Check for updates

SOFTWARE TOOL ARTICLE

REVISED MTaxi: A comparative tool for taxon identification of ultra low coverage ancient genomes

[version 3; peer review: 2 approved]

Gözde Atağ[1], Kıvılcım Başak Vural[1], Damla Kaptan[1], Mustafa Özkan[1], Dilek Koptekin[1,2], Ekin Sağlıcan[2], Sevcan Doğramacı[3], Mevlüt Köz[4], Ardan Yılmaz[5], Arda Söylev[3], İnci Togan[1], Mehmet Somel[1], Füsun Özer[6]

[1]Biological Sciences, Middle East Technical University, Ankara, Turkey
[2]Health Informatics, Middle East Technical University, Ankara, Turkey
[3]Computer Engineering, Konya Food and Agriculture University, Konya, Turkey
[4]Molecular Biology and Genetics, Konya Food and Agriculture University, Konya, Turkey
[5]Computer Engineering, Middle East Technical University, Ankara, Turkey
[6]Anthropology, Hacettepe University, Ankara, Turkey

v3 **First published:** 22 Aug 2022, **2**:100
https://doi.org/10.12688/openreseurope.14936.1

Second version: 07 Jul 2023, **2**:100
https://doi.org/10.12688/openreseurope.14936.2

Latest published: 29 Sep 2023, **2**:100
https://doi.org/10.12688/openreseurope.14936.3

Abstract

A major challenge in zooarchaeology is to morphologically distinguish closely related species' remains, especially using small bone fragments. Shotgun sequencing aDNA from archeological remains and comparative alignment to the candidate species' reference genomes will only apply when reference nuclear genomes of comparable quality are available, and may still fail when coverages are low. Here, we propose an alternative method, MTaxi, that uses highly accessible mitochondrial DNA (mtDNA) to distinguish between pairs of closely related species from ancient DNA sequences. MTaxi utilises mtDNA transversion-type substitutions between pairs of candidate species, assigns reads to either species, and performs a binomial test to determine the sample taxon. We tested MTaxi on sheep/goat and horse/donkey data, between which zooarchaeological classification can be challenging in ways that epitomise our case. The method performed efficiently on simulated ancient genomes down to 0.3x mitochondrial coverage for both sheep/goat and horse/donkey, with no false positives. Trials on n=18 ancient sheep/goat samples and n=10 horse/donkey samples of known species identity also yielded 100% accuracy. Overall, MTaxi provides a straightforward approach to classify closely related species that are difficult to distinguish through zooarchaeological methods using low coverage aDNA data, especially when similar quality reference genomes are unavailable. MTaxi is freely available at https://github.com/goztag/MTaxi.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 3 (revision) 29 Sep 2023		✓ view
version 2 (revision) 07 Jul 2023	✓ view	? view
version 1 22 Aug 2022	? view	? view

1. **Kevin G. Daly**, Trinity College Dublin, Dublin, Ireland

2. **Burak Yelmen**, Paris-Saclay University, Gif-sur-Yvette, France
University of Tartu, Tartu, Estonia

Any reports and responses or comments on the article can be found at the end of the article.

Keywords
ancient DNA, zooarcheology, mtDNA, bioinformatics, genomics

This article is included in the European Research Council (ERC) gateway.

This article is included in the Genetics and Genomics gateway.

This article is included in the Evolution and Ecology gateway.

This article is included in the Horizon 2020 gateway.

This article is included in the Evolutionary Biology collection.

Corresponding authors: Gözde Atağ (atag.gozde@gmail.com), Füsun Özer (fusun.ozer@gmail.com)

Author roles: Atağ G: Conceptualization, Data Curation, Formal Analysis, Methodology, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Vural KB**: Conceptualization, Data Curation, Formal Analysis, Resources, Writing – Review & Editing; **Kaptan D**: Conceptualization, Writing – Review & Editing; **Özkan M**: Conceptualization, Resources, Writing – Review & Editing; **Koptekin D**: Conceptualization, Methodology, Writing – Review & Editing; **Sağlıcan E**: Conceptualization, Writing – Review & Editing; **Doğramacı S**: Software, Validation; **Köz M**: Software, Validation; **Yılmaz A**: Methodology, Software, Validation; **Söylev A** : Conceptualization, Supervision, Writing – Review & Editing; **Togan İ**: Conceptualization, Supervision, Writing – Review & Editing; **Somel M**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing; **Özer F**: Conceptualization, Methodology, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 772390).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2023 Atağ G *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Atağ G, Vural KB, Kaptan D *et al.* **MTaxi: A comparative tool for taxon identification of ultra low coverage ancient genomes [version 3; peer review: 2 approved]** Open Research Europe 2023, **2**:100
https://doi.org/10.12688/openreseurope.14936.3

First published: 22 Aug 2022, **2**:100 https://doi.org/10.12688/openreseurope.14936.1

REVISED **Amendments from Version 1**

We thank Dr. Yelmen for his suggestions. The main changes we made in the article are explained below.

Dr. Yelmen had pointed out ambiguities in the definition and calculation of the performance metrics. Following this, we rewrote the formulas more clearly and described which cases true positives, false positives and false negatives correspond to. We also clarified that we have three possible classes in total; correctly identified cases, misidentified cases and unidentified cases. Since these classes do not include true negatives, our calculations lack this metric. For the accuracy metric, we now use the term "accuracy" rather than F1 score in the calculation part same as the main text, since in the absence of true negative cases, F1 score corresponds to "accuracy". We further stated that our metrics include the results for both target species. These clarifications are made under the section titled "Performance metric calculation".

Any further responses from the reviewers can be found at the end of the article

Introduction

Archaeological faunal remains have been widely used to address various questions in biology and social sciences. The scope of these range from the demographic history of wild populations, which can inform about ecological dynamics and conservation biology, to animal management and breeding practices, providing insights into the subsistence strategies and lifeways of prehistoric human societies that exploited animals[1–6]. A key step here is the accurate taxonomic identification of animal remains. However, distinguishing morphologically similar species in zooarchaeological material is a prevailing challenge, constrained by the high level of similarity between skeletal elements, the fragmented state of excavated specimens (possibly with missing fragments), and the absence of morphological markers in subadults[7,8]. The need for an effective approach to identify species' remains accurately has thus led to the development of several alternative methods, including isotope analyses, protein fingerprinting, and ancient DNA (aDNA) analyses[9–14].

The majority of non-human aDNA data today is produced using shotgun DNA sequencing on Illumina platforms[15]. Beyond species identification, such data from well-preserved zooarchaeological samples can yield a wealth of information to study demographic and evolutionary history. However, relatively old (e.g. >1000 years old) zooarchaeological samples from regions with humid, temperate or warmer environments are mostly poorly preserved[16]. Cooking and other forms of heat treatment before human consumption may additionally degrade organic material[17]. In such poorly preserved samples, the proportion of endogenous DNA among the total DNA read pool will be low, down to 1% or even lower[18]. Accordingly, most experiments can produce only low amounts of DNA sequence data, if any, from zooarchaeological samples from temperate regions within reasonable budgets; such genomic data frequently remain at genome-wide depths of coverage <0.1x per sample[19,20].

Theoretically, even 0.1x coverage genome data could allow accurate taxonomic identification by comparative alignment, i.e. mapping reads to the reference genomes of alternative candidate species, such as sheep versus goat, or horse versus donkey, and comparing coverages or mismatch rates. However, this only applies to situations where both species have assembled nuclear reference genomes (e.g. no such reference is available for the donkey). Even in cases where nuclear genomes from both species are available (e.g. sheep and goat), the limited amount of shotgun sequencing data available from poorly preserved samples, quality differences between the reference genomes, the highly fragmented nature of aDNA hence short read lengths, and postmortem damage can introduce high levels of uncertainty in the alignment process (See Discussions in 21,22). The problem is further exacerbated when the sequence similarity between candidate reference genomes is high or when there exists strong differences in the genome assemblage qualities.

These call for new approaches for species identification with aDNA data. For instance, the Zonkey[23] pipeline was developed for distinguishing horse, donkey and their hybrids by using nuclear aDNA variants with a clustering approach, but is only applicable for equid taxa from which there exists large datasets of genetic variation. Here we present a broadly applicable method, MTaxi, designed for distinguishing pairs of any closely related species using low amounts of shotgun aDNA sequencing data, whenever mitochondrial DNA (mtDNA) reference sequences are available. Our method focuses on mtDNA owing to its short size, haploid nature, having a lower rate of decay than nuclear DNA[16], and having multi-copies per cell, which increases its availability relative to autosomes[24], facilitating analyses. For example, across n=310 shotgun sequenced ancient DNA libraries from human, sheep, goat, horse and donkey generated by our group, each of which contained ≥0.01 endogenous DNA, the average ratio of mitochondrial DNA to nuclear DNA coverage was 87:1 (data not shown). The greater number of informative sites due to the high mitochondrial mutation rate is an additional advantage for taxon identification of closely related species[25]. Finally, the availability of mitochondrial reference sequences for a larger number of taxa (compared to a limited number of high quality reference genomes) allows our approach to be applied to a wider number of species, including extinct lineages. For instance, as of December 24 2021, the genome resources database from NCBI includes only 175 nuclear genomes for mammals[26], compared to 1453 mitochondrial genomes[27], an 8-fold difference. The MTaxi[28] approach uses roughly the same information as competitive mapping in species assignment, but has the advantage that it calculates statistical significance for the result.

To exemplify the use of MTaxi we chose the case of sheep (Ovis) versus goat (Capra), two closely related species belonging to the same subfamily Caprinae. The aforementioned constraints on morphological identification causes a large proportion of sheep and goat remains to be only identified at the subfamily level as Caprinae[8], and ambiguity which can significantly

constrain zooarchaeological analyses, especially in the study of animal husbandry. Here we first estimate MTaxi's accuracy using 1600 ancient mitogenome simulations with eight different coverages from both sheep and goat. We then test its performance with n=9 ancient sheep samples[19,20] and n=3 goat samples[29]. We further test MTaxi on the horse and donkey, a pair which also suffer from difficulties in morphological differentiation, using n=3 ancient horse, n=2 modern domestic horse, and n=5 modern domestic donkey samples adopted from the literature[30–34].

Methods

Overview of the method

MTaxi makes use of the mismatch positions, i.e. putative substitutions, between two alternative candidate taxa, such as sheep and goat. These "target sites" are obtained from pairwise alignment of mitochondrial reference genomes. Each read harbouring the target sites is classified according to the genotype compositions, and we identify the taxon using a binomial test for the read proportion of the sample (Figure 1).

Target sites

The method involves compiling a list of mtDNA target sites, representing likely substitutions between the species. To generate this list for sheep and goat, we first generated a pairwise alignment between sheep (Oar_v3.1) and goat (ARS1) mtDNA reference genomes via the R package *Biostrings* v.2.65.0[35] using default parameters, which yielded n=1699 single nucleotide substitutions. We then restricted these to transversions to avoid (a) confounding effects due to postmortem damage-induced transitions in ancient DNA, and (b) homoplasies that could arise by high-frequency transitions[36]. This yielded a set of n=197 transversion substitutions, which we refer to as the target site list 1.

We also created a subset of this, that we call target sites type 2, by removing polymorphisms in either species, which we reasoned might increase power by avoiding ambiguities. For this, we obtained a list of polymorphic sites using the software *snp-sites* v.2.4.1[37] from a data set assembled by Shi and colleagues[38], which contains pairwise alignments (each sequence aligned to the reference genome) for mtDNA sequences belonging to n=47 domestic sheep and n=35 domestic goats. In this dataset, we identified n=57 and n=40 polymorphic single nucleotide positions overlapping with the n=197 target sites in domestic sheep and goat, respectively. After eliminating these polymorphisms we were left with n=120 positions, which we refer to as the target site list 2.

We applied the same procedures to horse and donkey by using mtDNA references NC_001640.1 and NC_001788.1. This resulted in n=1264 substitution sites, and restricting these to transversions yielded n=117 positions (target sites type 1). The positions are concentrated around the D-loop, but are also represented across the mitochondrial genome following similar patterns between the two species (Figure 2).

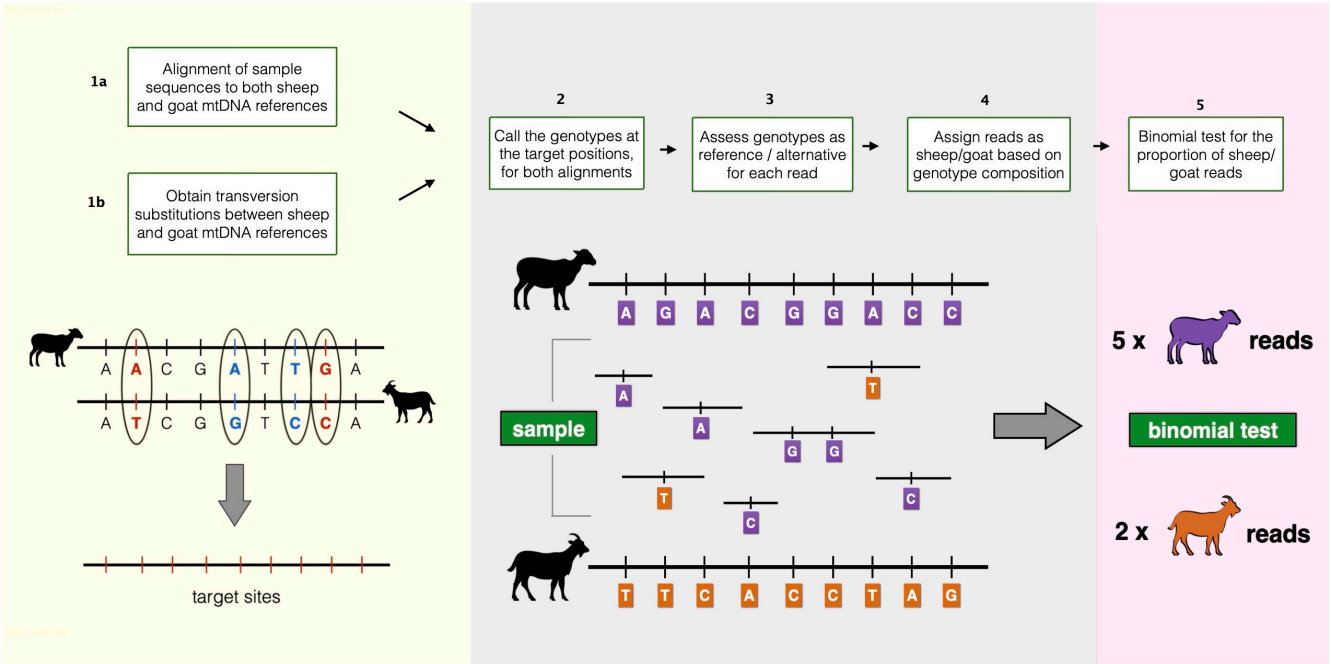

Figure 1. Overview of the MTaxi pipeline. Flowchart and representations of the steps to determine the sample taxon. Here sheep and goat stand for the candidate species pair, but MTaxi can be applied to any pair of species where mitochondrial DNA reference sequences are available. Target sites represent mismatches (candidate substitutions) between the reference genomes, restricted to transversions. Blue coloured alleles represent transitions, red coloured alleles represent transversions. Reads are assigned to either taxon based on target sites. Purple coloured alleles represent sheep alleles, orange coloured alleles represent goat alleles. Reads may be assigned to the wrong taxon due to homoplastic mutations, technical error, or incomplete lineage sorting.

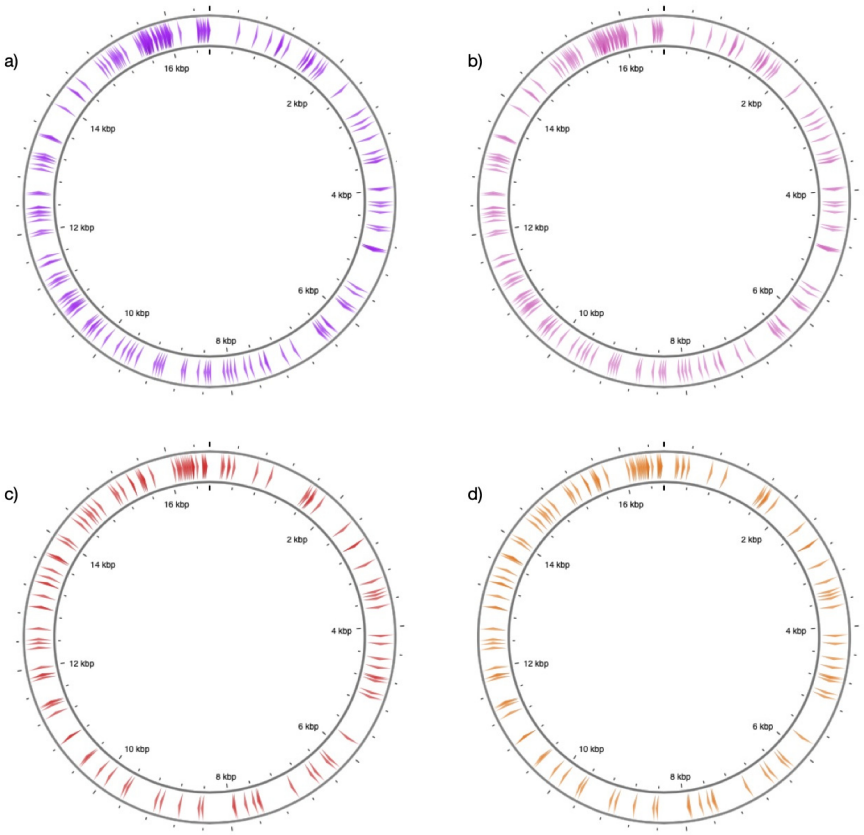

Figure 2. Distribution of target sites along reference mitochondrial genomes. The figure shows the position of target sites along (**a**) sheep and (**b**) goat (**c**) horse (**d**) donkey reference mitochondrial genomes. The sites represent transversion-type substitutions (n=197 for sheep and goat, and n=117 for horse and donkey). The figure was generated through CGview[39].

Alignment and genotyping

Ancient DNA data processing. *AdapterRemoval* v.2.3.1[40] was used for trimming residual adapter sequences and merging paired-end sequencing reads with an overlap of at least 11 bp between the pairs. Whole genome sheep and goat data (FASTQ files) were aligned to both sheep (Oar_v3.1) and goat (ARS1) reference genomes. BWA mem is designed as a highly sensitive aligner for reads >70 bp[41]. It has been also shown that BWA aln outperforms BWA mem in ancient DNA sequence alignment in terms of precision and proportion of the reads mapped[42]. We therefore used BWA aln for the alignment of ancient DNA sequences and mem for modern DNA sequences. *BWA aln* v.0.7.15[43] was run with parameters: "-n 0.01 -o 2" and disabling the seed with "-l 16500". Mitochondrial goat data (Gilat10, Shiqmim9, Kov27, Uiv17) were aligned to both sheep (NC_001941.1) and goat (NC_005044.2) mitochondrial references with the same parameters as well. To prevent the influence of the PCR duplicates, reads with identical start and end positions were removed using "FilterUniqSAMCons_cc.py"[44]. After the removal of PCR duplicates, reads with mapping quality scores (MAPQ) lower than 30 were filtered out using *samtools* (v.1.9)[45]. The reads mapping to the reference genome with >10% mismatches and having a length <35 bp were filtered out. In order to confirm the authenticity of the ancient DNA sequences from non-UDG-treated libraries (TEP03, TEP62, TEP83, ULU26, ULU31), we evaluated damage patterns which are characteristic of aDNA were estimated using *PMDtools*[46] "--deamination" parameter; all libraries had >20% C->T and G->A damage at 5' and 3' ends, respectively. The reads aligned to mtDNA were extracted from whole genome alignments using *samtools* (v.1.9)[45]. For each alignment, we called the genotypes of the target sites using *pysamv* v.0.16.0.1, which also runs *samtools* (v.1.9)[45]; genotyping was performed with parameters "-B" and "-A". As default, MTaxi uses both the reads that aligned only to one of the species' references and the ones that aligned to both species' references in the analysis (which we refer to as "all reads" below). Additionally, we included an option ("shared reads"), by which the reads are restricted to those that are aligned to both species' references (the reads with the same IDs between the two BAM files); this is a conservative approach that could eliminate the possible effects of quality differences between the two reference genomes. Using *pybedtools* v.0.8.1[47,48], we obtained the reads overlapping with the target sites. We note that aligning reads to both nuclear and mitochondrial genomes is superior to alignment only to the mitochondrial genome, because the latter can cause misalignment of nuclear mitochondrial DNA sequences (NUMTs) to the mitochondrial genome.

The alignment and genotyping procedures for whole genome ancient horse data were applied in the same way as described for sheep and goat data. However, for the alignment, due to lack of a nuclear reference genome for the donkey, equid reads were mapped only to mtDNA references of the two species.

For the comparison of whole genome mapping frequencies, the total number of bases aligned to both reference genomes were calculated using *samtools stats*[45], and we calculated the number of mismatches for each alignment.

Modern DNA data processing. After removing residual adapter sequences using *AdapterRemoval* v.2.3.1[40], we mapped the whole genome data of modern horse and donkey at pair-ended mode to both horse (NC_001640.1) and donkey (NC_001788) mitochondrial reference genomes using *BWA mem* (version 0.7.15)[43] module with the parameter '-M', and sorted the output using *samtools* (v.1.9) *sort*[45]. Duplicates were removed using *Picard MarkDuplicates*. Reads with mapping quality scores lower than 20 were filtered out using *samtools* (v.1.9)[45]. Libraries from the same individual were merged using *samtools* (v.1.9) *merge*[45], and then the same filtering and genotyping procedures described in Ancient DNA data processing section process were applied on all modern equid data.

Taxon assignment

Each read can carry either reference or alternative alleles at its target sites. MTaxi uses this data to assign reads to either taxon, species 1 (SP1) or species 2 (SP2), based on whether they carry the alternative allele or not. If an SP1 read was aligned to the SP1 genome, we expect no alternative alleles at target sites, and if aligned to the SP2 genome, we expect all alternative alleles. For the reads which overlap more than one target site, MTaxi retains reads carrying only alternative alleles or carrying only reference alleles, thus excluding reads with inconsistent alleles (i.e. alternative and reference alleles mixed) at target sites. Such inconsistent variants could represent PCR or sequencing errors, convergent mutations, or incomplete lineage sorting. Having thereby assigned reads as SP1 or SP2, MTaxi uses the proportion of these two classes of reads to determine the sample taxon using a two-tailed binomial test with the null hypothesis of p=0.5.

Throughout the study, for the evaluation of the power of MTaxi using samples with known species identity, we use one-tailed binomial tests since we have an a priori hypothesis about each sample's taxon. Meanwhile, in the method itself we use a two-tailed binomial test assuming both target species are equally likely.

Ancient mitogenome simulations

We simulated 1600 ancient sheep and goat mitochondrial genomes (100 sheep and 100 goats for each coverage) at eight different coverages (0.1x, 0.3x, 0.5x, 1x, 2x, 3x, 4x, 5x) using *gargammel*[49], and tested the accuracy of the method. The sequencing error was set to ~1% using the parameters "qs -10 qs2 -10". The simulations for horse and donkey (100 horse and 100 donkey for each coverage) were run with the same parameters above, again at eight different coverages (0.1x, 0.3x, 0.5x,1x,2x,3x,4x,5x). The same alignment and genotyping procedures described in Ancient DNA data processing section were applied to the simulated data, except that they were mapped only to the mtDNA references of the species.

Ancient and modern samples

We used published FASTQ files to study the performance of MTaxi on real data. For sheep, we used FASTQ files of n=5 ancient sheep individuals (TEP03, TEP62, TEP83, ULU26, ULU31) from Yurtman *et al.*[19] downloaded from the European Nucleotide Archive (ENA) database (Table 1), and n=4 (OBI013, OBI014, OBI017, OBI018) ancient sheep individuals from Taylor *et al.*[20] downloaded from ENA (Table 1). All data had been produced with Illumina sequencing using either whole genome shotgun sequencing or using SNP capture followed by sequencing. For goat, we used n=5 (Acem1, AP45, Azer3, Direkli1, Direkli6) ancient whole genome FASTQ files, and n=4 (Gilat10, Shiqmim9, Kov27, Uiv17) ancient mitochondrial capture FASTQ files, produced by shotgun Illumina sequencing and mtDNA capture-sequencing, and published by Daly *et al.*[29], downloaded from ENA (Table 1). For equids, we used n=3 ancient[30] and n=2 modern domestic horses[31,32] and n=5 modern domestic donkey[33,34] FASTQ files, downloaded from ENA (Table 1). We randomly downsampled the equid files to mtDNA coverages ranging from ~0.3x to ~4x using *samtools view* with the option "-s"[45].

Performance metric calculation

We measured the performance of MTaxi using three metrics; precision, recall and accuracy. Precision was calculated as (TP)/(TP + FP), recall was calculated as (TP)/(FN + TP), where TP (true positives) correspond to correctly identified cases, FP (false positives) to misidentified cases, and FN (false negatives) correspond to unidentified cases. All three metrics include the identifications made for both target species pairs (SP1 + SP2). Accuracy was estimated using the formula ((TP) / (TP + FP + FN)). Here, since we have three categories in total, comprising correctly identified, misidentified and unidentified cases, we do not include the category of TN (true negatives).

Results

Application to simulated ancient mitogenomes

We first studied the performance of MTaxi using ancient-like mtDNA read data simulations. We produced n=1600 mtDNA read datasets at varying coverage, n=800 for sheep and n=800 for goat (Materials and Methods). Using n=197 transversion substitutions (target sites type 1), MTaxi assigned BAM files to their respective taxa with 100% precision (i.e. no false positives) across all mtDNA coverages from 0.1x–5x using the default ("all reads") approach (Figure 3a). All simulated data had a recall (i.e. true positive rate) of 100% and no false positives, at mtDNA coverages ≥0.5x (Figure 3a).

For 0.3x and 0.1x, the recall was 93.5% and 51% respectively, while the precision was again 100% at both coverages. We also tested the performance of two more conservative approaches

Table 1. Genome data used in the study. The table lists ancient and modern-day genomes downloaded from European Nucleotide Archive (ENA), with the study accession IDs and sample aliases.

Sample Aliases	Species	Age	Study Accession	Publication
TEP03, TEP62, TEP83, ULU26, ULU31	Sheep	Ancient	PRJEB36540	(19)
OBI013,OBI014, OBI017, OBI018	Sheep	Ancient	PRJEB41594	(20)
Acem1, AP45, Azer3, Direkli1, Direkli6, Gilat10, Shiqmim9, Kov27, Uiv17	Goat	Ancient	PRJEB26011	(29)
Au6, Et1, Ke14, Sp5	Donkey	Modern	PRJNA431818	(34)
Willy	Donkey	Modern	PRJEB24845	(33)
FM1798	Horse	Modern	PRJEB10098	(32)
Twilight	Horse	Modern	PRJNA205517	(31)
VHR031, VHR102, CdY2	Horse	Ancient	PRJEB31613	(30)

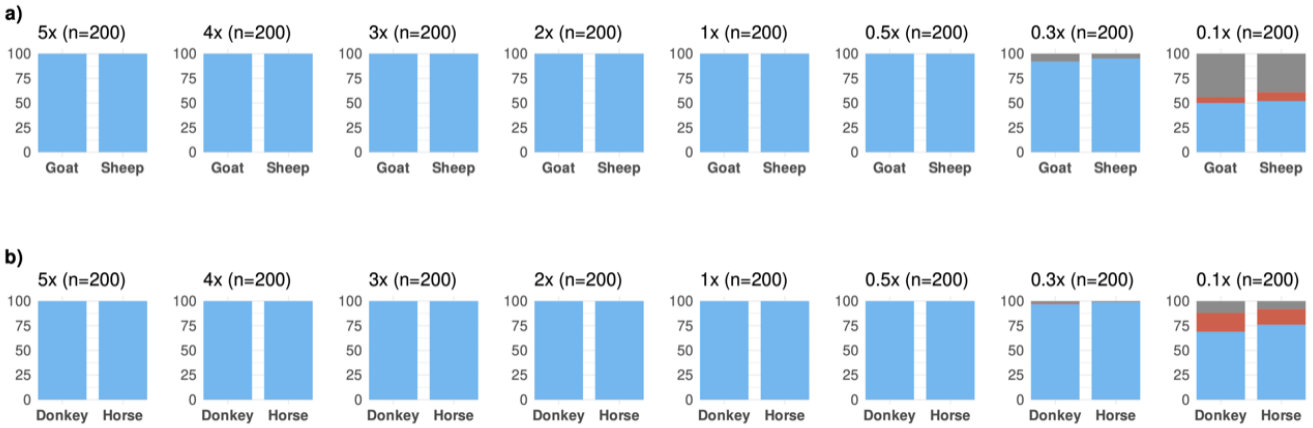

Figure 3. Results of the method applied to simulated ancient genomes at different coverages. Binomial test p-values for comparing the proportions of reads assigned to (**a**) sheep versus goat, and (**b**) and horse versus donkey. For sheep and goat, results are based on transversion substitutions, obtained through the default approach. n refers to the number of simulated genomes in each case (100 for each species in a pair). The height of the blue bar represents the number of simulated goat/donkey and sheep/horse genomes identified correctly with $p<0.05$, and the height of the red bar represents the number of unidentified cases. No cases were misidentified. The height of the grey bar represents the trials that did not contain any reads aligned to target sites, and thus could not be evaluated.

at coverages from 0.5x-5x. First, we tried the "shared reads" option, which uses only a subset of reads aligned to both genomes; here the recall was >52% at 0.5x, but reached >97% at 1x coverage (Figure S1b in Extended Data). This low recall appears to be caused by the lack of power to reject the null hypothesis due to the majority reads being eliminated by the "shared reads" approach.

Second, we repeated the analysis after eliminating polymorphic sites from the target substitution set (target sites type 2 with n=120 positions), and obtained 100% precision at 0.5x coverage (Figure S1c in Extended Data). Through the "shared" reads approach and eliminating polymorphic sites (target sites type 2), we could assign taxa with 99% recall only at 3x coverage (Figure S1d in Extended Data). This was again apparently caused by reduced power due to using fewer sites and fewer reads.

We also performed the same analysis for n=1600 datasets of horse or donkey. Using n=117 transversion substitutions (target sites type 1), we again achieved 100% precision and recall in taxonomic assignment at coverages ≥0.5x (Figure 3b). We could assign 0.3x and 0.1x coverage data with 98% and 72.5% recall, and 100% precision. Using the "shared reads" option, we obtained 97% recall at 0.5x, and 100% recall at >0.5x coverage (Figure S1f in Extended Data).

Overall, the simulations suggest that MTaxi can achieve full accuracy at mtDNA coverages ≥0.5x, and 97.5% accuracy at ≥0.3x. Conversely, limiting the analysis to subsets of reads aligned to both genomes or to non-polymorphic substitution positions reduces power, and does not increase accuracy.

Application to samples of known species identity

Sheep and goat. We tested MTaxi on n=9 published ancient sheep (*Ovis aries* / *Ovis orientalis*) samples with mtDNA coverages >0.2x, and n=9 published ancient goat (*Capra hircus* / *Capra aegagrus*) samples with mtDNA coverages >0.8x (Table 2). The samples were produced in three different laboratories and varied in their mtDNA coverage. MTaxi yielded 100% accuracy for all 18 samples using the default approach ("all reads") (Table 2). The probability of correct assignment by chance across all 18 MTaxi-classified specimens would be only 0.00038%, indicating the overall accuracy of our method (one-sided binomial test p = 3e-06).

As observed in the simulations, using the "shared reads" approach did not improve accuracy, and we could correctly assign only 15 samples, while 3 samples with the lowest coverage had too few reads for assignment at p<0.05 (Table S1, Extended Data). One sheep sample (ULU31), with mtDNA coverage at 1x, had no reads overlapping the target sites, and thus could not be analysed at all.

Interestingly, we observed 1-26% of reads misassigned with the default ("all reads") approach. These could represent homoplasy, shared polymorphism, or PCR/sequencing error. However, they do not influence the final outcome.

Since the positions that remain polymorphic within species can introduce noise in downstream analyses, for sheep and goat, we also studied the performance of the method using target sites type 2 (excluding polymorphisms; n=120 sites). Using the default approach ("all reads"), this yielded 97% accuracy; one sheep with a coverage of ~0.2x could not be identified (Table S2, Extended Data). With the "shared reads" approach, correct identification was made for only n=4 sheep and n=8 goat samples, yielding 80% accuracy (Table S3, Extended Data). Meanwhile, one goat and three sheep samples did not have any reads aligned to sheep and goat references that contained the target sites. We also noted that species-misassigned reads identified in the samples were not eliminated by this procedure (Tables S1–S4 in Extended Data). This result resonates with the above result from simulations, that removing polymorphic sites lowers statistical power but does not improve accuracy, at least in the case of sheep/goat assignment.

Horse and donkey. Applying MTaxi on n=5 horse and n=5 donkey samples, our method yielded 100% accuracy with both approaches (Table 3, Table S4 in Extended Data). The overall rate of correct assignment in this sample set appears significant (one-sided binomial test p=0.001).

Performance under alternative scenarios

In order to investigate how MTaxi would deal with extreme misidentifications, i.e. where the real taxa is not amongst the candidate pair of species, we mapped n=3 (Cdy2, Vhr031, Vhr102) ancient horse sample sequences to sheep and goat reference genomes, and ran the pipeline. The results did not indicate a significant affinity to either of the species (Table S5 in Extended Data).

We also tested MTaxi with species more closely related than sheep/goat or horse/donkey pairs, namely a pair of goat species Capra nubiana and Capra aegagrus. n=2 ancient C. aegagrus (Direkli1 and Direkli2) samples were aligned to these species' reference mitogenomes (reference assemblies - C. nubiana : https://www.ncbi.nlm.nih.gov/nuccore/OW568908.1, C. aegagrus : https://www.ncbi.nlm.nih.gov/nuccore/LR884221.1). Both individuals were identified correctly as C. aegagrus (Table S6 in Extended Data).

Whole genome comparative alignment

Comparative alignment can theoretically be a simple alternative to MTaxi when nuclear reference genomes are available. Here we explored the performance of comparative alignment using sheep/goat assignment as a model.

First, we observed that among ancient sheep BAM files used in this study, mapping results revealed inconsistencies in terms of the total number of bases aligning to each reference genome (Figure 4). Out of 9 sheep datasets with known species identity, only 4 showed a higher number of bases aligning to the sheep reference relative to the goat reference. However, we did not observe a similar inconsistency for the ancient goat samples, all of which had a higher number of bases mapped to the goat reference genome, most likely due to higher assembly quality of the goat reference. Unsurprisingly, the number of bases aligned to the nuclear genomes may not be an appropriate statistic for taxon identification between closely related taxon pairs (see Discussion).

We then analysed mismatch proportions of reads aligned to either genome. In all n=10 sheep and n=5 goat samples, including the lowest coverage samples with 0.0004x nuclear coverage, we found lower proportions of mismatches to their own reference genome. Again, this result is unlikely to happen by chance (one-sided binomial test p=3e-05). We note, however, that 3 of the sheep samples (TEP83, ULU26, ULU31) differ only marginally (by ~0.7%) in their sheep vs. goat mismatch proportions. This suggests that comparative alignment can be an alternative to mitochondrial analysis in species identification, although our results raise the possibility that its success may not be guaranteed in all circumstances.

Discussion

We showcase a simple but effective method to distinguish between closely related taxa using low coverage ancient DNA data, utilising mtDNA substitutions. Focusing on mtDNA is advantageous both in terms of high copy number and of

Table 2. MTaxi results on sheep/goat genome data of known species identity (default approach). The analysis was performed with n=197 transversion substitutions between sheep and goat, (target sites type 1). "Taxon" stands for known identity based on full genome data of the same sample (Table 1); "mtDNA coverage" shows coverage when mapping reads to the mtDNA reference of the original species (e.g. mtDNA coverage using the sheep reference for sheep data) after the duplicates have been removed; "Total assigned reads" refers to reads that could be mapped to both mtDNAs and the ones that could map only to one of the species' references with high quality, overlapped target sites, and could be assigned to either species; "Sheep reads" and "Goat reads" show the number of reads that could be unambiguously assigned to either species; "p-value" shows the two-sided binomial test p-value for the proportion of sheep and goat reads being equal, and "Identified taxon" shows the final taxon assignment.

Samples	Taxon	mtDNA coverage	Total assigned reads	Sheep reads	Goat reads	p value	Identified taxon
TEP03	Sheep	3.73327	285	279	6	<0.001	Sheep
TEP62	Sheep	45.4534	3181	2555	626	<0.001	Sheep
TEP83	Sheep	21.7152	1474	1236	238	<0.001	Sheep
OBI014 (OB20-06)	Sheep	9.97996	1043	1037	6	<0.001	Sheep
OBI018 (OB20-04)	Sheep	2.39606	235	232	3	<0.001	Sheep
OBI013 (OB20-01)	Sheep	1.25211	110	108	2	<0.001	Sheep
ULU26	Sheep	1.39023	178	171	7	<0.001	Sheep
ULU31	Sheep	1.01956	112	107	5	<0.001	Sheep
OBI017 (OB21-06)	Sheep	0.21726	29	29	0	<0.001	Sheep
Acem1	Goat	77.4073	7969	751	7218	<0.001	Goat
AP45	Goat	2.00415	217	0	217	<0.001	Goat
Azer3	Goat	49.7898	5351	144	5207	<0.001	Goat
Direkli1	Goat	58.1783	5579	488	5091	<0.001	Goat
Direkli6	Goat	106.625	10536	928	9608	<0.001	Goat
Gilat10	Goat	1.71545	171	1	170	<0.001	Goat
Shiqmim9	Goat	0.799651	84	1	83	<0.001	Goat
Kov27	Goat	2.99934	325	5	320	<0.001	Goat
Uiv17	Goat	0.909304	98	1	97	<0.001	Goat

greater variability (see Introduction). Also, the mtDNA/nuclear DNA ratio has been stated to correlate positively with a decrease in endogenous DNA content[24], suggesting that it should be more likely to obtain higher amounts of mtDNA than nuclear DNA in particularly poorly preserved samples.

Simulation results showed that MTaxi can distinguish sheep vs. goat with full accuracy at mtDNA coverages ≥0.5x and 96% accuracy at ≥0.3x. We also obtained 100% correct results with 18 ancient samples of known identity, (one sided binomial test p = 3e-06). n=2 C. aegagrus samples were also assigned correctly when compared with another goat species, C. nubiana. Likewise, simulations of ancient horse and donkey data yielded 100% and 99% accurate results at mtDNA coverages ≥0.5x and ≥0.3x, while downsampled modern and ancient domestic equid samples (n=10) of known species identity were also assigned fully correctly (one-sided binomial test p=0.001). We also ran MTaxi with n=3 ancient horse samples that were treated as ovicaprids (sheep/goat), to test the outcome in cases of possible zooarchaeological misidentifications, which yielded no significant affinity to either species. Overall, MTaxi appears as a simple and efficient tool for correct taxon identification using ultra-low coverage shotgun sequencing data.

Meanwhile, our results suggested that conservative modifications of the pipeline that involve limiting the analysis to "shared reads" or excluding polymorphic sites did not improve performance, but on the contrary reduced statistical power and recall.

Although the target sites are restricted to transversion substitutions as a means to prevent the effects of post-mortem transitions and possible homoplasies, this restriction also results in a considerable reduction in statistical power. Alternatively,

Table 3. MTaxi results on horse/donkey genome data of known species identity (default approach). The analysis was performed with n=117 transversion substitutions between horse and donkey (target sites type 1). "Taxon" stands for known identity based on full genome data of the same sample (Table 1); "mtDNA coverage" shows coverage when mapping reads to the mtDNA reference of the original species (e.g. mtDNA coverage using the horse reference for horse data) after the duplicates have been removed. Here, the present day genomes Au6, Et1, Ke14, Sp5, Willy, FM1798 and Twilight were downsampled in order to obtain c.4x, two c.2x, two c.1x, and two 0.5x genomes, while the ancient samples VHR031, VHR102, CdY2 were used as is. "Total assigned reads" refers to reads that could be mapped to both mtDNAs and the ones that could map only to one of the species' references with high quality, overlapped target sites, and could be assigned to either species; "Horse reads" and "Donkey reads" show the number of reads that could be unambiguously assigned to either species; "p-value" shows the two-sided binomial test p-value for the proportion of horse and donkey reads being equal, and "Identified taxon" shows the final taxon assignment.

Sample ID	Taxon	mtDNA coverage	Total assigned reads	Horse reads	Donkey reads	p value	Identified taxon
Au6	Donkey	0.843071	54	1	53	<0.001	Donkey
Et1	Donkey	1.96281	123	1	122	<0.001	Donkey
Ke14	Donkey	4.03893	208	5	203	<0.001	Donkey
Sp5	Donkey	0.981224	57	2	55	<0.001	Donkey
Willy	Donkey	0.490042	27	1	26	<0.001	Donkey
FM1798	Horse	1.83667	97	97	0	<0.001	Horse
Twilight	Horse	0.608403	30	29	1	<0.001	Horse
VHR031	Horse	0.319088	47	45	2	<0.001	Horse
VHR102	Horse	0.72599	86	83	3	<0.001	Horse
CdY2	Horse	1.29874	174	168	6	<0.001	Horse

MTaxi users could choose to use both transition and transversion substitutions (a) by using aDNA libraries prepared with Uracil-DNA Glycosylase (UDG) treatment to remove the uracils from ancient molecules[50], (b) by masking those positions susceptible to post-mortem damage at read ends (https://github.com/etkayapar/bamRefine), or (c) by accounting for ancient DNA damage patterns using probabilistic models[46,51].

We note that MTaxi is successful even at mitochondrial coverages of 0.3x, which is a level frequently reached in low-coverage sequencing experiments when there exists 1% endogenous DNA. For example, in the aforementioned shotgun sequencing dataset (see Introduction), we had n=226 ancient mammalian samples with 1-10% endogenous DNA (median = 3%), and each library sequenced to a size of up to 50 million total reads per sample (median = 415,146 reads); within this set 62% of the libraries reached mitochondrial DNA coverages ≥0.3x, sufficient for effective identification by MTaxi (while 50% and 84% of the libraries reached ≥0.5x and ≥0.1x, respectively).

Our observations on comparative alignment were also notable. The comparison of total number of bases mapped to sheep and goat reference genomes showed that mapping frequencies can be deceiving, even when analysing whole genome data. A sheep FASTQ file can align more widely to the goat reference genome, and the degree to which this occurs seems to vary among samples. Meanwhile, all the goat samples had higher numbers of bases mapped to the goat reference genome. The reason for the observed differences between the performance of goat and sheep samples in alignment of their respective genomes could be related to variability in reference genome qualities and/or polymorphism between the species (indeed, the N50 of the goat genome ARS1 is 26,244,591 while that for the sheep genome Oar_v3.1 is 40,376). More generally, this result indicates that taxon identification using only the number of aligned bases in comparative alignment is not reliable, although competitive mapping (aligning to the both genomes simultaneously) or using mismatch frequencies could be a more effective alternative.

Indeed, the comparison of mismatch proportions in comparative alignment appears to be a more robust approach based on our empirical sample of 15 sheep and goat samples, even at a nuclear coverage of 0.0004x (also a value typically displayed in low coverage sequencing experiments). This could be a simple solution for taxon identification if reference genomes are available for both taxa. Still, our observation that mismatch proportions can vary only marginally in some sheep samples mapped to goat (e.g. TEP83 and ULU26 in Figure 4), calls for caution in using this strategy. Moreover, MTaxi has the

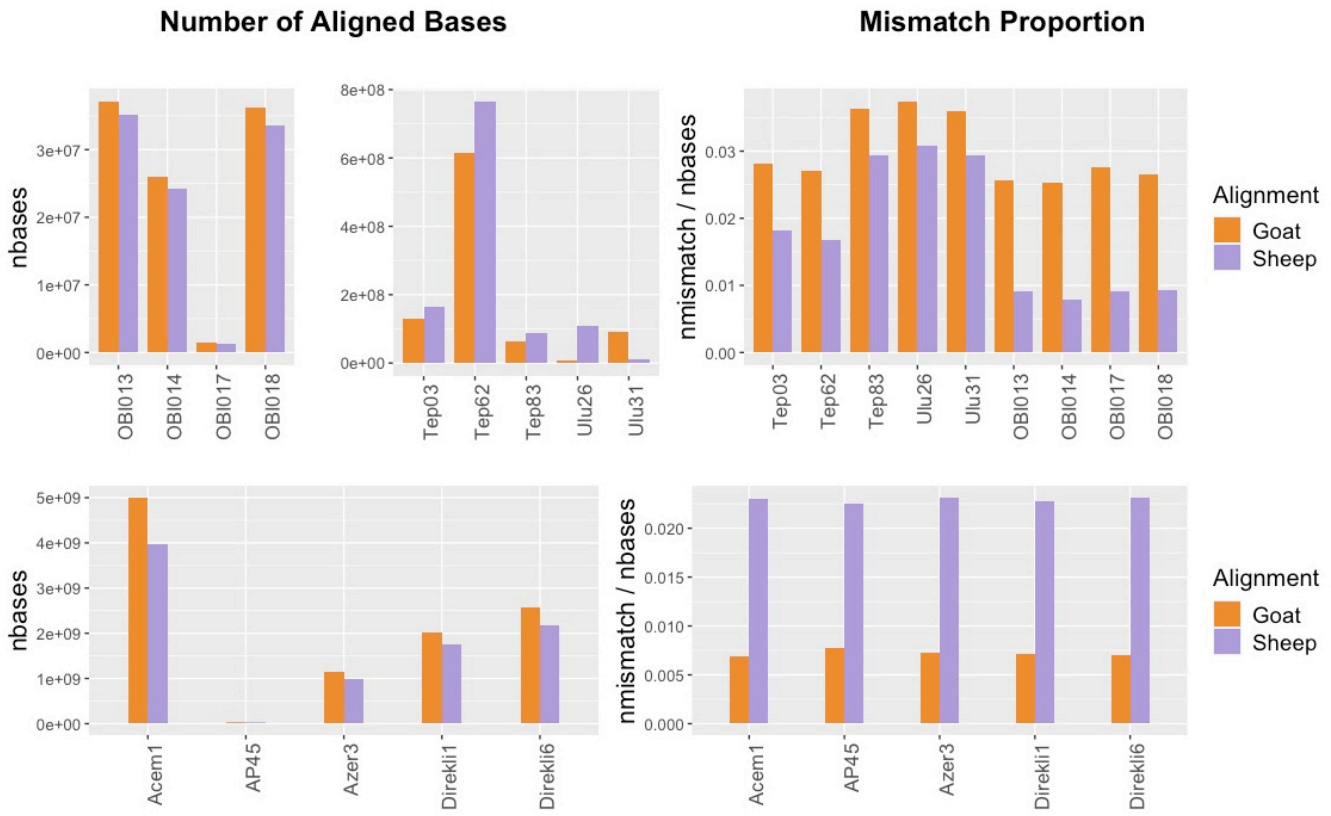

Figure 4. Whole Genome Comparative Alignment Results. Total number of bases aligned to sheep versus goat reference genomes (left panels) and the mismatch proportions (right panels) for whole genome ancient sheep (upper panels) and goat (lower panels) samples.

advantage over simple competitive mapping or mismatch frequency comparison approaches, such that it involves calculating statistical significance.

MTaxi would be expected to perform on any species pairs with a degree of divergence comparable to that of sheep and goat, and would be particularly convenient when reference nuclear genomes of one of the species is lacking, which precludes comparative alignment. Candidate taxa that pose challenges for zooarchaeological identification include several mammal species in families Cervidae (deer), Leporidae (rabbit/hare) and Bovidae (cattle/bison), and birds[52–54]. Horse and donkey are another such pair, on which we checked the performance of our method. Compared to the Zonkey pipeline[23], designed to classify ancient equid samples, MTaxi does not require a reference panel and is solely based on mitochondrial DNA data, hence an easier and faster method of classification.

In summary, the performance of MTaxi will depend on various factors, including evolutionary divergence and reference genome qualities of the species pairs, but we expect it to be an effective tool in various settings, as long as mitochondrial introgression between the species pairs can be excluded. We also note that its parameters and the data processing steps can be fine tuned to adjust for particularities of the species in question, such as the exclusion of polymorphic sites. It may be especially interesting to investigate how MTaxi performs depending on the divergence time between sister taxa.

Data availability

Underlying data

European Nucleotide Archive: Archaeogenetic analysis of Neolithic sheep from Anatolia suggests a complex demographic history since domestication. Accession number: PRJEB36540[19].

European Nucleotide Archive: Evidence for early dispersal of domestic sheep into Central Asia. Accession number: PRJEB41594[20].

European Nucleotide Archive: Ancient Goat Genomics. Study Accession: PRJEB26011[29].

European Nucleotide Archive: Dezhou Donkey de novo sequencing and assembly. Accession number: PRJNA431818[34].

European Nucleotide Archive: Improved de novo Genomic Assembly for the Domestic Donkey. Accession number: PRJEB24845[33].

European Nucleotide Archive: Evolutionary genomics and conservation of the endangered Przewalski's horse. Accession number: PRJEB10098[32].

European Nucleotide Archive: Equus Genome sequencing and assembly. Accession number: PRJNA205517[31].

European Nucleotide Archive: Tracking five millennia of horse management with extensive ancient genome time-series. Accession number: PRJEB31613[30].

Extended data

Zenodo: Extended Data for MTaxi, https://doi.org/10.5281/zenodo.8051744.

This project contains the following extended data within the file "Supplementary_tables.pdf":

- Table S1 : MTaxi results on sheep/goat genome data of known species identity using target sites type 1 using the "shared reads" approach
- Table S2 : MTaxi results on sheep/goat genome data of known species identity using target sites type 2 with the default approach ("all reads")
- Table S3 : MTaxi results on sheep/goat genome data of known species identity using target sites type 2 with the "shared reads" approach
- Table S4 : MTaxi results on horse/donkey genome data of known species identity using target sites type 1 with the "shared reads" approach
- Table S5: MTaxi results on horse genome data of known species identity (default approach), where horse is not among the candidate species.
- Table S6. MTaxi results on goat (Capra aegagrus) genome data of known species identity (default approach), where the candidate species are two goat species (Capra aegagrus and Capra nubiana)
- Figure S1. Results of the method applied to simulated ancient genomes at different coverages, including the conservative approaches.

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Software availability

Source code available from: https://github.com/goztag/MTaxi

Archived source code at time of publication: https://doi.org/10.5281/zenodo.6806307[28].

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Acknowledgements

We thank all the members of METU CompEvo group and the reviewers Dr. Kevin Daly and Dr. Burak Yelmen for their helpful suggestions.

An earlier version of this article can be found on bioRxiv (doi: https://doi.org/10.1101/2022.06.06.491147)

References

1. Braje TJ, Rick TC, Szpak P, *et al.*: **Historical ecology and the conservation of large, hermaphroditic fishes in Pacific coast kelp forest ecosystems.** *Sci Adv.* 2017; **3**(2): e1601759.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Costa T, Barri F: ***Lama guanicoe* remains from the *Chaco* ecoregion (Córdoba, Argentina): An osteological approach to the characterization of a relict wild population.** *PLoS One.* 2018; **13**(4): e0194727.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Murray MS: **Zooarchaeology and Arctic Marine Mammal Biogeography, conservation, and Management.** *Ecol Appl.* 2008; **18**(2 Suppl): S41–55.
PubMed Abstract | Publisher Full Text
4. Gifford-Gonzalez D: **Zooarchaeology and Ecology: Mortality Profiles, Species Abundance, Diversity.** In: *An Introduction to Zooarchaeology.* Cham: Springer; 2018; 475–501.
Publisher Full Text
5. Wolverton S, Lyman RL: **Applied Zooarchaeology History, Value, and Use.** In: *Conservation Biology and Applied Zooarchaeology.* Tucson: University of Arizona Press; 2012.
Reference Source
6. Steele TE: **The contributions of animal bones from archaeological sites: The past and future of zooarchaeology.** *J Archaeol Sci.* 2015; **56**: 168–76.
Publisher Full Text
7. LeFebvre MJ, Sharpe AE: **Contemporary challenges in zooarchaeological specimen identification.** *Zooarchaeology in Practice.* 2017; 35–57.
Publisher Full Text
8. Wolfhagen J, Price MD: **A probabilistic model for distinguishing between sheep and goat postcranial remains.** *J Archaeol Sci Rep.* 2017; **12**: 625–31.
Publisher Full Text
9. Parson W, Pegoraro K, Niederstätter H, *et al.*: **Species identification by means of the cytochrome b gene.** *Int J Legal Med.* 2000; **114**(1–2): 23–8.
PubMed Abstract | Publisher Full Text
10. Newman ME, Parboosingh JS, Bridge PJ, *et al.*: **Identification of archaeological animal bone by PCR/DNA analysis.** *J Archaeol Sci.* 2002; **29**(1): 77–84.
Publisher Full Text
11. Kahila Bar-Gal G, Ducos P, Kolska Horwitz L: **The application of ancient DNA analysis to identify neolithic Caprinae: A case study from the site of Hatoula, Israel.** *Int J Osteoarchaeol.* 2003; **13**(3): 120–31.
Publisher Full Text
12. Balasse M, Ambrose SH: **Distinguishing sheep and goats using dental morphology and stable carbon isotopes in C_4 Grassland Environments.** *J Archaeol Sci.* 2005; **32**(5): 691–702.
Publisher Full Text
13. Buckley M, Kansa SW: **Collagen fingerprinting of archaeological bone and teeth remains from Domuztepe, south eastern Turkey.** *Archaeol Anthropol Sci.* 2011; **3**(3): 271–80.
Publisher Full Text
14. Lan T, Lin Y, Njaramba-Ngatia J, *et al.*: **Improving species identification of ancient mammals based on next-generation sequencing data.** *Genes (Basel).* 2019; **10**(7): 509.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Orlando L, Allaby R, Skoglund P, *et al.*: **Ancient DNA analysis.** *Nat Rev Methods Primers.* 2021; **1**: 14.
Publisher Full Text
16. Allentoft ME, Collins M, Harker D, *et al.*: **The half-life of DNA in bone: Measuring decay kinetics in 158 dated fossils.** *Proc Biol Sci.* 2012; **279**(1748): 4724–33.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Ottoni C, Koon HE, Collins MJ, *et al.*: **Preservation of ancient DNA in thermally damaged archaeological bone.** *Naturwissenschaften.* 2009; **96**(2): 267–78.
PubMed Abstract | Publisher Full Text

18. Carpenter ML, Buenrostro JD, Valdiosera C, *et al.*: **Pulling out the 1%: Whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries.** *Am J Hum Genet.* 2013; **93**(5): 852–64.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Yurtman E, Özer O, Yüncü E, *et al.*: **Archaeogenetic analysis of neolithic sheep from Anatolia suggests a complex demographic history since domestication.** *Commun Biol.* 2021; **4**(1): 1279.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Taylor WT, Pruvost M, Posth C, *et al.*: **Evidence for early dispersal of domestic sheep into Central Asia.** *Nat Hum Behav.* 2021; **5**(9): 1169–79.
PubMed Abstract | Publisher Full Text

21. de Filippo C, Meyer M, Prüfer K: **Quantifying and reducing spurious alignments for the analysis of ultra-short ancient DNA sequences.** *BMC Biol.* 2018; **16**(1): 121.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Poullet M, Orlando L: **Assessing DNA Sequence Alignment Methods for Characterizing Ancient Genomes and Methylomes.** *Front Ecol Evol.* 2020; **8**.
Publisher Full Text

23. Schubert M, Mashkour M, Gaunitz C, *et al.*: **Zonkey: A simple, accurate and sensitive pipeline to genetically identify equine F1-hybrids in archaeological assemblages.** *J Archaeol Sci.* 2017; **78**: 147–57.
Publisher Full Text

24. Furtwängler A, Reiter E, Neumann GU, *et al.*: **Ratio of mitochondrial to nuclear DNA affects contamination estimates in ancient DNA analysis.** *Sci Rep.* 2018; **8**(1): 14075.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Brown WM, George M Jr, Wilson AC: **Rapid evolution of animal mitochondrial DNA.** *Proc Natl Acad Sci U S A.* 1979; **76**(4): 1967–71.
PubMed Abstract | Publisher Full Text | Free Full Text

26. **NCBI Genome Data viewer**. National Center for Biotechnology Information. U.S. National Library of Medicine; [cited 2021Dec24].
Reference Source

27. **Genome list - genome - NCBI**. National Center for Biotechnology Information. U.S. National Library of Medicine; [cited 2021Dec24].
Reference Source

28. Goztag AY: **goztag/MTaxi: (v1.0).** *Zenodo.* [Software] 2022.
http://www.doi.org/10.5281/zenodo.6806307

29. Daly KG, Delser PM, Mullin VE, *et al.*: **Ancient goat genomes reveal mosaic domestication in the Fertile Crescent.** *Science.* 2018; **361**(6397): 85–8.
PubMed Abstract | Publisher Full Text

30. Fages A, Hanghøj K, Khan N, *et al.*: **Tracking five millennia of horse management with extensive ancient genome time series.** *Cell.* 2019; **177**(6): 1419–1435.e31.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Orlando L, Ginolhac A, Zhang G, *et al.*: **Recalibrating *equus* evolution using the genome sequence of an early middle pleistocene horse.** *Nature.* 2013; **499**(7456): 74–8.
PubMed Abstract | Publisher Full Text

32. Der Sarkissian C, Ermini L, Schubert M, *et al.*: **Evolutionary genomics and conservation of the endangered przewalski's horse.** *Curr Biol.* 2015; **25**(19): 2577–83.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Renaud G, Petersen B, Seguin-Orlando A, *et al.*: **Improved de novo genomic assembly for the domestic donkey.** *Sci Adv.* 2018; **4**(4): eaaq0392.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Wang C, Li H, Guo Y, *et al.*: **Donkey genomes provide new insights into domestication and selection for coat color.** *Nat Commun.* 2020; **11**(1): 6014.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Pagès H, Aboyoun P, Gentleman R, *et al.*: **Biostrings:Efficient manipulation of biological strings**. R package version 2.54.0. 2019.

36. Broughton RE, Stanley SE, Durrett RT: **Quantification of homoplasy for nucleotide transitions and transversions and a reexamination of assumptions in weighted phylogenetic analysis.** *Syst Biol.* 2000; **49**(4): 617–27.
PubMed Abstract | Publisher Full Text

37. Page AJ, Taylor B, Delaney AJ, *et al.*: ***SNP-sites*: rapid efficient extraction of SNPs from multi-FASTA alignments.** *Microb Genom.* 2016; **2**(4): e000056.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Shi NN, Fan L, Yao YG, *et al.*: **Mitochondrial genomes of domestic animals need scrutiny.** *Mol Ecol.* 2014; **23**(22): 5393–7.
PubMed Abstract | Publisher Full Text

39. Grant JR, Stothard P: **The CGView Server: a comparative genomics tool for circular genomes.** *Nucleic Acids Res.* 2008; **36**(Web Server): W181–4.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Schubert M, Lindgreen S, Orlando L: **Adapterremoval v2: Rapid adapter trimming, identification, and read merging.** *BMC Res Notes.* 2016; **9**(1): 88.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Li H: **Aligning sequence reads, clone sequences and assembly contigs with BWA-Mem**. 2013; [cited 2023 Jun 8].
Publisher Full Text

42. Oliva A, Tobler R, Llamas B, *et al.*: **Additional evaluations show that specific *BWA-aln* settings still outperform *BWA-mem* for ancient DNA data alignment.** *Ecol Evol.* 2021; **11**(24): 18743–18748.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2010; **26**(5): 589–95.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Kircher M: **Analysis of high-throughput ancient DNA sequencing data.** *Methods Mol Biol.* 2012; **840**: 197–228.
PubMed Abstract | Publisher Full Text

45. Li H, Handsaker B, Wysoker A, *et al.*: **The sequence alignment/map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–9.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Skoglund P, Malmström H, Omrak A, *et al.*: **Genomic diversity and admixture differs for stone-age Scandinavian foragers and Farmers.** *Science.* 2014; **344**(6185): 747–50.
PubMed Abstract | Publisher Full Text

47. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–2.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Dale RK, Pedersen BS, Quinlan AR: **Pybedtools: A flexible python library for manipulating genomic datasets and Annotations.** *Bioinformatics.* 2011; **27**(24): 3423–4.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Renaud G, Hanghøj K, Willerslev E, *et al.*: **gargammel: a sequence simulator for ancient DNA.** *Bioinformatics.* 2017; **33**(4): 577–579.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Rohland N, Harney E, Mallick S, *et al.*: **Partial uracil–DNA–glycosylase treatment for screening of ancient DNA.** *Philos Trans R Soc Lond B Biol Sci.* 2015; **370**(1660): 20130624.
PubMed Abstract | Publisher Full Text | Free Full Text

51. Link V, Kousathanas A, Veeramah K, *et al.*: **Atlas: Analysis tools for low-depth and ancient samples**. Cold Spring Harbor Laboratory; 2017; [cited 2023 Jun 8].

52. Wolverton S: **Data quality in zooarchaeological faunal identification.** *J Archaeol Method Theory.* 2013; **20**: 381–96.
Publisher Full Text

53. Yang DY, Cannon A, Saunders SR: **DNA species identification of archaeological salmon bone from the Pacific Northwest Coast of North America.** *J Archaeol Sci.* 2004; **31**(5): 619–31.
Publisher Full Text

54. Bochenski ZM: **Identification of skeletal remains of closely related species: the pitfalls and solutions.** *J Archaeol Sci.* 2008; **35**(5): 1247–50.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 3

Reviewer Report 10 October 2023

https://doi.org/10.21956/openreseurope.18011.r35391

© **2023 Yelmen B.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Burak Yelmen**
[1] Laboratoire Interdisciplinaire des Sciences du Numérique, Paris-Saclay University, Gif-sur-Yvette, Île-de-France, France
[2] Estonian Biocentre, Institute of Genomics, University of Tartu, Tartu, Tartu County, Estonia

I thank the authors for their clarifications, I believe they improved the manuscript. Although the performance metrics used were defined differently (such as false negatives, which should correspond to misidentified cases for the negative class in binary classification), since the terminology is now clearly defined in the methods, I'm in favour of indexing the article.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My main research fields are population genetics, genomics, deep learning and generative modelling. My previous work involved demographic analysis of human populations based on modern and ancient genomic data. In addition, one of my ongoing research projects involve ancient DNA simulations.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 30 August 2023

https://doi.org/10.21956/openreseurope.17517.r33586

© **2023 Yelmen B.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Burak Yelmen
[1] Laboratoire Interdisciplinaire des Sciences du Numérique, Paris-Saclay University, Gif-sur-Yvette, Île-de-France, France
[2] Estonian Biocentre, Institute of Genomics, University of Tartu, Tartu, Tartu County, Estonia
[3] Laboratoire Interdisciplinaire des Sciences du Numérique, Paris-Saclay University, Gif-sur-Yvette, Île-de-France, France
[4] Estonian Biocentre, Institute of Genomics, University of Tartu, Tartu, Tartu County, Estonia

Revised manuscript is substantially improved with additional analyses and authors clarified almost all of the questions raised previously. However, there are two minor points I believe still need clarification for the performance metric calculation.

Authors' accuracy metric defined in the methods sections seems to actually correspond to F1 score (harmonic mean of recall and precision). In the manuscript, authors use the term "accuracy", which could be confusing to the readers since these two metrics are different and F1 score doesn't take into account true negatives (TN) so one class is omitted. I believe accuracy might be a more suitable metric in this case because correct predictions from both classes are taken into account.

Secondly, authors define recall as #(correctly identified cases)/#(unidentified + identified cases) and precision as (#correctly identified cases)/(#identified cases). Usually, recall is calculated as TP/(TP + FN) and precision as TP/(TP + FP). I'm not sure if I understand what corresponds to what here. I believe the confusion probably stems from the fact that these metrics are normally used for binary classification with a positive and a negative class and in authors' case, there is also the unidentified class. However, I think it's important to clarify this confusion because "recall" in authors' definition becomes "accuracy" if #(correctly identified cases) corresponds to TP + TN (in the case of sheep/goat, correctly identified sheep + correctly identified goat) and the denominator (unidentified + identified cases) seems to cover all cases. Or if #(correctly identified cases) corresponds to TP only, then recall would not be meaningful (since denominator seems to cover everything). Or maybe there is some issue with the denominator? Similarly, there is an ambiguity for the precision too. These would also affect the calculation of F1 score.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My main research fields are population genetics, genomics, deep learning and generative modelling. My previous work involved demographic analysis of human populations based on modern and ancient genomic data. In addition, one of my ongoing research projects involve ancient DNA simulations.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 18 July 2023

https://doi.org/10.21956/openreseurope.17517.r33585

© 2023 Daly K. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Kevin G. Daly
[1] Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland
[2] Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland

The changes made by the authors have improved the manuscript and accessibility of the MTaxi tool.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I work in the palaeogenomics of domestic animals, particularly goat. I have experience working with samples of extremely low endogenous DNA and sequencing reads, and have worked on species assignment of such samples. I was also responsible for the goat genomic data included in analyses in the text.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 23 May 2023

https://doi.org/10.21956/openreseurope.16140.r31295

© 2023 Yelmen B. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Burak Yelmen
[1] Laboratoire Interdisciplinaire des Sciences du Numérique, Paris-Saclay University, Gif-sur-Yvette, Île-de-France, France
[2] Estonian Biocentre, Institute of Genomics, University of Tartu, Tartu, Tartu County, Estonia
[3] Laboratoire Interdisciplinaire des Sciences du Numérique, Paris-Saclay University, Gif-sur-Yvette, Île-de-France, France
[4] Estonian Biocentre, Institute of Genomics, University of Tartu, Tartu, Tartu County, Estonia
[5] Laboratoire Interdisciplinaire des Sciences du Numérique, Paris-Saclay University, Gif-sur-Yvette, Île-de-France, France
[6] Estonian Biocentre, Institute of Genomics, University of Tartu, Tartu, Tartu County, Estonia

In this study, authors propose a method for taxonomic assignment of a sample to either of two

closely related species using ancient mitochondrial DNA (mtDNA). The method focuses on mtDNA positions with transversion substitutions between the two candidate species, assigns mtDNA reads from the sample in question to either candidate based on the read composition and uses a binomial test on these assignments to identify the true taxon. Authors test their approach both with simulated and real ancient mtDNA with varying coverages for sheep/goat and donkey/horse candidate pairs and compare their results with comparative alignment.

The manuscript is well written and covers the related literature background with appropriate references. It characterizes the problem related to the taxonomic identification of ancient samples and provides a useful solution for specific cases. I think the study is sound overall and can benefit the aDNA field, especially since the proposed method does not require high quality reference genomes and utilises more abundant mtDNA. However, some minor/moderate revisions could improve the manuscript. Below I provide my comments/questions and propose possible revisions, mainly for clarification purposes. The comments which start with an asterisk (Methods 4, Methods 6, and Results 1) are the points I believe are the most important and especially need to be addressed.

Abstract

1. "*...to classify closely related species that are compelling to distinguish...*" -> that are difficult to distinguish?

Introduction

1. Some citations could be helpful for the paragraph where alignment issues are discussed [1,2].

2. "*...and ambiguity which can significantly...*" -> typographical error

Methods

1. Figure 1: Different colors for the nucleotides in the first and second panel can be confusing to the reader. Authors might consider captioning the colors or changing the visual representation of transversion/transition substitutions in the first panel.

2. "*We call these “target sites” and are obtained from pairwise alignment of mitochondrial reference genomes.*" -> grammatical error

3. It sounds reasonable to restrict the analysis to transversions to reduce the effects of DNA damage and possible homoplasy but this causes a significant reduction in statistical power. For future studies, authors may consider including more substitutions to see how the identification is affected. One idea might be to use models for ancient DNA damage patterns (such as [3], based on non-independent C-T substitutions) to select the target positions for analysis.

4. *****Related to the previous comment, PMDtools with --deamination parameter was used in data processing but it is not clear for what purposes. Was filtering performed for likely deamination with this step?

5. Why different algorithms, BWA mem and BWA aln, were used for ancient and modern genomes? Is it based on literature [4]?

6. *****It could be useful to elaborate on the one-tailed and two-tailed binomial tests used in the

study and how unidentified cases were defined based on this in the methods section.

7. For ancient mitogenome simulations, why did authors opt not to simulate with lower coverage (<0.5x)? It could have been interesting to see how the performance changes with lower coverage, since authors also state in the discussion that lower coverage mtDNA is quite common.

8. Could authors elaborate on the rationale to downsample the equid mtDNA to 0.3x to 4x coverages?

Results

1. *****Precision and recall measurements might be confusing since there are also unidentified cases. Authors specify no misidentified cases (no false positives) but later in the results report 98% and 99% precision (referring to figure 3c and 3d). What are the false positives in this case? I think there needs to be clear definitions for all the performance metrics used.

2. "*...even at mtDNA coverages* ≥0.5x" -> Did authors mean =0.5x? Or "even" could be omitted.

3. "*...here the recall was >50% at 0.5x...*" Authors could give the exact percentage for recall here. Relatedly, please also refer to my first comment for the results.

4. Table 2. Caption ("Total assigned reads") does not match the table ("Total reads").

5. "*...would be only 0.0003%...*" I assume this is 0.5^18? I would prefer writing either ~0.00038% or ~0.0004%, not to confuse the readers.

6. "*...except one sheep, which had a coverage lower than 0.3x...*" -> In Table S2, this sheep sample seems to have ~0.1x coverage, why did authors emphasize <0.3?

7. In the comparative alignment section, authors wrote "*...differ only marginally (by ~0.7%) in their sheep vs. goat mismatch proportions.*" and concluded that "*...our results imply that its success may not be guaranteed in all circumstances.*" I believe this conclusion might not be appropriate without statistical backing.

Discussion

1. I think the degree of divergence is an interesting aspect, it could be a future direction to investigate how well the method performs with various degrees of relatedness between reference species.

2. Related to the previous comment, authors could emphasize the limitations of their method more thoroughly in the discussion. For example, how could the identification be affected if the target sample belongs to a third species? Could it be misidentified or fall under the unidentified category?

References

1. de Filippo C, Meyer M, Prüfer K: Quantifying and reducing spurious alignments for the analysis of ultra-short ancient DNA sequences.*BMC Biol*. 2018; **16** (1): 121 PubMed Abstract | Publisher Full Text
2. Poullet M, Orlando L: Assessing DNA Sequence Alignment Methods for Characterizing Ancient

Genomes and Methylomes. *Frontiers in Ecology and Evolution*. 2020; **8**. Publisher Full Text
3. Peyrégne S, Peter BM: AuthentiCT: a model of ancient DNA damage to estimate the proportion of present-day DNA contamination.*Genome Biol*. 2020; **21** (1): 246 PubMed Abstract | Publisher Full Text
4. Oliva A, Tobler R, Llamas B, Souilmi Y: Additional evaluations show that specific BWA-aln settings still outperform BWA-mem for ancient DNA data alignment.*Ecol Evol*. 2021; **11** (24): 18743-18748 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My main research fields are population genetics, genomics, deep learning and generative modelling. My previous work involved demographic analysis of human populations based on modern and ancient genomic data. In addition, one of my ongoing research projects involve ancient DNA simulations.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 18 Jun 2023

Gözde Atağ

Abstract "...to classify closely related species that are compelling to distinguish..." -> that are difficult to distinguish?

We thank Dr. Yelmen for this suggestion, we now use “difficult” instead of “compelling”.

Introduction Some citations could be helpful for the paragraph where alignment issues are discussed [1,2].

We thank Dr. Yelmen for this suggestion, we added the suggested citations where the

alignment issues are discussed.

"...and ambiguity which can significantly..." -> typographical error

We corrected the sentence as "...an ambiguity which can significantly...".

Methods Figure 1: Different colors for the nucleotides in the first and second panel can be confusing to the reader. Authors might consider captioning the colors or changing the visual representation of transversion/transition substitutions in the first panel.

We captioned the colors to avoid confusion.

"We call these "target sites" and are obtained from pairwise alignment of mitochondrial reference genomes." -> grammatical error

We corrected the sentence as follows : These "target sites" are obtained from pairwise alignment of mitochondrial reference genomes.

It sounds reasonable to restrict the analysis to transversions to reduce the effects of DNA damage and possible homoplasy but this causes a significant reduction in statistical power. For future studies, authors may consider including more substitutions to see how the identification is affected. One idea might be to use models for ancient DNA damage patterns (such as [3], based on non-independent C-T substitutions) to select the target positions for analysis.

We agree with Dr. Yelmen about the reduction in statistical power, we discuss this and possible alternative approaches in the Discussion part. The manuscript now reads as follows : Although the target sites are restricted to transversion substitutions as a means to prevent the effects of post-mortem transitions and possible homoplasies, this restriction also results in a considerable reduction in statistical power. Alternatively, MTaxi users could choose to use both transition and transversion substitutions by (a) by using aDNA libraries prepared with Uracil-DNA Glycosylase (UDG) treatment to remove the uracils from ancient molecules (1), (b) by masking those positions susceptible to post-mortem damage at read ends (https://github.com/etkayapar/bamRefine), or (c) by accounting for ancient DNA damage patterns using probabilistic models (2, 3).

*Related to the previous comment, PMDtools with --deamination parameter was used in data processing but it is not clear for what purposes. Was filtering performed for likely deamination with this step?

Thank you for pointing out this omission. PMDtools was only used on non-UDG-treated libraries for the purpose of checking the authenticity of the ancient sequences. We now clarify this point under Methods section: "In order to confirm the authenticity of the ancient DNA sequences from non-UDG-treated libraries (TEP03, TEP62, TEP83, ULU26, ULU31), we evaluated damage patterns characteristic of aDNA using PMDtools with the "--deamination" parameter; all libraries had >20% C->T and G->A damage at 5'

and 3' ends, respectively."

Why different algorithms, BWA mem and BWA aln, were used for ancient and modern genomes? Is it based on literature [4]?

We thank Dr. Yelmen for pointing this out, and we now explain our rationale in the Ancient DNA Data Processing section, based on literature. The manuscript now reads as follows : BWA mem is designed as a highly sensitive aligner for reads >70 bp (4). It has been also shown that BWA aln outperforms BWA mem in ancient DNA sequence alignment in terms of precision and proportion of the reads mapped (5). We therefore used BWA aln for the alignment of ancient DNA sequences and mem for modern DNA sequences.

*It could be useful to elaborate on the one-tailed and two-tailed binomial tests used in the study and how unidentified cases were defined based on this in the methods section.

We thank Dr. Yelmen for this suggestion. We now elaborate the use of one-tailed and two-tailed binomial tests in the Methods section as follows : Here and throughout the study, for the evaluation of the power of MTaxi using samples with known species identity, we use one-tailed binomial tests since we have an a priori hypothesis about each sample's taxon. Meanwhile, in the method itself we use a two-tailed binomial test assuming both target species are equally likely.

For ancient mitogenome simulations, why did authors opt not to simulate with lower coverage (<0.5x)? It could have been interesting to see how the performance changes with lower coverage, since authors also state in the discussion that lower coverage mtDNA is quite common.

We thank Dr. Yelmen for this suggestion. We now added results from simulations of ancient mitogenomes at 0.3x and 0.1x, and tested the performance; the results are presented in the updated Figure 3 (see Figure R1 below), and in Results and Discussion. Results showed 100% precision at both coverages, >90% and >50% recall for 0.3x and 0.1x respectively. We also updated the Discussion to suggest that MTaxi can be used with sufficient accuracy at 0.3x coverage. Figure R1: Binomial test p-values for comparing the proportions of reads assigned to (a) sheep versus goat, and (b) and horse versus donkey.

Could authors elaborate on the rationale to downsample the equid mtDNA to 0.3x to 4x coverages?

Thank you for this point. In fact in Table 3, we only downsampled seven present-day equid genomes, while the three ancient genomes were used as is. When downsampling we aimed at obtaining c.4x, two c.2x, two c.1x, and two 0.5x genomes. We now explain this in the table legend and in Methods.

Results *Precision and recall measurements might be confusing since there are also unidentified cases. Authors specify no misidentified cases (no false positives) but later in the

results report 98% and 99% precision (referring to figure 3c and 3d). What are the false positives in this case? I think there needs to be clear definitions for all the performance metrics used.

We agree with Dr. Yelmen about this ambiguity. We now define our metrics under Methods , in the newly added "Performance metric calculation" section as follows : We measured the performance of MTaxi using three metrics; precision, recall and accuracy. Precision was calculated as (#correctly identified cases)/(#identified cases), recall was calculated as #(correctly identified cases)/#(unidentified +identified cases). Accuracy was estimated using the F-scores, calculated as (2 × Precision × Recall)/(Precision + Recall)).

We also corrected the results related to the current Figure S1c (previous Figure 3c), which indeed had precision 100%. We thank the reviewer for noticing this error.

"...even at mtDNA coverages ≥0.5x" -> Did authors mean =0.5x? Or "even" could be omitted.

We omitted "even" from the sentence.

"...here the recall was >50% at 0.5x..." Authors could give the exact percentage for recall here. Relatedly, please also refer to my first comment for the results.

We now report the exact percentages as "..here the recall was 52% at 0.5x, but reached 97% at 1x coverage."

Table 2. Caption ("Total assigned reads") does not match the table ("Total reads").

We thank Dr. Yelmen for this suggestion. We corrected the table caption as "Total assigned reads".

"...would be only 0.0003%..." I assume this is 0.5^18? I would prefer writing either ~0.00038% or ~0.0004%, not to confuse the readers.

We now provide the percentage as ~0.00038%.

"...except one sheep, which had a coverage lower than 0.3x..." -> In Table S2, this sheep sample seems to have ~0.1x coverage, why did authors emphasize <0.3?

We thank Dr. Yelmen for pointing this out. The link in the manuscript is not updated to the latest version of the supplementary tables, in which the coverage is reported as 0.22x. We now updated the link to the latest version in Zenodo (https://doi.org/10.5281/zenodo.8051744).

In the comparative alignment section, authors wrote "...differ only marginally (by ~0.7%) in their sheep vs. goat mismatch proportions." and concluded that "...our results imply that its success may not be guaranteed in all circumstances." I believe this conclusion might not be appropriate without statistical backing.

Evaluating the power of comparative mapping would require a study on its own. Here we just wanted to point out that this approach may have limitations. However, we see Dr. Yelmen's point. We therefore changed the sentence as "...our results raise the possibility that its success may not be guaranteed in all circumstances."

Discussion I think the degree of divergence is an interesting aspect, it could be a future direction to investigate how well the method performs with various degrees of relatedness between reference species.

Thank you for this suggestion. We added a sentence that reads as: "It may be especially interesting to investigate how MTaxi performs depending on the divergence time between sister taxa."

Related to the previous comment, authors could emphasize the limitations of their method more thoroughly in the discussion. For example, how could the identification be affected if the target sample belongs to a third species? Could it be misidentified or fall under the unidentified category?

Thank you for this point. Following Dr. Daly's suggestion, we aligned n=3 ancient horse samples to sheep and goat reference genomes and ran MTaxi. None of the horse samples showed a significant affinity to either sheep or goat (p>0.05). The results are presented in the newly added section titled "Performance under alternative scenarios" (Table R1, Table S5). Table R1. MTaxi results on horse genome data of known species identity (default approach), where horse is not among the candidate species.

Horse samples were aligned to sheep and goat references, and the analysis was performed with transversion substitutions between sheep and goat.

Sample ID Taxon mtDNA coverage Total assigned reads Goat reads Sheep reads p value Identified taxon
Cdy2 Horse 0.319088 22 14 8 **>0.05 -** Vhr031 Horse 0.72599 4 2 2 **>0.05 -** Vhr102 Horse 1.29874 29 17 12 **>0.05 -**

References

1. Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D. Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. Philosophical Transactions of the Royal Society B: Biological Sciences. 2015;370(1660):20130624. doi:10.1098/rstb.2013.0624
2. Link V, Kousathanas A, Veeramah K, Sell C, Scheu A, Wegmann D. Atlas: Analysis tools for low-depth and ancient samples [Internet]. Cold Spring Harbor Laboratory; 2017 [cited 2023 Jun 8]. Available from: https://www.biorxiv.org/content/10.1101/105346v1
3. Skoglund P, Malmström H, Omrak A, Raghavan M, Valdiosera C, Günther T, et al. Genomic diversity and admixture differs for stone-age Scandinavian foragers and Farmers. Science. 2014;344(6185):747–50.
4. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-Mem [Internet]. 2013 [cited 2023 Jun 8]. Available from:

https://doi.org/10.48550/arXiv.1303.3997
5. Oliva A, Tobler R, Llamas B, et al.: Additional evaluations show that specific BWA-aln settings still outperform BWA-mem for ancient DNA data alignment. Ecol Evol. 2021; 11 (24): 18743-18748

Competing Interests: No competing interests were disclosed.

Reviewer Report 25 August 2022

https://doi.org/10.21956/openreseurope.16140.r29949

© 2022 Daly K. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Kevin G. Daly
[1] Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland
[2] Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland
[3] Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland

Atağ and colleagues report on MTaxi, a bioinformatic tool that compares the alignments of a single sample to two different mitochondrial sequences. A likely species identity is assigned among the pair based on a binomial test of reads overlapping transversion variant sites between the two sequences. Analysis can also be restricted to reads which map to both reference sequences, although improvement of false positive rates appears negligible in the sequence pairs tested. The tool is developed to aid species identification when working with archaeological material which may be one of two difficult-to-distinguish species, and when endogenous DNA preservation is poor.

The manuscript is exceptionally well written in its description of the broader field and the specific research question to be addressed. Citations are appropriate and support the arguments made in the text. The tool reported has use in solving specific problems in the genetic analysis of archaeozoological material when resources are limited, which is growing increasingly important as the field democratizes and broadens. The method has an excellent false positive rate and the incorrect species is unlikely to be assigned even for low coverage samples, assuming one of the two species is the correct one. There are also obvious ways that this approach could be extended in the future e.g. testing a set of reference mitochondria rather than just a pair. This could be very useful in large screening studies e.g. analyses of sediment samples.

As the manuscript stands, I would recommend several changes that can be made to the text to improve clarity to the readers regarding the method. The authors might consider a revision of their manuscript to take on some or all of these suggestions, but the method itself appears robust. A major point of clarity is the input files required; while bam and fastqs are self explanatory, the files defining the transversion variant sites are less so. Should each file (for

species 1 and 2) have the same number of variant sites, with line *n* in file 1 matching the same variant site in line *n* of file2? This could be clarified in the Github repository (https://github.com/goztag/MTaxi).

One additional request is for the inclusion in the Github repository of example data e.g. a small subset of data used in the text. This might include bam files, the reference fastqs used, and list of target sites. This would allow users to test if the software is working correctly. The github repository could also provide a more detailed description on the expected output.

The authors might consider highlighting the limitations of their approach. For example, as it stands the method is most appropriate when only two taxa are possible. It is unclear how the approach would deal with a sample outside of the taxa pair e.g. if a horse bone was misidentified during morphological assignment as either sheep or goat. If the authors were interested, they could align a small number of the ancient horse samples to the sheep and goat genome and describe the observed results. The method may also struggle with samples showing cross contamination or when the genome pairs are closely related species (e.g. *Capra aegagrus* and *Capra nubiana* whose ranges overlap in the Early Holocene Levant). Again, this could be explored further if the authors were interested.

Comments

Method

- Clarify in the text under “Target Sites” - the horse and donkey target sites are “type 1” only, correct?

- Clarify how “shared reads” between alignments are defined. Are they reads that share start and end positions between alignments? Are they reads which share id (first field in bam file)?

- Clarify the use of PMDtools in the pipeline - was it just used to check that the ancient samples were indeed ancient?

- In “Ancient DNA data processing”, there is some ambiguity as to how genotypes are called. For each target site, is the genotype of the *site* called or the “genotype” of the *read*? I personally would not describe reads as having “genotypes” - they carry bases representing an underlying genotype of the organism or genome. This also occurs under “Taxon assignment” - quoting the text “Each read can carry either reference or alternative genotypes at its target sites”. Alleles are also used to describe the base observed in a rad at a varying site, but genotype usually refers to the genome itself.

- Similarly under “Taxon assignment”, some clarity might be given that “the frequency of alternative alleles per read” refers to reads which overlap *more than one target site*. This is at least how I interpreted the sentence.

Results

- In the “Application to samples of known species identity”, there is ambiguity in the use of “accuracy”. Accuracy appears to refer to sample species assignment across the 18 tested ancient samples (“MTaxi yielded 100% accuracy for all 18 samples using the default

approach"). Accuracy is also used to refer to a subset of the sample ("yielded 100% accuracy for all samples except one"; "100% accuracy was achieved for only n=4 sheep and n=8 goat"). Accuracy should be reported for the entire sample set e.g. 17/18=94.4% and 12/18=66.6%. An alternative reading was that the accuracy was the percentage of reads assigned to the correct species e.g. for shared reads, only 12 samples showed 100% of their reads assigned to their true species. Please clarify in the manuscript.

Discussion

- The method is described as both "elaborate" and "simple". I would suggest the authors choose one! "Effective" might be a better word choice than "elaborate".

- In the final paragraph, please clarify that "mitochondrial introgression between the pairs of species" is what is meant.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I work in the palaeogenomics of domestic animals, particularly goat. I have experience working with samples of extremely low endogenous DNA and sequencing reads, and have worked on species assignment of such samples. I was also responsible for the goat genomic data included in analyses in the text.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 18 Jun 2023

Gözde Atağ

As the manuscript stands, I would recommend several changes that can be made to the text to improve clarity to the readers regarding the method. The authors might consider a revision of their manuscript to take on some or all of these suggestions, but the method itself appears robust. A major point of clarity is the input files required; while bam and fastqs are self explanatory, the files defining the transversion variant sites are less so.

Should each file (for species 1 and 2) have the same number of variant sites, with line n in file 1 matching the same variant site in line n of file2? This could be clarified in the Github repository (https://github.com/goztag/MTaxi).

We thank Dr. Daly for pointing this out, we provided a more detailed explanation about the input files containing transversion variant sites in the Github repository as follows : "transv_poly_sp1 and transv_poly_sp2 should have the same number of variant sites, each line corresponding to the same site in both files".

One additional request is for the inclusion in the Github repository of example data e.g. a small subset of data used in the text. This might include bam files, the reference fastqs used, and list of target sites. This would allow users to test if the software is working correctly. The github repository could also provide a more detailed description on the expected output.

We now added sample data which includes all necessary files and results to the Github repository. We also provided an example of expected output in README.

The authors might consider highlighting the limitations of their approach. For example, as it stands the method is most appropriate when only two taxa are possible. It is unclear how the approach would deal with a sample outside of the taxa pair e.g. if a horse bone was misidentified during morphological assignment as either sheep or goat. If the authors were interested, they could align a small number of the ancient horse samples to the sheep and goat genome and describe the observed results. The method may also struggle with samples showing cross contamination or when the genome pairs are closely related species (e.g. Capra aegagrus and Capra nubiana whose ranges overlap in the Early Holocene Levant). Again, this could be explored further if the authors were interested.

We thank Dr. Daly for this suggestion, we aligned n=3 ancient horse samples to sheep and goat references and tested how MTaxi behaves, the samples did not show any significant affinity to either species (p>0.05). We also tested how MTaxi deals with the pair of Capra aegagrus and Capra nubiana, n=2 Capra aegagrus samples were identified correctly (p<0.05) . We now added a section titled "Performance under alternative scenarios", under which the results are presented. (Tables S5,S6) Table R1. MTaxi results on horse genome data of known species identity (default approach), where horse is not among the candidate species.

Horse samples were aligned to sheep and goat references, and the analysis was performed with transversion substitutions between sheep and goat.

Sample ID Taxon mtDNA coverage Total assigned reads Goat reads Sheep reads p value Identified taxon Cdy2 Horse 0.319088 22 14 8 **>0.05 -** Vhr031 Horse 0.72599 4 2 2 **>0.05 -** Vhr102 Horse 1.29874 29 17 12 **>0.05 -**
Table R2. MTaxi results on goat (*Capra aegagrus*) genome data of known species identity (default approach), where the candidate species are two goat species (*Capra aegagrus* and *Capra nubiana*) *C. aegagrus* samples were aligned to *C. aegagrus* and *C. nubiana* references, and the analysis was performed with transversion substitutions

between *C. aegagrus* and *C. nubiana*.
Sample ID Taxon mtDNA coverage Total assigned reads *C.nubiana* reads *C.aegagrus* reads p value Identified taxon Direkli1 *Capra aegagrus* 6.80549 1645 331 1314 **<0.001** *Capra aegagrus* Direkli6 *Capra aegagrus* 10.2286 2263 445 1818 **<0.001** *Capra aegagrus*

Comments Method Clarify in the text under "Target Sites" - the horse and donkey target sites are "type 1" only, correct?

We now specify that only 'type 1' target sites were used for horse and donkey.

Clarify how "shared reads" between alignments are defined. Are they reads that share start and end positions between alignments? Are they reads which share id (first field in bam file)?

We thank Dr. Daly for this point; we now define 'shared reads' as ones that share the same IDs between the two BAM files, under the "Ancient DNA data processing" section of the manuscript.

Clarify the use of PMDtools in the pipeline - was it just used to check that the ancient samples were indeed ancient?

We now state that PMDtools was just used to check the authenticity of the ancient sequences in the Methods section as follows: "In order to confirm the authenticity of the ancient DNA sequences from non-UDG-treated libraries (TEP03, TEP62, TEP83, ULU26, ULU31), we evaluated damage patterns characteristic of aDNA using PMDtools (41) with the "--deamination" parameter; all libraries had >20% C->T and G->A damage at 5' and 3' ends, respectively."

In "Ancient DNA data processing", there is some ambiguity as to how genotypes are called. For each target site, is the genotype of the site called or the "genotype" of the read? I personally would not describe reads as having "genotypes" - they carry bases representing an underlying genotype of the organism or genome. This also occurs under "Taxon assignment" - quoting the text "Each read can carry either reference or alternative genotypes at its target sites". Alleles are also used to describe the base observed in a rad at a varying site, but genotype usually refers to the genome itself.

We agree with Dr. Daly about this ambiguity, we limited the use of the term 'genotype' to sites only, and used the term allele for the bases observed in a read.

Similarly under "Taxon assignment", some clarity might be given that "the frequency of alternative alleles per read" refers to reads which overlap more than one target site. This is at least how I interpreted the sentence.

We rewrote that section to avoid confusion. The manuscript now reads as follows: "MTaxi uses this genotype data to assign reads to either taxon, species 1 (SP1) or species 2 (SP2), based on whether they carry the alternative allele or not. If an SP1 read was aligned to the SP1 genome, we expect no alternative alleles at target sites,

and if aligned to the SP2 genome, we expect all alternative alleles. For the reads which overlap more than one target site, MTaxi retains reads carrying only alternative alleles or carrying only reference alleles, thus excluding reads with inconsistent alleles (i.e. alternative and reference alleles mixed) at target sites."

Results In the "Application to samples of known species identity", there is ambiguity in the use of "accuracy". Accuracy appears to refer to sample species assignment across the 18 tested ancient samples ("MTaxi yielded 100% accuracy for all 18 samples using the default approach"). Accuracy is also used to refer to a subset of the sample ("yielded 100% accuracy for all samples except one"; "100% accuracy was achieved for only n=4 sheep and n=8 goat"). Accuracy should be reported for the entire sample set e.g. 17/18=94.4% and 12/18=66.6%. An alternative reading was that the accuracy was the percentage of reads assigned to the correct species e.g. for shared reads, only 12 samples showed 100% of their reads assigned to their true species. Please clarify in the manuscript.

We thank Dr. Daly for pointing out this ambiguity. We now report accuracy for the entire sample set. We also added a section called "Performance metrics" to Methods, in which the calculation of accuracy is explained.

Discussion The method is described as both "elaborate" and "simple". I would suggest the authors choose one! "Effective" might be a better word choice than "elaborate".

We thank Dr. Daly for this suggestion, we now describe the method as simple and effective.

n the final paragraph, please clarify that "mitochondrial introgression between the pairs of species" is what is meant.

We clarified that we mean mitochondrial introgression between the pairs of species.

Competing Interests: No competing interests were disclosed.