	Alignment of the unfiltered mitochondrial reads to the cymt reference to sample JCK10711
`JCK10711_unfiltered_numt.sam`	Alignment of the unfiltered mitochondrial reads to the numt reference to sample JCK10711
`JCK10712_unfiltered_cymt.sam`	Alignment of the unfiltered mitochondrial reads to the cymt reference to sample JCK10712
`JCK10712_unfiltered_numt.sam`	Alignment of the unfiltered mitochondrial reads to the numt reference to sample JCK10712

Filtered Alignments

Datasets comprised of the cymt-specific reads for each sample, each generated using a different method to identify and remove reads of numt origin:

1. `BLAST`: numt reads are identified based on their sequence identity (defined by the e-value) to the numt reference using `BLAST`
2. `BWA`: numt reads are identified based on the primary/secondary alignments as defined by `BWA mem` + `SAMtools`
3. `Numt Parser`: numt reads identified based their percent mismatch to the numt reference using `Numt Parser`

Refer to de Flamingh et al. (2022) for the specific filtering paramers. Each filtered dataset is in BAM alignmed format, and shows the alignment of cymt-specific reads to the lion cymt reference (KP202262.1).

File Name	Description
`JCK10711_filtered_BLAST.bam`	filtered cymt reads (using the `BLAST` method) for sample JCK10711
`JCK10711_filtered_BWA.bam`	filtered cymt reads (using the `BWA`+`SAMtools` method) for sample JCK10711
`JCK10711_filtered_NumtParser.bam`	filtered cymt reads (using the `Numt Parser` method) for sample JCK10711
`JCK10712_filtered_BLAST.bam`	filtered cymt reads (using the `BLAST` method) for sample JCK10712
`JCK10712_filtered_BWA.bam`	filtered cymt reads (using the `BWA`+`SAMtools` method) for sample JCK10712
`JCK10712_filtered_NumtParser.bam`	filtered cymt reads (using the `Numt Parser` method) for sample JCK10712

Authors

Alida de Flamingh
Program in Ecology, Evolution, and Conservation Biology
Carl R. Woese Institute for Genomic Biology
University of Illinois at Urbana-Champaign
deflami2@illinois.edu

Angel G. Rivera-Colon
Department of Evolution, Ecology, and Behavior
University of Illinois at Urbana-Champaign
angelgr2@illinois.edu