# APPLIED USES OF ANCIENT DNA

EDITED BY: Nic Rawlence, Michael David Martin, Nathan Wales and Michael Knapp

PUBLISHED IN: Frontiers in Ecology and Evolution

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-933-2
DOI 10.3389/978-2-88966-933-2

## About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

# APPLIED USES OF ANCIENT DNA

Topic Editors:
**Nic Rawlence,** University of Otago, New Zealand
**Michael David Martin,** Norwegian University of Science and Technology, Norway
**Nathan Wales,** University of York, United Kingdom
**Michael Knapp,** University of Otago, New Zealand

**Citation:** Rawlence, N., Martin, M. D., Wales, N., Knapp, M., eds. (2021). Applied Uses of Ancient DNA. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88966-933-2

# Table of Contents

frontiers in Ecology and Evolution

EDITORIAL
published: 29 April 2021
doi: 10.3389/fevo.2021.679489

# Editorial: Applied Uses of Ancient DNA

*Nicolas J. Rawlence [1]*, Michael Knapp [2,3], Michael D. Martin [4] and Nathan Wales [5]*

[1] Otago Palaeogenetics Laboratory, Department of Zoology, University of Otago, Dunedin, New Zealand, [2] Department of Anatomy, University of Otago, Dunedin, New Zealand, [3] Coastal People Southern Skies Centre of Research Excellence, University of Otago, Dunedin, New Zealand, [4] Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, [5] Department of Archaeology, University of York, York, United Kingdom

**Keywords: ecological restoration, human history, method development, museums, taxonomy**

**Editorial on the Research Topic**

**Applied Uses of Ancient DNA**

The advent of ancient DNA research 35 years ago, with the publication of DNA sequences from the extinct quagga (Higuchi et al., 1984), changed the way scientists look at ecology and evolution. For the first time, evolution could be studied in "real-time" rather than relying on contemporaneous samples to study the past. The invention and continued development of high-throughput sequencing technologies (Marguiles et al., 2005) led to further technological advances in the field, making the sequencing of whole ancient genomes possible (e.g., Van der Valk et al., 2021), in turn allowing macro- and micro-evolutionary processes to be examined in ever finer detail (Mitchell and Rawlence, 2021). Ancient DNA, combined with archaeological, paleontological and paleoecological approaches, has been shown to be critical to the understanding of how ecosystems function, and how these processes may play out into the future (Stiller et al., 2010; Rawlence et al., 2012; Knapp, 2019; Thomas et al., 2019).

While the majority of ancient DNA research is focused on big-picture, curiosity driven or "blue skies" questions (e.g., Slon et al., 2018), there is a growing appreciation that ancient DNA can be used for more applied aspects of science. Metcalf et al. (2012) used ancient DNA from cutthroat trout (*Oncorhynchus clarkii*) museum specimens, and historical records, to document recent human-driven extinctions, previously cryptic lineages, translocations and range fluctuations, to highlight the importance of historical records for informing evidence-based conservation management. Increasingly, applied applications of ancient DNA include conservation paleontology (e.g., Shepherd et al., 2012), conservation archaeogenomics (Hofman et al., 2015), elucidating the impacts of over-harvesting (Oosting et al., 2019), museum collections management (e.g., Boessenkool et al., 2010), ecological restoration (e.g., Wilmshurst et al., 2014), tracking the history of invasive species (e.g., Martin et al., 2014), biosecurity (e.g., Di Donato et al., 2018), and taxonomy (e.g., Forin et al., 2018).

In this Research Topic, we have compiled 13 different contributions from across the field of applied ancient DNA pertaining to museum collections, ecological restoration, human and cultural history, and methodological development. Museum specimens are increasingly valued and sought after in genetic research (e.g., McCormack et al., 2017; Nakahama, 2020). Verry et al. highlight the importance of accurate specimen data when using historical records to inform conservation translocation decisions, where deliberate falsification or poor record keeping was

OPEN ACCESS

***Edited and reviewed by:***
*Franco Biondi,
University of Nevada, Reno,
United States*

****Correspondence:***
*Nicolas J. Rawlence
nic.rawlence@otago.ac.nz*

***Specialty section:***
*This article was submitted to
Paleoecology,
a section of the journal
Frontiers in Ecology and Evolution*

***Received:** 11 March 2021*
***Accepted:** 22 March 2021*
***Published:** 29 April 2021*

***Citation:***
*Rawlence NJ, Knapp M, Martin MD
and Wales N (2021) Editorial: Applied
Uses of Ancient DNA.
Front. Ecol. Evol. 9:679489.
doi: 10.3389/fevo.2021.679489*

common (e.g., Boessenkool et al., 2010; Miskelly, 2012; Rawlence et al., 2014). Casteneda-Rico et al. resolve the controversial taxonomy of the critically endangered Puebla deer mouse (*Peromyscus mekisturus*) using historical specimens, with consequent conservation management implications, further highlighting the importance of archival specimens when contemporary individuals are rare or non-existent. Sarkissian et al. demonstrate that mollusc shells, which are ideal indicators of environmental change and have been seemingly ignored by the ancient DNA community until recently, can in fact preserve ancient DNA up to ∼100,000 years old.

Biodiversity cannot be conserved and ecosystems restored effectively without knowledge of the pre-disturbance ecological and genetic landscape. This is especially true in tropical ecosystems, which are undergoing collapse (Barlow et al., 2018). Dommain et al. show it is possible to retrieve sedimentary ancient DNA from swamps in tropical Uganda, which can provide an additional tool to the fossil and historical record, especially where these are rare.

Ancient DNA has also allowed scientists and archaeologists to better understand our past, often in ways that are not previously discernible from the archaeological record. Vai et al. review the potential applications and limitations of experimental and computational methodologies for kinship determination. Wasef et al. show that hypothesized inequalities in Australian Aboriginal burial practices based on status, sex, and age, may instead relate to local versus non-local origins. In their review, Przelomska et al. argue a multidisciplinary approach offers an unprecedented insight into how plants and humans evolved together, how plants were used, and how ancient DNA can play an important role in the sustainability of the world's food system. Finally, Arriola et al. review the field of palaeomicrobiology which has provided a unique window to the evolution of (non)pathogenic microbes, many of which have had a profound impact on human history, and what the future may hold for this burgeoning area of ancient DNA research.

Methodological development is an important part of applied ancient DNA research, as the technique is pushed to its limits (e.g., Van der Valk et al., 2021). The methodological contributions to this Research Topic are no exception, ranging from the application of hybridization capture techniques to forensic science (Young et al.) to developing more accurate algorithms for taxonomic assignment of sedimentary ancient DNA data (Cribdon et al.), and reconstructing signatures of methylation from DNA damage in ancient botanical remains (Wagner et al.), to improving the reference-alignment of DNA sequences derived from ancient genomes and methylomes (Poullet and Orlando), and finally the development of an archival and management system for ancient DNA data (Dolle et al.).

With funding bodies, and the public, increasingly interested in the expected real-world impacts that will justify the costs of scientific research, practical applications of ancient DNA become increasingly important.

## AUTHOR CONTRIBUTIONS

NJR, MK, MDM, and NW contributed to writing and editing. All authors contributed to the article and approved the submitted version.

## FUNDING

NJR was supported by a Royal Society of New Zealand Marsden FastStart grant (Grant no. 16-UOO-096). MK was supported by a Royal Society of New Zealand Rutherford Discovery Fellowship (Grant no. 14-UOO-007). MDM was supported by the Research Council of Norway (Award 287327). NW was supported by the European Research Council Horizon 2020 Research and Innovation Programme (Grant no. 842577).

## REFERENCES

Barlow, J., Franca, F., Gardner, T. A., Hicks, C. C., Lennox, G. D., Berenguer, E., et al. (2018). The future of hyperdiverse tropical ecosystems. *Nature* 559, 517–526. doi: 10.1038/s41586-018-0301-1

Boessenkool, S., Star, B., Seddon, P. J., and Waters, J. M. (2010). Lost in translation of deliberate falsification? Genetic analyses reveal erroneous museum data for historic penguin specimens. *Proc. B* 277, 1057–1064. doi: 10.1098/rspb.2009.1837

Di Donato, A., Filippone, E., Ercolano, M. R., and Frusciante, L. (2018). Genome sequencing of ancient plants remains: findings, uses and potential applications for the study and improvement of modern crops. *Front. Plant Sci.* 9:441. doi: 10.3389/fpls.2018.00441

Forin, N., Nigris, S., Voyron, S., Girlanda, M., Vizzini, A., Casadoro, G., et al. (2018). Next generation sequencing of ancient fungal specimens: the case of the Saccardo Mycological Herbarium. *Front. Ecol. Evol.* 6:129. doi: 10.3389/fevo.2018.00129

Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A., and Wilson, A. C. (1984). DNA sequences from the quagga, and extinct member of the horse family. *Nature* 312, 282–284. doi: 10.1038/312282a0

Hofman, C. A., Rick, T. C., Fleischer, R. C., and Maldonado, J. E. (2015). Conservation archaeogenomics: ancient DNA and biodiversity in the Anthropocene. *Trends Ecol. Evol.* 30, 540–549. doi: 10.1016/j.tree.2015.06.008

Knapp, M. (2019). From a molecules' perspective – contributions of ancient DNA research to understanding cave bear biology. *Hist. Biol.* 31, 442–447. doi: 10.1080/08912963.2018.1434168

Marguiles, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380. doi: 10.1038/nature03959

Martin, M. D., Zimmer, E. A., Olsen, M. T., Foote, A. D., Gilbert, M. T. P., Brush, G. S., et al. (2014). Herbarium specimens reveal a historical shift in phylogeographic structure of common ragweed during native range disturbance. *Mol. Ecol.* 23, 1701–1716. doi: 10.1111/mec.12675

McCormack, J. E., Rodriguez-Gomez, F., Tsai, W. L. E., and Faircloth, B. C. (2017). Transforming museum specimens into genomic resources. In: Webster, M. S. (ed) The extended specimen: emerging frontiers in collections based ornithological research. *Stud. Avian Biol.* 50, 143–156.

Metcalf, J. L., Stowell, S. L., Kennedy, C. M., Rogers, K. B., McDonald, D., Keepers, J. E. K., et al. (2012). Historical stocking data and 19th Century DNA reveal human-induced changes to native diversity and distribution of cutthroat trout. *Mol. Ecol.* 21, 5194–5207. doi: 10.1111/mec.12028

Miskelly, C. M. (2012). Discovery and extinction of the South Island snipe (*Coenocorypha iredalei*) on islands around Stewart Island. *Notornis* 59, 15–31.

Mitchell, K. J., and Rawlence, N. J. (2021). Examining natural history through the lens of palaeogenomics. *Trends Ecol. Evol.* 36, 258–267. doi: 10.1016/j.tree.2020.10.005

Nakahama, N. (2020). Museum specimens: an overlooked and valuable material for conservation genetics. *Ecol. Res.* 36, 13–23. doi: 10.1111/1440-1703.12181

Oosting, T., Star, B., Barrett, J. H., Wellenreuther, M., Ritchie, P. A., Rawlence, N. J. (2019). Unlocking the potential of ancient fish DNA in the genomic era. *Evol. Appl.* 12, 1513–1522. doi: 10.1111/eva.12811

Rawlence, N. J., Kennedy, M., Waters, J. M., and Scofield, R. P. (2014). Morphological and ancient DNA analyses reveal inaccurate labels on two of Buller's bird specimens. *J. Roy. Soc. New Zeal.* 44, 163–169. doi: 10.1080/03036758.2014.972962

Rawlence, N. J., Metcalf, J., Wood, J. R., Worthy, T. H., Austin, J. A., Cooper, A. (2012). The effect of climate and environmental change on the megafaunal moa of New Zealand in the absence of humans. *Quat. Sci. Rev.* 50, 141–153. doi: 10.1016/j.quascirev.2012.07.004

Shepherd, L. D., Worthy, T. H., Tennyson, A. J. D., Scofield, R. P., Ramstad, K. M., Lambert, D. M. (2012). Ancient DNA analyses reveal contrasting phylogeographic patterns amongst kiwi (*Apteryx* spp.) and a recently extinct lineage of spotted kiwi. *PLoS ONE* 7:e42384. doi: 10.1371/journal.pone.0042384

Slon, V., Mafessoni, F., Vernot, B., de Filippo, C., Grote, S., Viola, B., et al. (2018). The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature* 561, 113–116. doi: 10.1038/s41586-018-0455-x

Stiller, M., Baryshnikov, G., Bocherens, H., Grandal d'Anglade, A., Hilpert, B., Münzel, S. C., et al. (2010). Withering away – 25,000 years of genetic decline preceded cave bear extinction. *Mol. Biol. Evol.* 27, 975–978. doi: 10.1093/molbev/msq083

Thomas, J. E., Carvalho, G. R., Haile, J., Rawlence, N. J., Martin, M. D., Ho, S. Y. W., et al. (2019). Demographic reconstruction from ancient DNA supports rapid extinction of the Great Auk. *eLife* 8:e47509. doi: 10.7554/eLife.47509.sa2

Van der Valk, T., Pecnerova, P., Diez-del-Molino, D., Bergstrom, A., Oppenheimer, J., Hartmann, S., et al. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. *Nature* 591, 265–269. doi: 10.1038/s41586-021-03224-9

Wilmshurst, J. M., Moar, N. T., Wood, J. R., Bellingham, P. J., Findlater, A. M., Robinson, J. J., et al. (2014). Use of pollen and ancient DNA as conservation baselines for offshore islands in New Zealand. *Conserv. Biol.* 28, 202–212. doi: 10.1111/cobi.12150

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 20 November 2019
doi: 10.3389/fevo.2019.00450

# Hybridization Enrichment to Improve Forensic Mitochondrial DNA Analysis of Highly Degraded Human Remains

*Jennifer M. Young, Denice Higgins and Jeremy J. Austin**

*Australian Centre for Ancient DNA, School of Biological Sciences, University of Adelaide, Adelaide, SA, Australia*

Forensic mitochondrial DNA analysis of degraded human remains using PCR-based Sanger sequencing of the control region can be challenging when endogenous DNA is highly fragmented, damaged and at very low concentration. Hybridization enrichment coupled with massively parallel sequencing (MPS) offers an effective alternative for recovering DNA fragments as small as 30 base pairs (bp) from poorly preserved samples. Here, we apply this methodology on a range of degraded human skeletal remains that have previously been analyzed using PCR-based Sanger sequencing with variable success. Our results reaffirm the benefit of targeted enrichment for analysis of degraded remains and highlight the importance of using optimized library preparation and enrichment techniques. We provide an indication of the sequencing depth required to obtain full mtDNA genomes given the complexity of the library and confirm that a second enrichment and/or a very high sequencing effort may be required to obtain full mtDNA genomes for some degraded samples.

**Keywords: forensic, DNA, mitochondrial genome, massively parallel sequencing, degraded remains, hybridization enrichment**

OPEN ACCESS

***Edited by:***
*Nathan Wales,*
*University of York, United Kingdom*

***Reviewed by:***
*Diana Ivette Cruz Dávalos,*
*Université de Lausanne, Switzerland*
*Daniel W. Foerster,*
*Leibniz Institute for Zoo and Wildlife Research (LG), Germany*

****Correspondence:***
*Jeremy J. Austin*
*jeremy.austin@adelaide.edu.au*

***Specialty section:***
*This article was submitted to Paleoecology, a section of the journal Frontiers in Ecology and Evolution*

***Received:*** *14 August 2019*
***Accepted:*** *07 November 2019*
***Published:*** *20 November 2019*

***Citation:***
*Young JM, Higgins D and Austin JJ (2019) Hybridization Enrichment to Improve Forensic Mitochondrial DNA Analysis of Highly Degraded Human Remains. Front. Ecol. Evol. 7:450. doi: 10.3389/fevo.2019.00450*

## INTRODUCTION

Mitochondrial DNA (mtDNA) analysis is useful in forensic cases involving degraded human remains as it allows inference of maternal biogeographic ancestry and comparison to maternal relatives (Gill et al., 1994; Melton et al., 2005; Nelson and Melton, 2007). Forensic mtDNA analysis typically uses PCR amplification and Sanger sequencing of the hypervariable regions of the control region, or D-loop (Kim et al., 2013; Lyons et al., 2013; Daud et al., 2014). However, massively parallel sequencing (MPS) offers the potential to obtain full mtDNA genomes to increase resolution between closely related haplogroups (Børsting et al., 2014; King et al., 2014; Yang et al., 2014; Davis et al., 2015; Parson et al., 2015; Zhou et al., 2016; Holland et al., 2017). Human identification cases that would most benefit from mtDNA analysis, such as long-term missing persons, human rights investigations and victims of disasters often involve DNA that is highly fragmented (<100 base pairs, bp). Consequently poor PCR amplification is observed due to high levels of DNA fragmentation (below target amplicon size), DNA damage and abasic sites (Gilbert et al., 2003; Maciejewska et al., 2013; Chaitanya et al., 2015). Just et al. correlated mtDNA PCR success to DNA input highlighting the failure rate of PCR based methods due to low quantity DNA (Just et al., 2015).

Hybridization enrichment (targeted in-solution enrichment, in-solution capture) coupled with MPS offers a number of benefits over PCR based methods for analysis of degraded DNA (Templeton et al., 2013; Hofreiter et al., 2015). These include the ability to retrieve sequence information from DNA fragments as short as 30 bp, well below the minimum PCR threshold, and to identify unique DNA fragments, thus removing issues related to PCR duplication. Hybridization enrichment requires conversion of fragmented genomic DNA into a DNA library by ligation of barcoded adapters (**Figure 1A**). PCR primers, complimentary to the adapters, are then used to immortalize the DNA prior to hybridization enrichment. Hybridization baits constructed from biotinylated, single-stranded RNA or DNA are subsequently used to isolate the sequences of interest (human mtDNA in this case) from the library. MtDNA genome hybridization enrichment and MPS has been tested and applied in forensic research on high quality DNA, mock degraded DNA, chemically treated DNA, telogen hairs, degraded human skeletal remains and archaeological samples (Templeton et al., 2013; Marshall et al., 2017; Shih et al., 2018). Eduardoff et al. (2017) used primer extension capture (PEC) to enrich and sequence the mtDNA control region from high quality DNA, human hairs and ancient human bones.

Despite the benefits of hybridization enrichment, low DNA quality and quantity in degraded samples still remains a limiting factor. This requires optimization of the library preparation and hybridization conditions to maximize endogenous DNA recovery. A number of studies have shown that substantial amounts of DNA in degraded samples can be lost during DNA extraction (Benoit et al., 2013; Dabney et al., 2013; Barta et al., 2014; Kemp et al., 2014; Pajnic, 2016; Glocke and Meyer, 2017). Consequently, improved extraction methods have been developed (Glocke and Meyer, 2017). DNA can also be lost during the purification steps used in DNA library preparation (DeAngelis et al., 1995; Fisher et al., 2011), which is a major concern for samples with only trace amounts of DNA present. A number of recent studies have tested alternative library preparation conditions. Fisher et al. (2011), Li et al. (2013), and Carøe et al. (2018) recommended modifications to the library preparation steps to reduce or eliminate tube transfers, replacing one or both silica spin-column clean-ups with a heat-kill step or solid phase reversible immobilization (SPRI) bead clean-ups. Single stranded library preparation methods have been advocated to reduce DNA loss, as these protocols do not include size selection steps (Gansauge et al., 2017; Glocke and Meyer, 2017). However, single stranded library methodologies are more complex and expensive than double stranded preparations. Also, it has been noted that the benefits of single stranded library methods over double stranded ones aren't as evident when examining moderately degraded samples rather than ancient samples (DNA fragments below 30 bp and endogenous content below 3%) (Sandoval-Velasco et al., 2017). Hence double stranded methods may still be preferable for forensic purposes.

Optimization of the hybridization protocol can increase the retention of target molecules. Hybridization reaction efficiency has been shown to be influenced by two key factors—hybridization temperature and annealing time—which directly impact, enrichment specificity and sensitivity (Paijmans et al., 2016). Cruz-Dávalos et al. (2017) investigated probe concentration, DNA library input amounts, annealing temperatures and incubation times and suggested that for low endogenous DNA content samples, lower (55°C) annealing temperatures and longer incubation times produced greater target enrichment. Furthermore, Brotherton et al. (2013) and Templeton et al. (2013) demonstrated that multiple rounds of enrichment can increase the number of on-target reads.

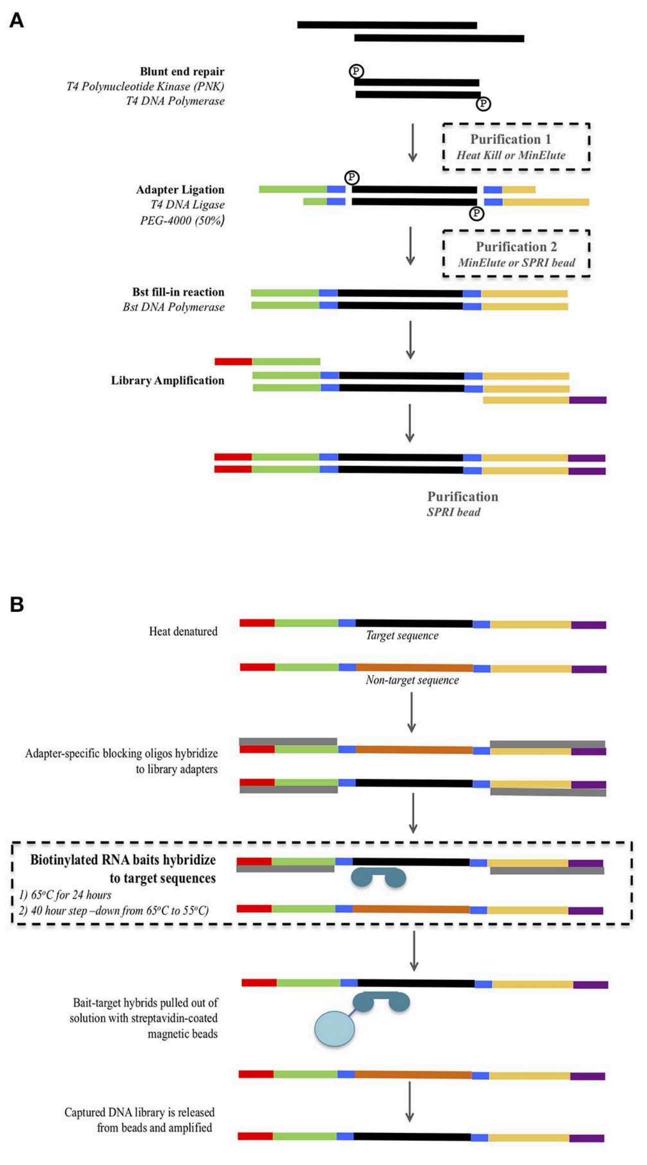


**FIGURE 1 |** Overview of the library preparation **(A)** and mtDNA hybridization enrichment **(B)** protocol. Modifications to the protocol are highlighted inside dashed boxes. For the library preparation **(A)** three combinations of purification 1 and 2 were tested: Minelute + Minelute; Heat Kill + MinElute; and Heat Kill + SPRI bead.

In this study we investigate the effects of three purification protocols during library preparation and two hybridization enrichment protocols to compare the endogenous DNA quantity, average fragment length retained, and enrichment efficiency obtained from forensic samples. The best performing protocol was then applied to a range of degraded human forensic samples (bone and teeth), which had previously been tested using traditional PCR amplification and Sanger sequencing of the mtDNA control region with varying success. For eleven samples that did not generate whole mtDNA genomes we explored the benefits of undertaking a second round of enrichment. The results presented reinforce the benefits of hybridization enrichment for highly degraded remains, such as those encountered in missing persons' cases. For severely degraded samples a second round of enrichment and/or a very high sequencing effort may be required to obtain full mtDNA genomes.

# MATERIALS AND METHODS

## Samples

Thirty-six degraded human bone and tooth samples (**Table S1**) were analyzed as part of on-going attempts to identify skeletal remains recovered from Europe and south-east Asia. Samples included femurs, humeri, "long bones" (most likely fragments of femur or humerus), and molars. All samples were recovered from soil environments, were mostly fragmentary and were all >70 years post-mortem.

## Ancient DNA Precautions Against Contamination

Contamination of samples with contemporary DNA and previously amplified mtDNA PCR products was controlled by conducting all pre-PCR work at dedicated ancient DNA facilities at the Australian Centre for Ancient DNA, University of Adelaide. No contemporary human samples or DNA had ever been present in the pre-PCR laboratory. The ancient DNA laboratory is physically separate from post-PCR laboratories and includes the use of dead-air glove boxes fitted with internal UV lights for DNA extraction, library preparation and PCR set-up, regular decontamination of all work areas and equipment with sodium hypochlorite, PPE including disposable clean-room body suit, face mask, face shield, shoe covers, and triple-gloving and strict one-way movement of personnel.

## DNA Extraction

To reduce surface contamination, the outer surfaces of the bones and teeth were UV irradiated (260 nm) for 30 min then ∼1–2 mm of the sample surface was removed using a Dremel tool with a carborundum cutting disc. Each sample was then ground to a fine powder using a Mikro-Dismembrator (Sartorius). DNA was extracted from 0.2 to 0.5 g of powdered bone or tooth as described by Brotherton et al. (2013). DNA extractions were conducted in batches of 1–7 samples with a negative extraction control.

## mtDNA Control Region PCR Amplification and Sequencing

For each sample, four short (160–187 bp, including primers) overlapping mtDNA control region amplicons (CR_S1−15,997–16,140, CR_S2−16,118–16,222, CR_S3−16,210–16,347, and CR-S4−16,288–16,409) were targeted spanning positions 15,997–16,409 that included the hypervariable region 1 (16,024–16,365). PCRs were done in 25 μL volumes containing 1× High Fidelity Buffer (ThermoFisher Scientific), 1 mg/mL Rabbit Serum Albumin (Sigma), 2 mM MgSO4, 250 μM each dNTP, 0.5 U Platinum Taq High Fidelity (ThermoFisher Scientific), 400 nM forward primer, 400 nM reverse primer, and 2 uL of DNA. Primer sequences were as published in Haak et al. (2005). Each primer included an M13 tag to enable sequencing of all amplicons with the same sequencing primers. Thermocycling conditions were 94°C for 2 min followed by 50 cycles of 94°C for 15 s, 55°C for 15 s, and 68°C for 30 s, followed by 10 min. at 68°C. All PCR attempts included negative extraction controls and a PCR negative control. PCR products were visualized via electrophoresis on a 3.5% agarose TBE gel. Samples with successful PCR amplification of three or four amplicons were sent to Australian Genome Research Facility (AGRF, Adelaide, South Australia) for purification and bi-directional Sanger sequencing. Sequence chromatograms were visualized in Geneious v9 (Biomatters) and aligned to the revised Cambridge Reference Sequence (rCRS) (Andrews et al., 1999). A consensus base was called only if covered by concordant forward and reverse reads. The sequencing success of each fragment and the resulting haplotypes are reported in **Table S1**.

## Library Preparation

Four samples (S1, S2, S3, and S4) for which all four control region fragments had been successfully PCR amplified and Sanger sequenced were subjected to library preparation using three different protocols that used different reaction clean-up steps (**Figure 1A**, see below). Double stranded libraries were constructed with truncated Illumina adapters containing dual 5-mer internal barcodes (Haak et al., 2015). For all protocols, the blunt end repair reaction, adapter ligation and *Bst* fill-in reactions were performed following the protocol from Meyer and Kircher (2010).

Library preparation includes two reaction clean-up steps: the first following the blunt end repair reaction, and the second following the adapter ligation reaction. The standard ancient DNA library preparation method uses spin-column purification (MinElute PCR purification kit, Qiagen) for both steps (Kircher et al., 2012). However, alternatives to these have been suggested in order to reduce the number of pipetting/transfer steps and potential for DNA loss. These alternatives include a heat-kill step following the blunt end repair (Fisher et al., 2011) and SPRI (solid phase reversible immobilization) beads (Fisher et al., 2011; Li et al., 2013). We tested both of these modified reaction clean-up steps as follows (**Figure 1A**). Protocol 1 used a heat-kill step (75°C for 20 min) after the blunt end repair and a spin-column purification (MinElute PCR purification kit, Qiagen) after the adapter ligation. Protocol 2 used a heat-kill step (75°C for 20 min) after the blunt end repair and an SPRI bead clean-up after the adapter ligation. Protocol 3 used the standard ancient DNA method of spin-column purification (MinElute PCR purification kit, Qiagen) for both steps. For all three protocols, the concentrations of reagents and reaction volumes in the blunt end repair, adapter ligation, and *Bst* fill-in were kept constant.

For the MinElute purification we followed the manufacturer's instructions adding a 5× volume of PB buffer to the blunt end

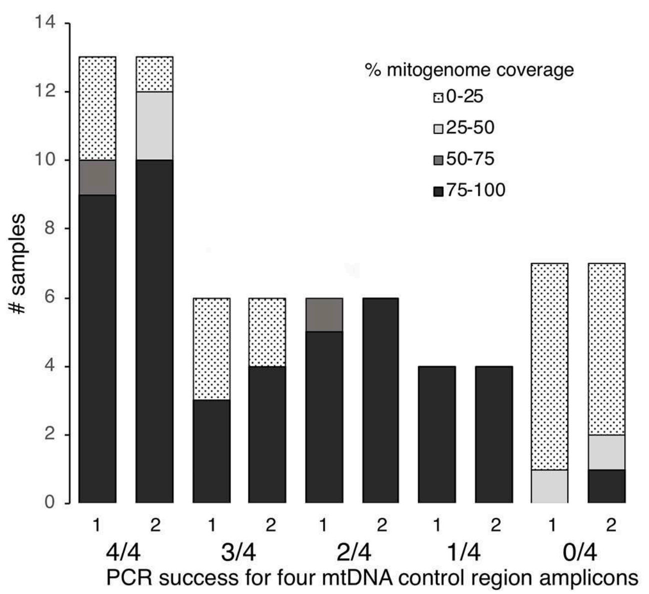


**FIGURE 2 |** MtDNA genome coverage (100–75, 75–50, 50–25, 25–0%) at >5× read depth for 36 degraded human bone and tooth samples following one or two rounds of hybridization enrichment and MPS, compared to control region PCR success (4/4, 3/4, 2/4, 1/4, or 0/4 fragments amplified).

repair or adapter ligation reaction. Purified DNA was eluted in 22.5 µL EB buffer + 0.05% Tween at 50°C. For the SPRI bead purification, we prepared a home-made bead solution containing 0.1% Sera-Mag Magnetic Speedbeads (FisherScientific), 18% PEG-8000, 1 M NaCl, 10 mM Tris, 1 mM EDTA, 0.05% Tween-20) as described by Rohland and Reich (2012). A 3× volume of the Sera-Mag/PEG solution was added to the ligation reaction, pipette mixed 10 times and incubated at room temperature for 10 min. The solution was placed on a magnetic stand for 5 min and the supernatant removed. The beads were washed twice with 150 µL of 80% ethanol, air-dried for 10 min and the purified DNA was eluted in 20 µL of EB buffer + 0.05% Tween.

Following library preparation, adapter-ligated DNA was amplified in eight separate 25 µL reactions containing 1× High Fidelity Buffer (ThermoFisher Scientific), 2 mM $MgSO_4$, 250 µM each dNTP, 500 nM IS7_short_amp.P5 (Meyer and Kircher, 2010), 500 nM IS8_short_amp.P7 (Meyer and Kircher, 2010), and 1 U of Platinum Taq DNA Polymerase, High Fidelity (ThermoFisher Scientific). Thermocycling conditions were: 94°C for 2 min, 13 cycles of 94°C for 15 s, 60°C for 15 s, 68°C for 30 s, followed by 68°C for 10 min. All eight reactions for each library were pooled and then purified using AmpureXP beads at a ratio of 1.8× as per manufacturer's instructions. We assessed the relative DNA yield of each library preparation protocol using a Qubit fluorometer (Thermo Fisher Scientific) and the dsDNA High Sensitivity Assay Kit (**Figure 2**). For each protocol, the average DNA yield and standard deviation was calculated to examine the variation across different samples.

## Hybridization Enrichment

Based on the DNA yields obtained from the three library preparation protocols, we used only the libraries generated using Protocol 3 (MinElute PCR Purification kit used at both purification steps) to examine different hybridization conditions for mtDNA genome enrichment (see Results). Libraries were enriched using Mitochondrial MYTObaits (MYcoarray) following the MYbaits V3.01 (August 2015) protocol (**Figure 1B**). Each sample was subjected to two different hybridization protocols varying in both temperature and time: (1) 65°C for 24 h and (2) a step-down approach at 65°C for 5 h, 60°C for 5 h, 55°C for 30 h (**Figure 1B**). Enriched libraries were eluted in 30 µL TLE buffer + 0.05% Tween-20. Enriched libraries were amplified in eight separate 25 µL reactions containing 1× GeneAmp PCR Buffer (ThermoFisher Scientific), 2 mM $MgCl_2$, 250 µM each dNTP, 1 U of AmpliTaq Gold DNA Polymerase (ThermoFisher Scientific), and 500 nM of forward [IS4_indPCR.P5 (Meyer and Kircher, 2010)] and reverse [7-mer indexing primer (Meyer and Kircher, 2010)] full-length Illumina adapter primers. Thermocycling conditions were: 94°C for 12 min, 13 cycles of 94°C for 30 s, 60°C for 30 s, 72°C for 45 s, followed by 72°C for 10 min. All eight reactions for each library were pooled and then purified using Ampure XP beads at a ratio of 1.8× as per manufacturer's instructions. Purified libraries were quantified using the Agilent Tapestation and samples were pooled at equimolar concentrations to form six final library pools. Library pools were quantified via real-time PCR using the KAPA Library Quantification kit before being sequenced on the Illumina MiSeq using a 300-cyle kit (150-cycle paired-end) at AGRF.

## MPS and Data Analysis

Libraries were initially de-multiplexed by the Illumina software into separate folders based on the index sequence. Sequences were de-multiplexed into specific samples using the dual P5/P7 5-mer internal barcodes and then processed using the PALEOMIX v1.0.1 pipeline (Schubert et al., 2014). AdapterRemoval v2 (Lindgreen, 2012) was used to trim adapter sequences, merge the paired reads, and eliminate all reads shorter than 25 bp. Collapsed reads were mapped to the revised Cambridge Reference Sequence (rCRS) mtDNA reference genome (NC_012920) (Andrews et al., 1999) with BWA v0.6.2. The minimum mapping quality was set to 25, seeding was disabled, and the maximum number or fraction of open gaps was set to 2. PCR duplicates (mapped reads that start and finish at the same location) were removed using rmdup_collapsed.py to retain only unique reads to avoid the effect of clonality overinflating read depths. Clonality was calculated as the percentage of mapped reads that were PCR duplicates. Unique mapped reads were visualized in Geneious v9 (Biomatters) to determine the mtDNA genome coverage and read depth for each sample and to generate a consensus sequence. For consensus calling a majority rule consensus approach was set using the "Highest Quality" option, "?" was called for bases with no coverage and "N" was called for bases with <5× read depth. Haplotypes and haplogroups were assigned from the consensus using MITOMASTER (Lott et al., 2013).

## Testing the Optimized Method on Degraded Human Remains

Based on the results obtained from the library preparation and hybridization comparisons on four degraded human samples,

**TABLE 1** | Number of retained reads, mapped reads, unique mapped reads, clonality, and average read depth for four degraded human bone samples for two different hybridization enrichment conditions.

| Sample | Hybridization conditions | Retained reads | Mapped reads (%) | Unique mapped reads (%) | Clonality (%) | Average read depth |
|---|---|---|---|---|---|---|
| S1 | 65°C/24 h | 475,361 | 31,812 (6.7) | 22,473 (4.7) | 29.4 | 130 |
| | 65–55°C/40 h | 599,380 | 182,038 (30.4) | 59,637 (10.0) | 67.2 | 318 |
| S2 | 65°C/24 h | 655,398 | 79,055 (12.1) | 50,541 (7.7) | 36.1 | 315 |
| | 65–55°C/40 h | 434,473 | 129,832 (29.9) | 73,179 (16.8) | 43.6 | 429 |
| S3 | 65°C/24 h | 1,836,038 | 107,764 (5.9) | 45,178 (2.5) | 58.1 | 264 |
| | 65–55°C/40 h | 180,761 | 16,939 (9.4) | 13,882 (7.7) | 18.1 | 96 |
| S4 | 65°C/24 h | 2,215,188 | 8,270 (0.4) | 6,086 (0.3) | 26.4 | 38 |
| | 65–55°C/40 h | 427,190 | 19,817 (4.6) | 12,376 (2.9) | 37.5 | 72 |

*All libraries were prepared using Protocol 3.*

we tested the best performing method on the remaining 32 samples, using protocol 3 (MinElute + MinElute) for the library preparation and the temperature step-down and extended time for the hybridization enrichment. Libraries were prepared in batches of eight as described in section DNA Extraction, and hybridization enrichment was performed on each library individually as described in Hybridization Enrichment. Using protocol 3, 29 of the 32 samples produced sufficient input DNA for hybridization enrichment (i.e., 100–500 ng as recommended by MYbaits). Three samples (S12, S33, and S34) produced <100 ng of input DNA (32.5, 41.2, and 69.7 ng, respectively), but were still included in the hybridization enrichment. The enriched libraries were pooled into six pools and sequenced on six Illumina MiSeq runs using a 300-cycle kit (150-cycle paired end) at the AGRF (Adelaide, Australia).

## Secondary Hybridization Enrichment

After primary enrichment and sequencing, several samples returned very low numbers of mtDNA reads. Previous work by Templeton et al. showed that a second hybridization (i.e., repeating the hybridization on the enriched DNA from the primary hybridization) can increase both the percentage and overall number of mapped mtDNA reads (Templeton et al., 2013). We replicated this work by performing a second hybridization enrichment on eleven samples (with 0–72% mitogenome coverage at >5× read depth after the first round of enrichment) using the same temperature step-down protocol as for the first enrichment (described in Hybridization Enrichment).

## Authentication of Sequencing Results

Haplotypes generated by MPS for each sample were compared to previous Sanger sequencing results, when available, and to other sequencing attempts for the same sample for concordance. The mitochondrial haplotype for each sample was also compared to each of the other samples and to our staff elimination database to detect any possible cross-contamination between samples and from staff working on the samples.

# RESULTS

## Control Region Sanger Sequencing

Of the 36 samples examined, 13 had successful PCR amplification and Sanger sequencing for all four fragments, six samples showed 75% success (i.e., three out of four fragments amplified and sequenced), six samples showed 50% success (i.e., two out of four fragments amplified) and four showed 25% success (i.e., one out of four fragments amplified) (**Table S1**). Seven samples failed to amplify for any of the four fragments (**Table S1**). Thus, nineteen of the 36 samples (53%) met our threshold of ≥75% PCR success to yield Sanger sequence data.

## Optimization of Library Preparation and Hybridization Enrichment Protocol

### Effect of Library Preparation on DNA Yield and Hybridization DNA Input

The efficiency of the three library preparation protocols was examined by quantifying the DNA yield post-library amplification. On average, protocol 1 (heat-kill + MinElute) resulted in the lowest DNA yield (9.5 ± 7.4 ng/μL), protocol 2 (heat-kill + SPRI) intermediate yield (62.8 ± 67.0 ng/μL), and protocol 3 (MinElute + MinElute) the highest (102.8 ± 37.3 ng/μL).

Low DNA yield from library preparation influenced the DNA input available for hybridization enrichment. MYbaits recommends 100–500 ng DNA input for hybridization enrichment. As a result, only libraries generated using protocol 3 were used in the subsequent hybridization enrichment experiments.

### Effect of Hybridization Enrichment Conditions on Retrieval of Mitochondrial DNA

The total number of reads, per sample, retained after quality filtering (retained reads) ranged from 180,761 to 2,215,188 with the number of unique mapped reads ranging from 6,086 to 73,179 (**Table 1**). For all samples, the 65–55°C/40 h hybridization approach increased the percentage of mapped reads by 1.6–11.5× and increased the percentage of unique mapped reads by 2.1–10.7×, compared to the 65°C/24 h approach. Average read length of mapped reads was lower (93.9 bp) using the 65–55°C/40 h hybridization compared to 65°C/24 h (100 bp). Clonality (the

**TABLE 2 |** Increase in mtDNA genome coverage and read-depth following a second round of enrichment for 11 degraded human bone and tooth samples.

| Sample | Enrichment round | Retained reads | Mapped reads | Unique mapped reads (%) | Clonality (%) | Coverage >5× (%) | Average read depth |
|---|---|---|---|---|---|---|---|
| S11 | 1 | 437,260 | 17,443 | 1,501 (<1) | 91.4 | 71.7 | 6.3× |
| | 2 | 1,492,571 | 1,210,673 | 2,093 (<1) | 99.8 | 86.2 | 9.4× |
| S12 | 1 | 183,035 | 6,476 | 659 (<1) | 89.8 | 12.3 | 2.4× |
| | 2 | 974,256 | 766,963 | 1,030 | 99.9 | 35.4 | 3.9× |
| S13 | 1 | 264,413 | 2,016 | 512 (<1) | 74.6 | 10.1 | 2.3× |
| | 2 | 1,475,117 | 889,387 | 946 (<1) | 99.9 | 44.1 | 4.5× |
| S14 | 1 | 256,187 | 9,063 | 205 (<1) | 97.7 | 0.4 | 1× |
| | 2 | 906,450 | 766,484 | 377 (<1) | 99.9 | 17.4 | 2.1× |
| S15 | 1 | 274,953 | 166 | 164 (<1) | 1.2 | 1 | 0.9× |
| | 2 | 371,271 | 293,942 | 32,016 (8.6) | 89.1 | 100 | 181× |
| S16 | 1 | 243,621 | 2,404 | 41 (<1) | 98 | 0 | 0.1× |
| | 2 | 256,002 | 120,135 | 61 (<1) | 99.9 | 0 | 0.2× |
| S19 | 1 | 204,449 | 10,492 | 252 (<1) | 97.6 | 1 | 1× |
| | 2 | 1,552,779 | 1,243,821 | 473 (<1) | 99.9 | 16 | 2.3× |
| S25 | 1 | 261,331 | 982 | 924 | 5.9 | 55.1 | 5.2× |
| | 2 | 424,533 | 363,518 | 13,607 (3.2) | 96.2 | 99.9 | 73.5× |
| S31 | 1 | 237,747 | 352 | 308 (<1) | 12.5 | 6 | 1.8× |
| | 2 | 1,915,711 | 480,379 | 2,348 (<1) | 99.5 | 98 | 12.9× |
| S33 | 1 | 176,241 | 87 | 11 (<1) | 87 | 0 | 0× |
| | 2 | 82,596 | 19,492 | 15 (<1) | 99.9 | 0 | 0.1× |
| S34 | 1 | 272,551 | 5 | 5 (<1) | 0 | 0 | 0× |
| | 2 | 498,921 | 114,621 | 508 (<1) | 99.5 | 11 | 2.2× |

percentage of mapped reads that were PCR duplicates) increased in three samples for the 65–55°C/40 h hybridization. The 65–55°C/40 h approach also increased the average read depth across the mtDNA genome by 1.4–2.5× for samples S1, S2, and S4. In contrast, the 65°C/24 h approach generated more unique reads for S3 and a higher average mtDNA genome read depth (264× compared to 96×). Based on these results, the step-down hybridization approach was selected to analyse the remaining 32 bone samples.

## mtDNA Genome Sequencing From Degraded Samples

Using protocol 3 (Minelute/Minelute cleanup) for library preparation and the 65–55°C/40 h hybridization for a single round of enrichment, we sequenced 52,697–955,346 raw reads per sample (mean = 591,400, **Table S1**). With PCR duplicates removed we obtained 1–108,585 unique mapped reads per sample (mean = 16,482, **Table S1**). Average fragment length for mapped reads varied almost 2-fold: 59.7–111.4 bp (mean = 81.2 bp, **Table S1**). With a minimum 5× read depth to call a base, we recovered full mitogenomes from 17 samples; 80–96% mitogenome coverage from four samples; 32–72% mitogenome coverage from three samples; and 0–12% mitogenome coverage from 12 samples (**Figure 2**, **Table S1**).

MtDNA haplogroups could be predicted for 24 samples with as low as 32% coverage (at >5× read depth), including 11 samples for which control region PCRs had failed on two or more fragments. The MPS results were concordant with the Sanger results for all samples where there was comparable sequence data and no sequences matched other samples or any of our staff elimination profiles (**Table S1**).

## Effect of Secondary Hybridization Enrichment on mtDNA Genome Recovery

Eleven samples with no to moderate (0–72%) mtDNA genome coverage after the first round of enrichment were subjected to a second round of enrichment. All samples showed an increase in the number of unique mapped reads and an increase in mtDNA genome coverage and average read-depth (**Table 2**, **Figure 2**). This increase in coverage appears to be related to the level of clonality following the first enrichment—the lower the clonality, the greater the increase in coverage following the second round of enrichment (**Figure S1**). Most noticeably, S15 and S31, which had 1.2 and 12.5% clonality, respectively after the first enrichment, showed an increase in mtDNA genome coverage (at >5× read depth) from 1 to 100% and 6 to 98% following the second enrichment. In contrast, samples S16, S19 and S33 with high clonality after the first enrichment (98, 97, and 87%, respectively), did not show a substantial increase in mtDNA genome coverage following a second round of enrichment. The improvement in coverage and read depth resulted in control region haplotypes and mtDNA haplogroup prediction (using coding region and control region data) for six samples that had too low coverage after a single round of enrichment.

### Effect of Sequencing Depth on mtDNA Genome Coverage

Sequencing read-depth and mtDNA genome coverage were related (**Figure 3**). Approximately 5,000 unique mapped reads were required to obtain a full mtDNA genome with >5× read depth (**Figure 3A**). As the number of unique reads increased above this threshold, there was a linear increase in average mtDNA genome read-depth (**Figure 3B**). The increased mtDNA genome coverage from 1 to 16% for sample S19 is likely due to the increased sequencing effort resulting in an 8× increase in the number of raw sequences obtained for the second enrichment compared to the first enrichment. Similarly, sample S31, had an 8× increase in the number of raw reads for the second enrichment, with a mtDNA genome coverage increase from 6 to 98%. All other samples had an increase in retained reads of between 0.1 and 1.8×.

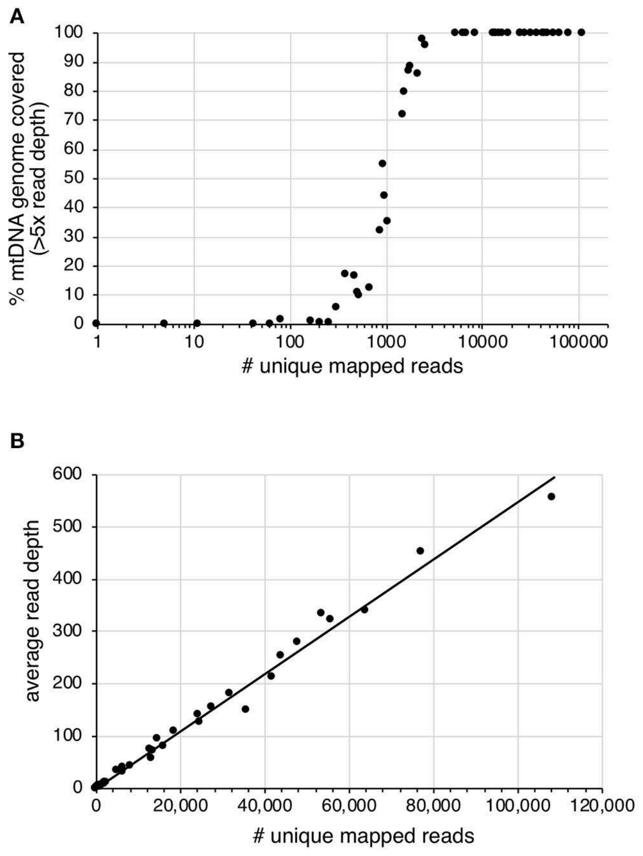


**FIGURE 3 |** The effect of increased number of unique mapped reads on **(A)** the percentage of mtDNA genome with >5× read depth and **(B)** the average read depth across the mtDNA genome.

## DISCUSSION

Human identification from degraded remains presents a number of technical issues for forensic science. The low quality and quantity of DNA present within skeletal remains, such as teeth and bone, can result in unsuccessful mtDNA sequencing using traditional approaches. Here, 17 out of 38 degraded samples (45%, **Figure 2**) showed low control region PCR amplification success and would have been excluded from further analysis under the criteria used by Just et al. (2015). This study confirms that hybridization enrichment can be used to obtain mtDNA genome data, and improve mtDNA typing success, from degraded human forensic samples. We tested alternative library preparation and hybridization conditions and applied the optimized methods, including a second round of enrichment on some samples. Sufficient mtDNA sequence was generated to produce a control region haplotype and/or predict the mtDNA haplogroup from 30 out of 36 samples. Compared to PCR and Sanger sequencing only six samples (16%) remained recalcitrant to mtDNA analysis.

MtDNA genome coverage is influenced by the proportion of on-target reads retrieved during hybridization enrichment, which in turn is affected by the endogenous DNA content of the sample and the sequenced fragment lengths. We examined three methods of improving mtDNA genome recovery (1) library preparation method, (2) hybridization conditions, and (3) a second round of enrichment. The use of enzymatic heat-kill steps and SPRI-bead clean-ups reduced the overall DNA yield from the library preparation, in many cases to levels below the minimum amount of input DNA recommended for hybridization enrichment. Despite potential for DNA loss associated with column purification, the MinElute reaction clean-ups yielded the highest amount of library DNA. The relaxed step-down hybridization conditions, combined with longer incubation times, increased on-target reads by 2–10× compared to incubation at 65°C, presumably due to more efficient hybridization of poor quality sequences to the probes at lower temperature (Wetmur, 1991; Carletti et al., 2006). The optimized method produced 0.2–17% on-target reads which was sufficient to recover complete or near-complete mtDNA genomes from highly degraded skeletal remains with a read-depth between 38 and 429×. This is an improvement on the method reported by Templeton et al. (2013) who recovered 2.8% on-target reads from a well-preserved cranium fragment, only after two-rounds of enrichment, to obtain 99.5% mtDNA at an average read depth of 20× (Templeton et al., 2013). Here, a second round of enrichment was used to increase the on-target reads and obtain complete, or near complete, mtDNA genomes from samples with low coverage after the first enrichment. We demonstrate that the efficiency of the second enrichment (in terms of the percentage of on-target reads) is influenced by the clonality at the first enrichment step. This observation will be useful in deciding whether a second enrichment would increase mtDNA genome coverage. Alternative strategies, such as low-coverage shotgun sequencing of samples combined with bioinformatic prediction of library complexity (e.g., using Preseq, http://smithlabresearch.org/software/preseq/) may assist to streamline laboratory analysis of highly degraded samples.

The percentage of on-target reads will increase the sequencing read depth across the mtDNA genome and the reliability for calling variants. Using the approach tested here, a minimum of ∼5,000 unique mapped reads with a mean read length of 81 bp were required to obtain a complete mtDNA genome with a minimum 5× read depth. Increasing unique mapped reads above this threshold increased the average sequence read depth. PCR based mtDNA genome studies have shown differential coverage, with consistent trends between samples (King et al., 2014; McElhoe et al., 2014; Parson et al., 2015). However, differential read-depth has been attributed to the amplification strategy and the positioning of the overlapping primers. There is currently no agreed minimum read-depth threshold for calling a variant in forensic mtDNA genome studies. For example, McElhoe et al. (2014) suggest a minimum read-depth threshold of 200×, King et al. (2014) applied a minimum threshold of 40× and Ring et al. use a minimum threshold of 10× (Ring et al., 2017). For PCR based approaches, these sequence read-depths are plausible as a high percentage of sequencing reads are on-target and PCR duplicates are not excluded from the analysis. In contrast, shotgun and hybridization enrichment approaches naturally have a much lower percentage of on-target sequences as the target fragments are not primarily amplified. Also, as unique molecules can be distinguished, PCR duplicates are excluded from analysis lowering apparent read-depth. Parson et al. reported >98% mapped reads with ∼70,000× read-depth using long amplicons (2–3 kb), ∼70% mapped reads with read-depth between 6,000 and 25,000× using midi-sized amplicons (62-amplicons of 300–500 bp), and <0.1% reads mapped with read-depth between 1 and 133× using shotgun sequencing (Parson et al., 2015). For the shotgun sequencing approach, a minimum read-depth of 48× was applied which allowed for 31 of the expected 34 variants to be called. Minimum sequencing read-depth thresholds should be explored to call variants from hybridization enrichment data from degraded remains. The nature of DNA within degraded remains differs from that within good quality forensic samples, and this should be considered when assessing the reliability of SNPs and calling of heteroplasmy, particularly where read-depth is low (Hanssen et al., 2017; Rathbun et al., 2017).

## CONCLUSIONS

Hybridization enrichment will deliver new capacity and capability in specialist DNA-based identification of trace and highly degraded DNA. This new approach will result in improved and more reliable identification from trace sources and decomposed human remains and will reduce costs and delays in the identification of unknown samples, improving outcomes in criminal and coronial investigations.

## DATA AVAILABILITY STATEMENT

The datasets generated in this study are available on Figshare under the following doi: https://doi.org/10.25909/5dbac1950942c.

## AUTHOR CONTRIBUTIONS

JY and JA conceived and designed the experiments. JY performed the experiments. JY, DH, and JA analyzed and interpreted the data. JA provided the samples, reagents, and equipment. JY and DH wrote the paper. JA edited the paper.

## FUNDING

This research was supported by an Australian Research Council Future Fellowship (FT100100108) and Discovery Project (DP150101664) to JA.

## ACKNOWLEDGMENTS

We thank past and present members of the Australian Centre for Ancient DNA for technical advice and assistance.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2019.00450/full#supplementary-material

**Table S1 |** PCR success of the four overlapping HVS-1 control region fragments, endogenous DNA and mtDNA genome coverage after library preparation (Protocol 3) and enrichment, and haplotypes. A PCR was regarded successful where a positive band in gel electrophoresis, however in some cases poor quality sequences were obtained (missing data is in brackets). *E2 indicates samples subject to a second round of enrichment*.

**Figure S1 |** Exponential decrease in second round enrichment improvement as the % clonality from the first enrichment increases.

## REFERENCES

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet*. 23:147. doi: 10.1038/13779

Barta, J. L., Monroe, C., Teisberg, J. E., Winters, M., Flanigan, K., and Kemp, B. M. (2014). One of the key characteristics of ancient DNA, low copy number, may be a product of its extraction. *J. Archaeol. Sci.* 46, 281–289. doi: 10.1016/j.jas.2014.03.030

Benoit, J. N., Quatrehomme, G., Carle, G. F., and Pognonec, P. (2013). An alternative procedure for extraction of DNA from ancient and weathered bone fragments. *Med. Sci. Law* 53, 100–106. doi: 10.1258/msl.2012.012026

Børsting, C., Fordyce, S. L., Olofsson, J., Mogensen, H. S., and Morling, N. (2014). Evaluation of the Ion Torrent$^{TM}$ HID SNP 169-plex: a SNP typing assay developed for human identification by second generation sequencing. *Forensic Sci. Int. Genet.* 12, 144–154. doi: 10.1016/j.fsigen.2014.06.004

Brotherton, P., Haak, W., Templeton, J., Brandt, G., Soubrier, J., Jane Adler, C., et al. (2013). Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. *Nat. Commun.* 4:1764. doi: 10.1038/ncomms2656

Carletti, E., Guerra, E., and Alberti, S. (2006). The forgotten variables of DNA array hybridization. *Trends Biotechnol.* 24, 443–448. doi: 10.1016/j.tibtech.2006.07.006

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S. T., Sinding, M. H. S., Samaniego, J. A., et al. (2018). Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* 9, 410–419. doi: 10.1111/2041-210X.12871

Chaitanya, L., Ralf, A., van Oven, M., Kupiec, T., Chang, J., Lagacé, R., et al. (2015). Simultaneous whole mitochondrial genome sequencing with short overlapping amplicons suitable for degraded DNA using the ion torrent personal genome machine. *Hum. Mutat.* 36, 1236–1247. doi: 10.1002/humu.22905

Cruz-Dávalos, D. I., Llamas, B., Gaunitz, C., Fages, A., Gamba, C., Soubrier, J., et al. (2017). Experimental conditions improving in-solution target enrichment for ancient DNA. *Mol. Ecol. Resour.* 17, 508–522. doi: 10.1111/1755-0998.12595

Dabney, J., Knapp, M., Glocke, I., Gansauge, M. T., Weihmann, A., Nickel, B., et al. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15758–15763. doi: 10.1073/pnas.1314445110

Daud, S., Shahzad, S., Shafique, M., Bhinder, M. A., Niaz, M., Naeem, A., et al. (2014). Optimization and validation of PCR protocol for three hypervariable regions (HVI, HVII and HVIII) in human mitochondrial DNA. *Adv. Life Sci.* 1, 165–170.

Davis, C., Peters, D., Warshauer, D., King, J., and Budowle, B. (2015). Sequencing the hypervariable regions of human mitochondrial DNA using massively parallel sequencing: enhanced data acquisition for DNA samples encountered in forensic testing. *Legal Med.* 17, 123–127. doi: 10.1016/j.legalmed.2014.10.004

DeAngelis, M. M., Wang, D. G., and Hawkins, T. L. (1995). Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res.* 23, 4742–4743. doi: 10.1093/nar/23.22.4742

Eduardoff, M., Xavier, C., Strobl, C., Casas-Vargas, A., and Parson, W. (2017). Optimized mtDNA control region primer extension capture analysis for forensically relevant samples and highly compromised mtDNA of different age and origin. *Genes* 8:E237. doi: 10.3390/genes8100237

Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T. M., et al. (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* 12:R1. doi: 10.1186/gb-2011-12-1-r1

Gansauge, M.-T., Gerber, T., Glocke, I., Korlević, P., Lippik, L., Nagel, S., et al. (2017). Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res.* 45:e79. doi: 10.1093/nar/gkx033

Gilbert, M. T., Willerslev, E., Hansen, A. J., Barnes, I., Rudbeck, L., Lynnerup, N., et al. (2003). Distribution patterns of postmortem damage in human mitochondrial DNA. *Am. J. Hum. Genet.* 72, 32–47. doi: 10.1086/345378

Gill, P., Ivanov, P. L., Kimpton, C., Piercy, R., Benson, N., Tully, G., et al. (1994). Identification of the remains of the Romanov family by DNA analysis. *Nat Genet.* 6, 130–135. doi: 10.1038/ng0294-130

Glocke, I., and Meyer, M. (2017). Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. *Genome Res.* 27, 1230–1237. doi: 10.1101/gr.219675.116

Haak, W., Forster, P., Bramanti, B., Matsumura, S., Brandt, G., Tanzer, M., et al. (2005). Ancient DNA from the first European farmers in 7500-year-old neolithic sites. *Science* 310, 1016–1018. doi: 10.1126/science.1118725

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522, 207–211. doi: 10.1038/nature14317

Hanssen, E. N., Lyle, R., Egeland, T., and Gill, P. (2017). Degradation in forensic trace DNA samples explored by massively parallel sequencing. *Forensic Sci. Int. Genet.* 27, 160–166. doi: 10.1016/j.fsigen.2017.01.002

Hofreiter, M., Paijmans, J. L., Goodchild, H., Speller, C. F., Barlow, A., Fortes, G. G., et al. (2015). The future of ancient DNA: technical advances and conceptual shifts. *BioEssays* 37, 284–293. doi: 10.1002/bies.201400160

Holland, M. M., Pack, E. D., and McElhoe, J. A. (2017). Evaluation of GeneMarker® HTS for improved alignment of mtDNA MPS data, haplotype determination, and heteroplasmy assessment. *Forensic Sci. Int. Genet.* 28, 90–98. doi: 10.1016/j.fsigen.2017.01.016

Just, R. S., Scheible, M. K., Fast, S. A., Sturk-Andreaggi, K., Rock, A. W., Bush, J. M., et al. (2015). Full mtGenome reference data: development and characterization of 588 forensic-quality haplotypes representing three US populations. *Forensic Sci. Int. Genet.* 14, 141–155. doi: 10.1016/j.fsigen.2014.09.021

Kemp, B. M., Winters, M., Monroe, C., and Barta, J. L. (2014). How much DNA is lost? Measuring DNA loss of short-tandem-repeat length fragments targeted by the PowerPlex 16(R) system using the Qiagen MinElute Purification Kit. *Hum. Biol.* 86, 313–329. doi: 10.13110/humanbiology.86.4.0313

Kim, N. Y., Lee, H. Y., Park, S. J., Yang, W. I., and Shin, K.-J. (2013). Modified Midi- and mini-multiplex PCR systems for mitochondrial DNA control region sequence analysis in degraded samples. *J. Forensic Sci.* 58, 738–743. doi: 10.1111/1556-4029.12062

King, J. L., LaRue, B. L., Novroski, N. M., Stoljarova, M., Seo, S. B., Zeng, X., et al. (2014). High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. *Forensic Sci. Int. Genet.* 12, 128–135. doi: 10.1016/j.fsigen.2014.06.001

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 40:e3. doi: 10.1093/nar/gkr771

Li, C., Hofreiter, M., Straube, N., Corrigan, S., and Naylor, G. J. (2013). Capturing protein-coding genes across highly divergent species. *Biotechniques* 54, 321–326. doi: 10.2144/000114039

Lindgreen, S. (2012). AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res. Notes* 5:337. doi: 10.1186/1756-0500-5-337

Lott, M. T., Leipzig, J. N., Derbeneva, O., Xie, H. M., Chalkia, D., Sarmady, M., et al. (2013). mtDNA variation and analysis using mitomap and mitomaster. *Curr. Protoc. Bioinformatics* 23, 21–26. doi: 10.1002/0471250953.bi0123s44

Lyons, E. A., Scheible, M. K., Sturk-Andreaggi, K., Irwin, J. A., and Just, R. S. (2013). A high-throughput Sanger strategy for human mitochondrial genome sequencing. *BMC Genomics* 14:881. doi: 10.1186/1471-2164-14-881

Maciejewska, A., Jakubowska, J., and Pawłowski, R. (2013). Whole genome amplification of degraded and nondegraded DNA for forensic purposes. *Int. J. Legal Med.* 127, 309–319. doi: 10.1007/s00414-012-0764-9

Marshall, C., Sturk-Andreaggi, K., Daniels-Higginbotham, J., Oliver, R. S., Barritt-Ross, S., and McMahon, T. P. (2017). Performance evaluation of a mitogenome capture and Illumina sequencing protocol using non-probative, case-type skeletal samples: implications for the use of a positive control in a next-generation sequencing procedure. *Forensic Sci. Int. Genet.* 31, 198–206. doi: 10.1016/j.fsigen.2017.09.001

McElhoe, J. A., Holland, M. M., Makova, K. D., Su, M. S.-W., Paul, I. M., Baker, C. H., et al. (2014). Development and assessment of an optimized next-generation DNA sequencing approach for the mtgenome using the Illumina MiSeq. *Forensic Sci. Int. Genet.* 13, 20–29. doi: 10.1016/j.fsigen.2014.05.007

Melton, T., Dimick, G., Higgins, B., Lindstrom, L., and Nelson, K. (2005). Forensic mitochondrial DNA analysis of 691 casework hairs. *J. Forensic Sci.* 50, 73–80. doi: 10.1520/JFS2004230

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protoc.* 2010:pdb.prot5448. doi: 10.1101/pdb.prot5448

Nelson, K., and Melton, T. (2007). Forensic mitochondrial DNA analysis of 116 casework skeletal samples. *J. Forensic Sci.* 52, 557–561. doi: 10.1111/j.1556-4029.2007.00407.x

Paijmans, J. L., Fickel, J., Courtiol, A., Hofreiter, M., and Forster, D. W. (2016). Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. *Mol. Ecol. Resour.* 16, 42–55. doi: 10.1111/1755-0998.12420

Pajnic, I. Z. (2016). Extraction of DNA from human skeletal material. *Methods Mol. Biol.* 1420, 89–108. doi: 10.1007/978-1-4939-3597-0_7

Parson, W., Huber, G., Moreno, L., Madel, M.-B., Brandhagen, M. D., Nagl, S., et al. (2015). Massively parallel sequencing of complete mitochondrial genomes from hair shaft samples. *Forensic Sci. Int. Genet.* 15, 8–15. doi: 10.1016/j.fsigen.2014.11.009

Rathbun, M. M., McElhoe, J. A., Parson, W., and Holland, M. M. (2017). Considering DNA damage when interpreting mtDNA heteroplasmy in deep sequencing data. *Forensic Sci. Int. Genet.* 26, 1–11. doi: 10.1016/j.fsigen.2016.09.008

Ring, J. D., Sturk-Andreaggi, K., Peck, M. A., and Marshall, C. (2017). A performance evaluation of Nextera XT, and KAPA HyperPlus for rapid Illumina library preparation of long-range mitogenome amplicons. *Forensic Sci. Int. Genet.* 29, 174–180. doi: 10.1016/j.fsigen.2017.04.003

Rohland, N., and Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* 22, 939–946. doi: 10.1101/gr.128124.111

Sandoval-Velasco, M., Lundstrøm, I. K. C., Wales, N., Ávila-Arcos, M. C., Schroeder, H., and Gilbert, M. T. P. (2017). Relative performance of two DNA extraction and library preparation methods on archaeological human teeth samples. *STAR* 3, 80–88. doi: 10.1080/20548923.2017.1388551

Schubert, M., Ermini, L., Der Sarkissian, C., Jonsson, H., Ginolhac, A., Schaefer, R., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* 9, 1056–1082. doi: 10.1038/nprot.2014.063

Shih, Y. S., Bose, N., Gonçalves, B. A., Erlich, A. H., and Calloway, D. C. (2018). Applications of probe capture enrichment next generation sequencing for whole mitochondrial genome and 426 nuclear SNPs for forensically challenging samples. *Genes* 9:E49. doi: 10.3390/genes9010049

Templeton, J. E., Brotherton, P. M., Llamas, B., Soubrier, J., Haak, W., Cooper, A., et al. (2013). DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. *Investig. Genet.* 4:26. doi: 10.1186/2041-2223-4-26

Wetmur, J. G. (1991). DNA probes: applications of the principles of nucleic acid hybridization. *Crit. Rev. Biochem. Mol. Biol.* 26, 227–259. doi: 10.3109/10409239109114069

Yang, Y., Xie, B., and Yan, J. (2014). Application of next-generation sequencing technology in forensic science. *Genomics Proteomics Bioinformatics* 12, 190–197. doi: 10.1016/j.gpb.2014.09.001

Zhou, Y., Guo, F., Yu, J., Liu, F., Zhao, J., Shen, H., et al. (2016). Strategies for complete mitochondrial genome sequencing on Ion Torrent PGM platform in forensic sciences. *Forensic Sci. Int. Genet.* 22, 11–21. doi: 10.1016/j.fsigen.2016.01.004

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 19 December 2019
doi: 10.3389/fevo.2019.00496

# Who, Where, What, Wren? Using Ancient DNA to Examine the Veracity of Museum Specimen Data: A Case Study of the New Zealand Rock Wren (*Xenicus gilviventris*)

*Alexander J. F. Verry*[1]*, *Lachie Scarsbrook*[1], *R. Paul Scofield*[2], *Alan J. D. Tennyson*[3], *Kerry A. Weston*[4], *Bruce C. Robertson*[5] and *Nicolas J. Rawlence*[1]

[1] Otago Palaeogenetics Laboratory, Department of Zoology, University of Otago, Dunedin, New Zealand, [2] Canterbury Museum, Christchurch, New Zealand, [3] National Museum of New Zealand Te Papa Tongarewa, Wellington, New Zealand, [4] Biodiversity Group, Department of Conservation, Christchurch, New Zealand, [5] Department of Zoology, University of Otago, Dunedin, New Zealand

OPEN ACCESS

**Edited by:**
Laura Parducci,
Uppsala University, Sweden

**Reviewed by:**
Cristiano Vernesi,
Fondazione Edmund Mach, Italy
Michael Hofreiter,
University of Potsdam, Germany

***Correspondence:**
Alexander J. F. Verry
verryalex94@gmail.com

**Specialty section:**
This article was submitted to
Paleoecology,
a section of the journal
Frontiers in Ecology and Evolution

**Received:** 01 July 2019
**Accepted:** 04 December 2019
**Published:** 19 December 2019

**Citation:**
Verry AJF, Scarsbrook L, Scofield RP, Tennyson AJD, Weston KA, Robertson BC and Rawlence NJ (2019) Who, Where, What, Wren? Using Ancient DNA to Examine the Veracity of Museum Specimen Data: A Case Study of the New Zealand Rock Wren (Xenicus gilviventris). Front. Ecol. Evol. 7:496. doi: 10.3389/fevo.2019.00496

Museum specimens provide a record of past species distribution and are an increasingly important resource for conservation genetic research. The scientific value of these specimens depends upon the veracity of their associated data and can be compromised by inaccurate details; including taxonomic identity, collection locality, and collector. New Zealand contains many endemic species that have been driven to extinction or reduced to relict distributions following the arrival of humans and mammalian predators, including the Acanthisittid wrens (of which only two of the eight described species presently persist). One of these is the New Zealand rock wren (*Xenicus gilviventris*), currently classified as an endangered species and experiencing ongoing population declines. Here we analyze ancient DNA retrieved from New Zealand rock wren museum skins to establish the veracity of their recorded collection localities—New Zealand rock wrens exhibit strong north-south genetic structuring along the Southern Alps of New Zealand's South Island. We include the only specimen reportedly collected from New Zealand's North Island, outside the known range of New Zealand rock wrens, specimens collected by Henry Hamersley Travers, a collector known for poor record keeping and potentially fraudulent specimen data, and type specimens of proposed *Xenicus* taxa. Multiple instances of inaccurate collection locality were detected, including that of the New Zealand rock wren reportedly collected from the North Island, which matches individuals from the southern South Island. Syntypes of *X. haasti*, and a syntype of *X. gilviventris* clustered with individuals belonging to the northern New Zealand rock wren lineage. Our results suggest that New Zealand rock wrens have not been historically extirpated from New Zealand's North Island, and that caution must be taken when utilizing museum specimens to inform conservation management decisions. Additionally, we describe the type locality of both *X. gilviventris* and *X. haasti*, with genetic and historical evidence suggesting that the specimens used to describe these taxa were collected from the headwaters of the Rakaia River. This study demonstrates that ancient DNA analysis can add value to museum specimens by revealing incorrect specimen data and inform the conservation management and taxonomy of endangered species.

**Keywords: conservation, translocation, re-introduction, taxonomy, museum skin**

## INTRODUCTION

Museum specimens constitute a valuable resource for conservation focused research, providing a record of past species' distributions (Shaffer et al., 1998). This record can be used to inform present-day conservation management of biodiversity. However, the reliability of this record can be compromised by inaccurate (or non-existent) data (e.g., collection locality or taxonomic status) due to poor record keeping, or deliberate falsification (Boessenkool et al., 2009). Detection of unreliable data can often be difficult if the only information readily available are the collector's notes (Barbanera et al., 2016), with specimen labels prone to error (Winker, 2000). The majority of errors concern taxonomic identity or collection locality, and can lead to misconceptions about a species' distribution (Graham et al., 2004). Advances in DNA extraction techniques and sequencing technologies have facilitated the use of historical museum specimens in conservation genetic studies (Wandeler et al., 2007). Retrieval of ancient DNA (aDNA) from such specimens can be used to examine temporal changes in genetic diversity (Thomas et al., 1990), test for congruence between collection locality and genetic data (Boessenkool et al., 2009), determine the provenance of specimens for which locality data are lacking (Shepherd et al., 2013), and assign mislabeled specimens to the correct taxonomic unit (Rawlence et al., 2014a).

Ancient DNA analysis is also an important tool for the conservation of biodiversity, as knowledge of past biodiversity, geographic range expansions/contractions, and the factors that lead to population declines or extinctions is crucial for making informed management decisions (Leonard, 2008; Grealy et al., 2017). One potential tool available to conservation managers are species re-introductions, a form of species translocation whereby individuals of a particular species are intentionally translocated to an area of their former distribution in order to establish a new population and re-introduce the species to their former range (IUCN, 2013). To determine whether a species translocation constitutes a re-introduction, or a novel introduction (the introduction of a species to an area outside their known range) an accurate record of the past and present distribution of the species is key. Cooper et al. (1996) used aDNA to demonstrate that the Laysan duck (*Anas laysanensis*), historically known only from Laysan Island, Hawaii, was formerly widespread throughout the Hawaiian Islands and argued for the re-introduction of the anatid to its former range. Species re-introductions often form an integral part of conservation management plans for isolated island ecosystems like Hawaii and New Zealand, as many species have been driven to extinction or reduced to relict populations following anthropogenic impact (Olson and James, 1982; Holdaway, 1989).

The New Zealand rock wren (hereafter NZ rock wren; *Xenicus gilviventris*) is a small alpine passerine, and one of only two extant members of the endemic New Zealand wrens (Acanthisittidae), the sister taxa to all other passerines (Barker et al., 2004; Hackett et al., 2008; Jarvis et al., 2014). NZ rock wrens have decreased in both range and abundance over the past 100 years (Michelsen-Heath and Gaze, 2007), and are currently threatened with extinction, primarily due to predation by introduced stoats (*Mustela erminea*) (Little et al., 2017; O'Donnell et al., 2017; Weston et al., 2018). Presently, NZ rock wrens are only found within alpine habitat throughout mountainous regions of New Zealand's South Island (Gill et al., 2010). Subfossil *Xenicus* remains [attributable to either rock wren or the congeneric bush wrens (*X. longipes* subspp.)] are known from throughout the North and South Islands of New Zealand, but have not been identified to species level (Worthy and Holdaway, 2002). It is currently unknown whether NZ rock wrens once inhabited the North Island. A study of present-day NZ rock wren phylogeography (covering the known geographical range of the species) revealed strong phylogeographic structure in both mitochondrial and nuclear (microsatellite) DNA, with two highly divergent northern and southern lineages which diverged two million years ago with minimal gene flow; the contact zone occurring in the central region of the South Island's Southern Alps near Aoraki/Mount Cook (Weston and Robertson, 2015). Individuals from southern Fiordland (Lake Roe and Lake MacArthur) also formed a monophyletic clade within the southern lineage (Weston and Robertson, 2015). Intriguingly, NZ rock wrens from southern Fiordland have been described previously as a separate subspecies (*X. gilviventris rineyi*) on the basis of morphological (Falla, 1953) and behavioral (Riney, 1953) characteristics, although this subspecies status has never been widely accepted (Gill et al., 2010; Weston and Robertson, 2015). No genetic analysis of the type series specimens has been undertaken, and it is unknown whether the specimens used to describe this subspecies belong to the same genetic clade as contemporary individuals from the same locality.

The NZ rock wren was originally described as *X. gilviventris* in 1867 by August Pelzeln, and later described as *X. haasti* by Sir Walter Buller in 1869. Both taxa were described using specimens collected by Julius Haast in the 1860s (Buller, 1869; Naturhistorisches Museum Wien, unpublished archive), however, the collection localities of these specimens are either unknown (for *X. gilviventris*) or vague (listed as "alpine heights of the South Island" for *X. haasti*; Buller, 1869; see also Tennyson and Bartle, 2008; Gill et al., 2010). Historical evidence suggests these specimens may have been collected from the headwaters of the Rakaia River (Naturhistorisches Museum Wien, unpublished archive). Genetic analysis of these specimens could be used to lend supporting evidence to this theory as this area is suitably north of the hybrid zone that specimens collected from this area are likely to belong to the northern NZ rock wren lineage.

The Natural History Museum at Tring (NHMUK) holds a purported NZ rock wren skin labeled as collected from the "Rimutaka" [= Remutaka] Ranges in the lower North Island prior to the 1930s (NHMUK 1939.12.9.75), outside the current distribution of NZ rock wren. The precise collection date, and collector, are unknown. This specimen was once part of Lord Walter Rothschild's collections, a portion of which were donated to NHMUK upon his death in 1937. Much of Rothschild's collection, including other NZ rock wren specimens, was sold to the American Museum of Natural History (AMNH) in 1932. If the taxonomic identity and collection data for this specimen is correct, then NZ rock wrens once occurred in the North Island and have since been extirpated from the region. As such, a translocation of NZ rock wren to the Remutaka Ranges would constitute a species re-introduction to their former range. Alternatively, NHMUK 1939.12.9.75 may not be a genuine NZ rock wren specimen and could potentially be the morphologically similar NZ bush wren, perhaps the North Island subspecies (*X. l. stokesii*) which possibly persisted within the Wellington region until 1918 (Stidolph, 1926). These two species are thought to have diverged approximately 15 Myrs ago (Mitchell et al., 2016). We would expect to see significant genetic divergence between NHMUK 1939.12.9.75 and the South Island lineages of NZ rock wrens if the taxonomic identity and collection locality of NHMUK 1939.12.9.75 are correct, given the presence of multiple avian taxa endemic to either the North or South Island (e.g., Trewick, 1996; Miller and Lambert, 2006; Murphy et al., 2006; Grosser et al., 2017).

However, even if the taxonomic identity of NHMUK 1939.12.9.75 is correct, its collection locality data may not be accurate. Erroneously-labeled yellow-eyed penguin (*Megadyptes antipodes*) specimens have been identified within the Lord Walter Rothschild collection held by the AMNH (Boessenkool et al., 2009). These specimens were reportedly collected from the sub-Antarctic Auckland and Campbell Islands by Henry Hammersley Travers (H. H. Travers) before being sold to Rothschild. Analysis of genetic microsatellites strongly suggest that these specimens originate from the South Island of New Zealand (Boessenkool et al., 2009). Additionally, H. H. Travers was involved with the mislabeling of South Island snipe (*Coenocorypha iredalei*) skins, including specimens within the AMNH Rothschild collection (Miskelly, 2012). As a result, Travers was probably falsely credited with the discovery of South Island snipe, as it is likely that he never visited the type locality of *C. iredalei* and that the type specimens were sourced from an unknown collector by H. H. Travers, before being sold to Lord Rothschild (Miskelly, 2012). Finally, it has also been suggested that Sir Walter Buller, who sold avian specimens to Rothschild (Bartle and Tennyson, 2009), mislabeled parts of his personal collection upon sale, with the misidentification of two Auckland Island shags (*Leucocarbo colensoi*) attributed to poor record keeping (Rawlence et al., 2014a). NHMUK 1939.12.9.75 could potentially have been supplied to Rothschild by a dealer such as Travers or Buller.

Here we use aDNA extracted from NZ rock wren specimens, and the strong phylogeographic structure present within NZ rock wrens, to infer their taxonomic identity and the accuracy of their recorded collection localities. We include the purported NZ rock wren recorded as collected from the Remutaka Ranges, to determine whether NZ rock wren historically inhabited the North Island of New Zealand. Type specimens of the putative NZ rock wren subspecies *X. gilviventris rineyi* were also examined to establish whether they belong to the same genetic clade as contemporary individuals from the same locality. Syntypes of both *X. gilviventris* and *X. haasti* were analyzed to ascertain the plausibility that they were collected from the headwaters of the Rakaia River. Finally, individuals collected throughout the late 1800s–early 1900s, many of which were associated with incomplete collection data, or obtained by potentially unreliable collectors, were examined to establish the veracity of their associated collection localities.

# METHODS

## Ancient DNA Extraction, Amplification, and Sequencing

A total of 31 NZ rock wren museum skins were sampled from multiple institutions (**Table 1**, **Figure 1**). A clean scalpel blade was used to remove a single toepad from each specimen, with gloves and face-mask worn throughout sampling. A new scalpel blade and gloves were used for each specimen to minimize inter-sample contamination and contamination via exogenous DNA. All DNA extractions and PCR setup were performed in a dedicated, physically isolated, aDNA laboratory (Otago Palaeogenetics Laboratory) following strict aDNA guidelines (see Cooper and Poinar, 2000; Fulton and Shapiro, 2019). No *Xenicus* specimens had been analyzed within this laboratory prior to this study. Ancient DNA extractions followed the methodology of Thomas et al. (2017) whereby toepad samples were minced using a scalpel blade and incubated at 55°C overnight within 1 mL of the extraction buffer of Gilbert et al. (2007); followed by purification and elution of DNA according to Dabney et al. (2013). Negative DNA extraction controls (i.e., reagents only, no sample) were processed alongside museum specimens (∼1 control per 7 specimens), and subjected to the same PCR conditions detailed below (as were negative PCR controls).

The mitochondrial cytochrome b (Cyt B) dataset of Weston and Robertson (2015), and the Cyt B sequences of partial NZ rock and bush wren mitochondrial genomes from Mitchell et al. (2016) (KX369035.1 and KX369033.1), were aligned and used to design PCR primer pairs that would amplify 172 bp of Cyt B, in two small overlapping fragments (overlap = 45 bp including primers, 5 bp sans primers), of NZ bush and rock wrens: Xenicus_cytb_Frag1F (5′-ATCCTAGTCC TCTTCCTCAG-3′) and Xenicus_cytb_Frag1R (5′-CTCCC GATTCATGTGAGGAT-3′) (128 bp including primers), and Xenicus_cytb_Frag2F (5′-TTGAATCCTAATCTCCAACC-3′) and Xenicus_cytb_Frag2R (5′-ATGGGGAATAGGATT

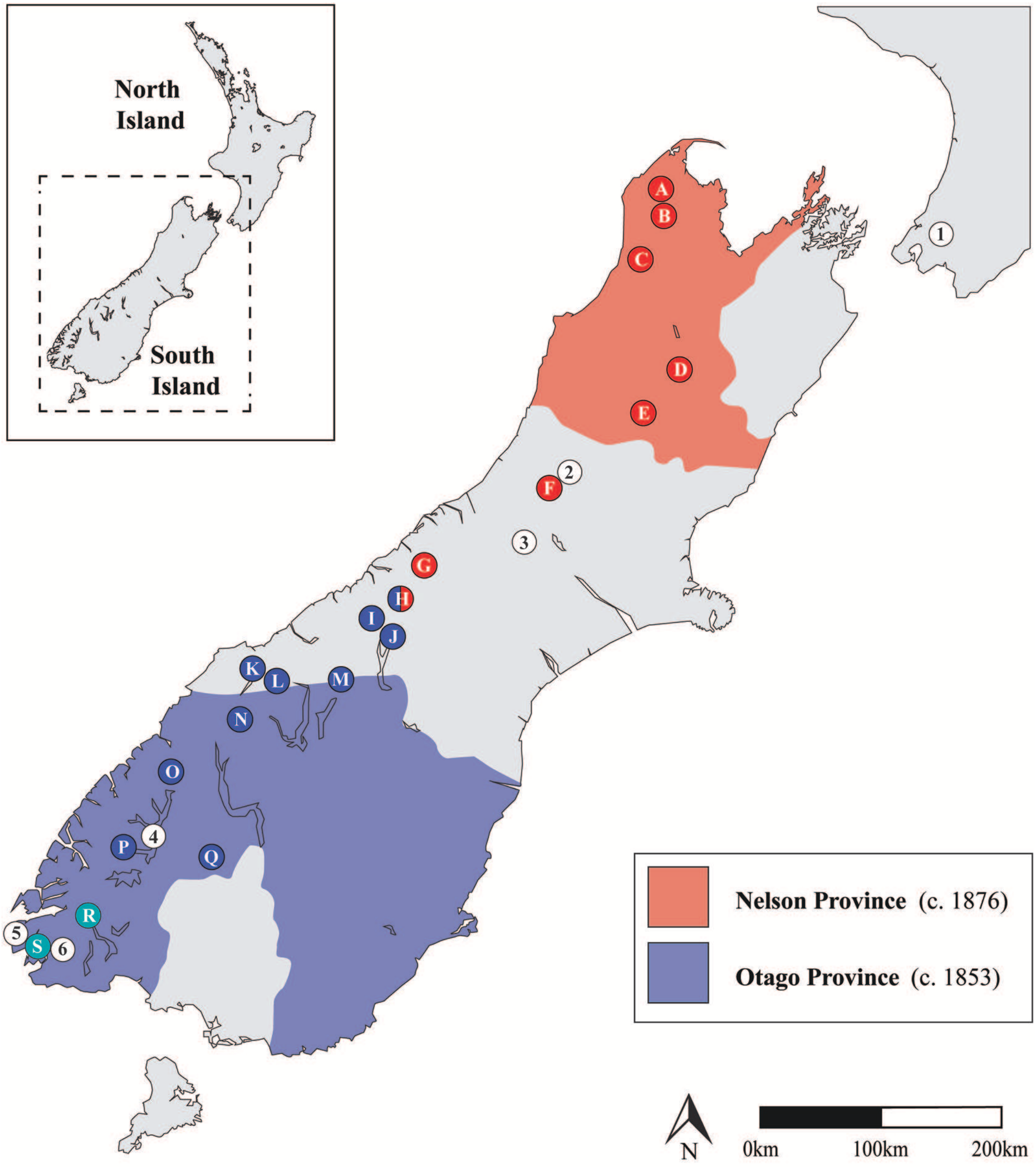
**TABLE 1 |** Registration and collection information of the New Zealand rock wren (*Xenicus gilviventris*) museum specimens examined by this study.

| Museum number | Recorded collection locality | Collection date | Collector | Additional Notes |
|---|---|---|---|---|
| AMNH 554486-88 | Otago Province | April 1897 | H. H. Travers | Collection data for specimens obtained by H. H. Travers may be erroneous |
| AMNH 554489-92 | Nelson Province | March 1897 | H. H. Travers | Collection data for specimens obtained by H. H. Travers may be erroneous |
| AMNH 554493-94 | Nelson Province | October 1896 | H. H. Travers | Collection data for specimens obtained by H. H. Travers may be erroneous |
| AMNH 554495-96 | Nelson Province | September 1896 | H. H. Travers | Collection data for specimens obtained by H. H. Travers may be erroneous |
| AMNH 554498 | Nelson Province | 20th September 1896 | H. H. Travers | Collection data for specimens obtained by H. H. Travers may be erroneous |
| AMNH 554499-500 | Nelson Province | September 1896 | H. H. Travers | Collection data for specimens obtained by H. H. Travers may be erroneous |
| CM Av240-1 | Arthur's Pass | – | E. F. Stead | |
| CM Av888-9 | – | – | – | |
| CM Av890 | Nelson Mountains | – | – | |
| NMNZ OR.2396 | Tops, south-west of Lake MacArthur, Fiordland | 18th April 1953 | T. Riney | *X. gilviventris rineyi* paratype |
| NMNZ OR.2397 | Tops, north of Lake MacArthur, Fiordland | 14th April 1953 | T. Riney | *X. gilviventris rineyi* holotype |
| NMNZ OR.5094 | – | March–April 1866 | J. Haast | *X. haasti* syntype |
| NMNZ OR.5095 | Nelson Province | May 1896 | – | |
| NMNZ OR.5096 | Nelson Province | March 1899 | – | |
| NMNZ OR.5097 | Nelson Province | May 1898 | – | |
| NMNZ OR.12586 | – | March–April 1866 | J. Haast | *X. haasti* syntype |
| NHMUK 1939.12.9.75 | Rimutaka Ranges | – | – | Supposedly collected outside the known range of the NZ rock wren. Part of the Lord Walter Rothschild collection |
| NHMUK 1903.5.13.3 | Lake Te Anau | March 1903 | Donald Ross? | Possibly collected by Donald Ross, presented to NHMUK by the earl of Ranfurly |
| NHMUK 1904.8.2.10 | Long Sound, Fiordland | March 1897 | – | Presented to NHMUK by the earl of Ranfurly, suggested to be *X. gilviventris rineyi* |
| NMW 51045 | – | March-April 1866 | J. Haast | *X. gilviventris* syntype |
| WML 18.10.98.9 | – | – | – | |

*Unknown details are denoted via a "–" AMNH, American Museum of Natural History; CM, Canterbury Museum; NHMUK, Natural History Museum at Tring; NMNZ, National Museum of New Zealand Te Papa Tongarewa; NMW, Naturhistorisches Museum Wein; WML, World Museum Liverpool.*

AGGAG-3′) (129 bp including primers). The NZ rock wren control region (CR) dataset of Weston and Robertson (2015) was used to design primer pairs that would amplify a 280 bp portion of the NZ rock wren mitochondrial control region, in two small overlapping fragments (overlap = 62 bp including primers, 15 bp sans primers): RW_CR_Frag1F (5′-AAATTATGTCCACGCTTGC-3′) and RW_CR_Frag1R (5′-GGTGTATTTTGGTRGATCATTGG-3′) (205 bp including primers), and RW_CR_Frag2F (5′-CTGATTAYTATAACARTCCTACC-3′) with RW_CR_Frag2R (5′-GATGACAATATTTGTCCTGC-3′) (175 bp including primers). Within the 15 bp overlap between the two CR amplicons there are three fixed single nucleotide polymorphisms (SNPs) that distinguish the northern, southern, and Fiordland clades of NZ rock wren.

Each PCR (20 μL) contained 2 μL of undiluted or 1:10 diluted DNA, 0.25 μM of each primer, 0.63 mM dNTPs, 4.0 mM $MgCl_2$, 1 M Betaine, 1 X PCR Buffer II (100 mM Tris-HCl, pH 8.3, 500 mM KCl), and 2 U AmpliTaq Gold (Life Technologies), made up to a total volume of 20 μL with UltraPure double-distilled water (ThermoFisher Scientific). PCR thermocycling conditions were 94°C for 5 min, 60 cycles of 94°C 30 s, 56°C 45 s, 72°C 60 s, and a final extension step of 72°C for 10 min. Unsuccessful PCRs were repeated using lower annealing temperatures (52–54°C). PCR amplification and downstream processes were carried out within a modern genetics laboratory. No modern *Xencius* specimens had been analyzed in this laboratory. Successful PCRs were identified via gel electrophoresis on a 2% agarose gel stained with SYBR safe (Thermo Fisher Scientific), visualized under blue light. Positive PCR products were purified using ExoSAP (1.5 U Exo1, 1 U SAP; GE Healthcare) following the manufacturer's instructions, then sequenced bidirectionally at the University of Otago using Big Dye terminator technology on an ABI 3730xl. When an inconsistency between sequences from an individual was observed (likely due to post-mortem DNA damage, i.e., G–A and C–T transitions; Hofreiter et al., 2001) additional PCRs and bidirectional sequencing were conducted, and a majority-rule consensus applied to the independent replicates (Brotherton et al., 2007; Winters et al., 2011). PCRs and DNA sequencing were



| **Museum Specimen Localities** ○ | | | |
|---|---|---|---|
| 1. Rimutaka Ranges<br>2. Arthur's Pass | 3. Headwaters of the Rakaia River | 4. Lake Te Anau<br>5. Lake McArthur | 6. Long Sound |
| **Sampling Localities of Contemporary Individuals** | | *Northern/Southern/Fiordland Lineages* ● ● ● | |
| A. Lake Clara<br>B. Henderson Basin<br>C. Wangapeka<br>D. Lake Thompson<br>E. Lewis Pass | F. Bealey<br>G. Whymper<br>H. Aoraki/Mount Cook<br>I. Landsborough<br>J. Twin Stream | K. Haast Range<br>L. Lake Crucible<br>M. Ahuriri<br>N. Matukituki<br>O. Homer | P. Murchison Mountains<br>Q. Eyres<br>R. Lake Roe<br>S. Lake MacArthur |

**FIGURE 1 |** Map of New Zealand denoting relevant regions and localities. The historic Otago and Nelson province boundaries are illustrated, collection localities for museum specimens are indicated via numbered circles, while sampling localities for contemporary individuals (see Weston and Robertson, 2015) are indicated via lettered circles.

also replicated for type specimens and specimens that exhibited discrepancies between collection locality and genetic sequence.

## Phylogenetic Analysis

DNA sequences were checked by eye, and aligned against the Cyt B and CR datasets of Weston and Robertson (2015) and Mitchell et al. (2016), within Geneious v11.0.5 (Kearse et al., 2012) using the Geneious alignment algorithm with default parameters. The CR and Cyt B datasets were trimmed to 284 and 163 bp, respectively. The 163 bp alignment of Cyt B contains 14 variable sites, 7 of which are parsimony informative. The Cyt B dataset was primarily used to distinguish between NZ rock and bush wrens to determine the taxonomic identity of the museum specimens analyzed (particularly NHMUK 1939.12.9.75). While the 285 bp CR alignment contains 60 variable sites, with 33 of these being parsimony informative, to provide increased discriminatory power and assign museum specimens to one of the three major NZ rock wren clades previously identified by Weston and Robertson (2015). The Akaike Information Criterion implemented within jModelTest v2.1.10 (Darriba et al., 2012) was used to determine the most appropriate nucleotide substitution model for each dataset. These were HKY + G for the CR dataset, and HKY for the Cyt B dataset. The nested sampling package within BEAST v2.5 (Russel et al., 2018) was used to determine the most appropriate tree prior for Bayesian phylogenetics by converting marginal likelihood values into Bayes factors. The constant population size coalescent tree prior was determined to be the most appropriate prior for both datasets. Bayesian phylogenies were constructed using BEAST2 v2.5 (Bouckaert et al., 2014). The corresponding bush wren cytochrome b sequence (KX369035.1; Mitchell et al., 2016) was used as the outgroup for the Cyt B phylogeny, while the rifleman (*Acanthisitta chloris*) (AY325307.1; Harrison et al., 2004) was used for the CR phylogeny, as the partial bush wren mitogenome of Mitchell et al. (2016) does not contain the control region. The BEAST analysis used 10 million Markov Chain Mote Carlo (MCMC) generations, sampled every 1,000 iterations with the first 10% discarded as the burn-in, a constant population size coalescent tree prior, and the HKY (Cyt B) or HKY + G (CR) nucleotide substitution model. Tracer v1.7 (Rambaut et al., 2018) was used to check effective sample sizes (>200) and MCMC convergence. Bayesian phylogenetic trees were summarized using TreeAnnotator and visualized using FigTree v1.4.3 (https://github.com/rambaut/figtree/releases). Median-joining haplotype networks were constructed using PopART (Leigh and Bryant, 2015).

# RESULTS

Amplifiable DNA was successfully extracted from 30 of the 31 NZ rock wren museum skins analyzed. The two Cyt B and CR fragments were successfully amplified from 30 and 28 specimens, respectively. No amplicons were produced by our negative DNA extraction or PCR controls. The singular specimen (CM Av889) that did not produce amplifiable DNA did not have locality data, and no discrepancies between collection locality and Cyt B sequence were detected for the further two specimens (CM Av888 and NMNZ OR.5097) that failed to produce CR amplicons. Both the CR and Cyt B phylogenies (**Figures 2**, **3**) showed strong support (posterior probability: 0.88–1) for the northern and southern NZ rock wren lineages (and the southern Fiordland clade) previously identified by Weston and Robertson (2015). The NZ rock wren specimen allegedly collected from the southern North Island (NHMUK 1939.12.9.75; Remutaka Ranges) does not cluster with the bush wren Cyt B sequence but sits within the southern lineage of NZ rock wren, specifically, the southern Fiordland clade (**Figure 2**; an uncollapsed version of this tree can be found within the Supplementary Material as **Figure S1**). This phylogenetic position is further supported by the CR data (**Figure 3**). Further discrepancies between collection data and genetic sequence were detected in additional specimens. Three specimens (AMNH 554486, AMNH 554487, and AMNH 554488) collected by Henry H. Travers labeled "Otago Province" in the southern South Island, clustered within the northern NZ rock wren lineage (**Figures 2**, **3**) as opposed to the southern lineage, indicating their associated collection locality data is incorrect. Conversely, a single specimen from "Nelson Province" in the northern South Island (NMNZ OR.5096) consistently grouped with individuals from southern Fiordland (**Figures 2**, **3**). The holotype and a paratype of *X. g. rineyi* (NMNZ OR.2396 and NMNZ OR.2397) also clustered within this southern Fiordland clade, alongside contemporary individuals collected from the same region.

Using the phylogenetic framework of Weston and Robertson (2015), we were able to assign individual specimens with no formal collection locality data (CM Av888, NMNZ OR.12586, NMNZ 5094, NMW 51045, and WML 18.10.98.9) to a broad geographical region based upon their genetic lineage. All specimens with unknown locality data, including a syntype of *X. gilviventris* (NMW 51045), and the syntypes of *X. haasti* (NMNZ OR.12586, NMNZ OR.5094), cluster with the northern lineage (**Figures 2**, **3**), suggesting they were collected from the northern South Island, potentially as far south as Aoraki/Mount Cook. In addition, historical evidence (discussed below) strongly suggests that the type specimens of *X. gilviventris* and *X. haasti* were collected from the headwaters of the Rakaia River, allowing us to formally identify the type locality of both *X. gilviventris* and *X. haasti*.

Median joining haplotype networks (**Figures 2**, **3**) also revealed strong north-south phylogeographic structuring in the CR and Cyt b datasets as previously reported by Weston and Robertson (2015), and the phylogenetic analyses presented here. There were fewer segregating sites within the Cyt B dataset. The northern lineage is dominated by one haplotype, with a singular museum specimen (CM Av890), exhibiting a single base pair difference from this common haplotype (**Figure 2**). All specimens representing the southern Fiordland clade have the same haplotype, while another four haplotypes form a haplogroup representing the remainder of the southern lineage. The CR dataset contained higher levels of genetic diversity with distinct northern and southern haplogroups separated by 12 mutational steps (**Figure 3**).

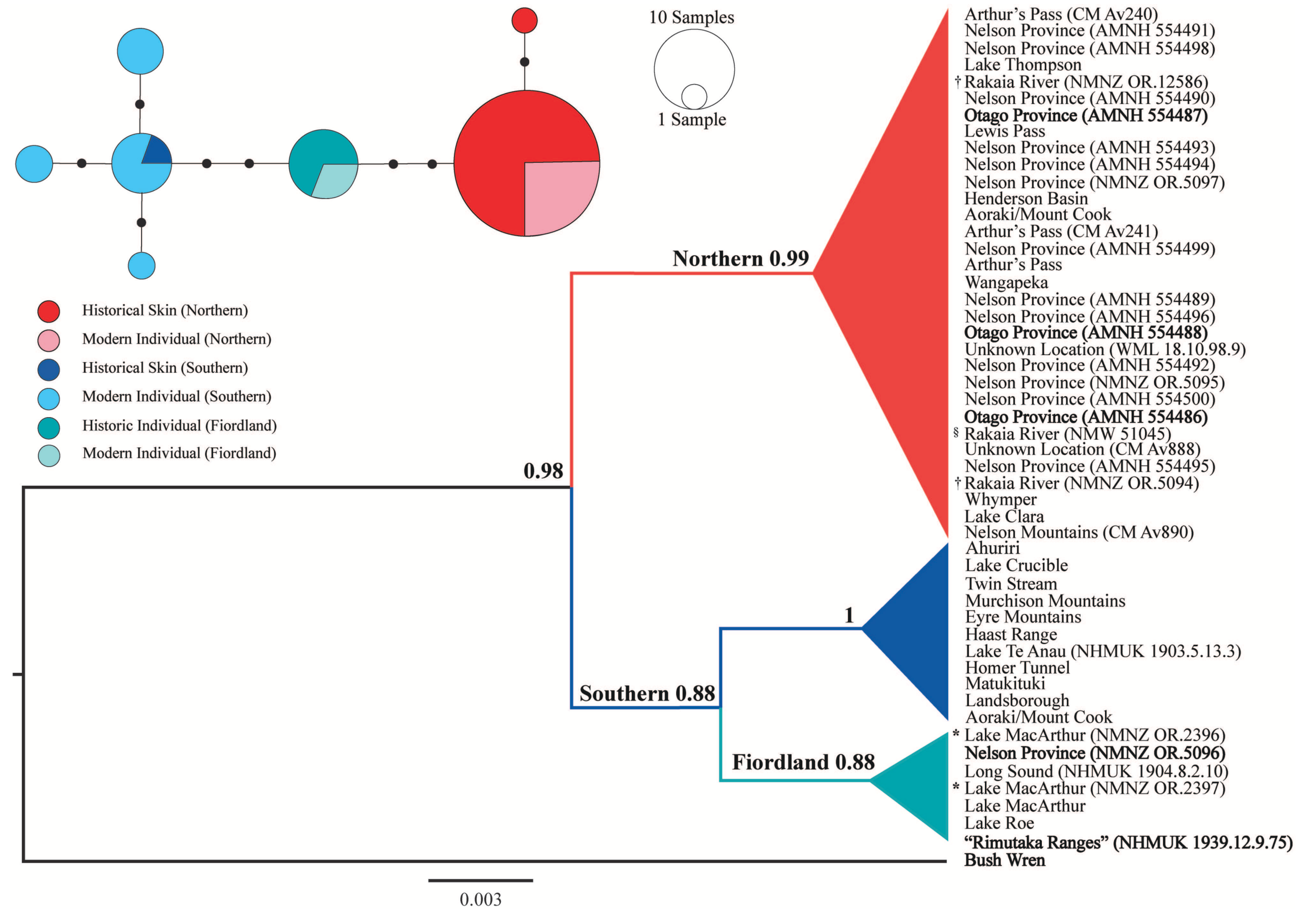

10 Samples
1 Sample
Historical Skin (Northern)
Modern Individual (Northern)
Historical Skin (Southern)
Modern Individual (Southern)
Historic Individual (Fiordland)
Modern Individual (Fiordland)
Northern 0.99
0.98
1
Southern 0.88
Fiordland 0.88
0.003
Arthur's Pass (CM Av240)
Nelson Province (AMNH 554491)
Nelson Province (AMNH 554498)
Lake Thompson
† Rakaia River (NMNZ OR.12586)
Nelson Province (AMNH 554490)
Otago Province (AMNH 554487)
Lewis Pass
Nelson Province (AMNH 554493)
Nelson Province (AMNH 554494)
Nelson Province (NMNZ OR.5097)
Henderson Basin
Aoraki/Mount Cook
Arthur's Pass (CM Av241)
Nelson Province (AMNH 554499)
Arthur's Pass
Wangapeka
Nelson Province (AMNH 554489)
Nelson Province (AMNH 554496)
Otago Province (AMNH 554488)
Unknown Location (WML 18.10.98.9)
Nelson Province (AMNH 554492)
Nelson Province (NMNZ OR.5095)
Nelson Province (AMNH 554500)
Otago Province (AMNH 554486)
§ Rakaia River (NMW 51045)
Unknown Location (CM Av888)
Nelson Province (AMNH 554495)
† Rakaia River (NMNZ OR.5094)
Whymper
Lake Clara
Nelson Mountains (CM Av890)
Ahuriri
Lake Crucible
Twin Stream
Murchison Mountains
Eyre Mountains
Haast Range
Lake Te Anau (NHMUK 1903.5.13.3)
Homer Tunnel
Matukituki
Landsborough
Aoraki/Mount Cook
* Lake MacArthur (NMNZ OR.2396)
Nelson Province (NMNZ OR.5096)
Long Sound (NHMUK 1904.8.2.10)
* Lake MacArthur (NMNZ OR.2397)
Lake MacArthur
Lake Roe
"Rimutaka Ranges" (NHMUK 1939.12.9.75)
Bush Wren

**FIGURE 2 |** Collapsed Bayesian phylogeny and median-joining haplotype network of NZ rock wren (*Xenicus gilviventris*) mitochondrial cytochrome b sequences (163 bp). Major clades mentioned within the text are identified and posterior probabilities are displayed next to main nodes. Bold labels denote museum specimens for which the recorded collection locality and genetic sequence do not match. **X. gilviventris rineyi* type series specimen, †*X. haasti* syntype, §*X. gilviventris* syntype. The scale bar depicts the distance corresponding to 0.003 nucleotide substitutions per site. Within the haplotype network, mutational steps are illustrated via small black circles, and haplotypes are colored according to genetic lineage and specimen type (contemporary individual or museum skin).

## DISCUSSION

The scientific value of museum specimens is dependent on their associated data, including collection locality and taxonomic identity (Shepherd et al., 2013). Inaccurate data can distort inferences of historical population connectivity, identification of conservation management units, and formulation of conservation management plans (Boessenkool et al., 2009). Here we detect five discrepancies between recorded collection locality and genetic sequence by analyzing NZ rock wren specimens held within museum collections worldwide, and combine historical records with our genetic data to describe the type locality of both *X. gilviventris* and *X. haasti*.

A discrepancy with important conservation implications is the mislabeling of a specimen likely collected from southern Fiordland (based on DNA), as "Rimutaka [= Remutaka] Ranges" in the southern North Island. This label data suggested that the past distribution of NZ rock wrens was previously much more extensive than the present-day, and that NZ rock wrens have been historically extirpated from the North Island of New Zealand; however, this position is not supported by genetic evidence and hence the data for this specimen is erroneous. Our results strongly suggest that (1) the taxonomic identity of NHMUK 1939.12.9.75 is indeed a NZ rock wren, not a bush wren, and (2) that the specimen has been mislabeled and was not collected in the southern North Island but was instead collected in the southern South Island. Therefore, there is currently no definitive record of NZ rock wrens inhabiting the North Island of New Zealand.

Additional erroneously-labeled specimens were also discovered, including three specimens from the AMNH Rothschild collection reportedly collected from the Otago Province in April 1897 by H. H. Travers. Mitochondrial DNA sequence analysis suggests that these specimens were collected north of Aoraki/Mount Cook in the central South Island, well outside the historical boundary of the Otago Province. Each specimen has a Rothschild Museum label, as well as a label from H. H. Travers. These labels state "Otago" (i.e., Otago Province)

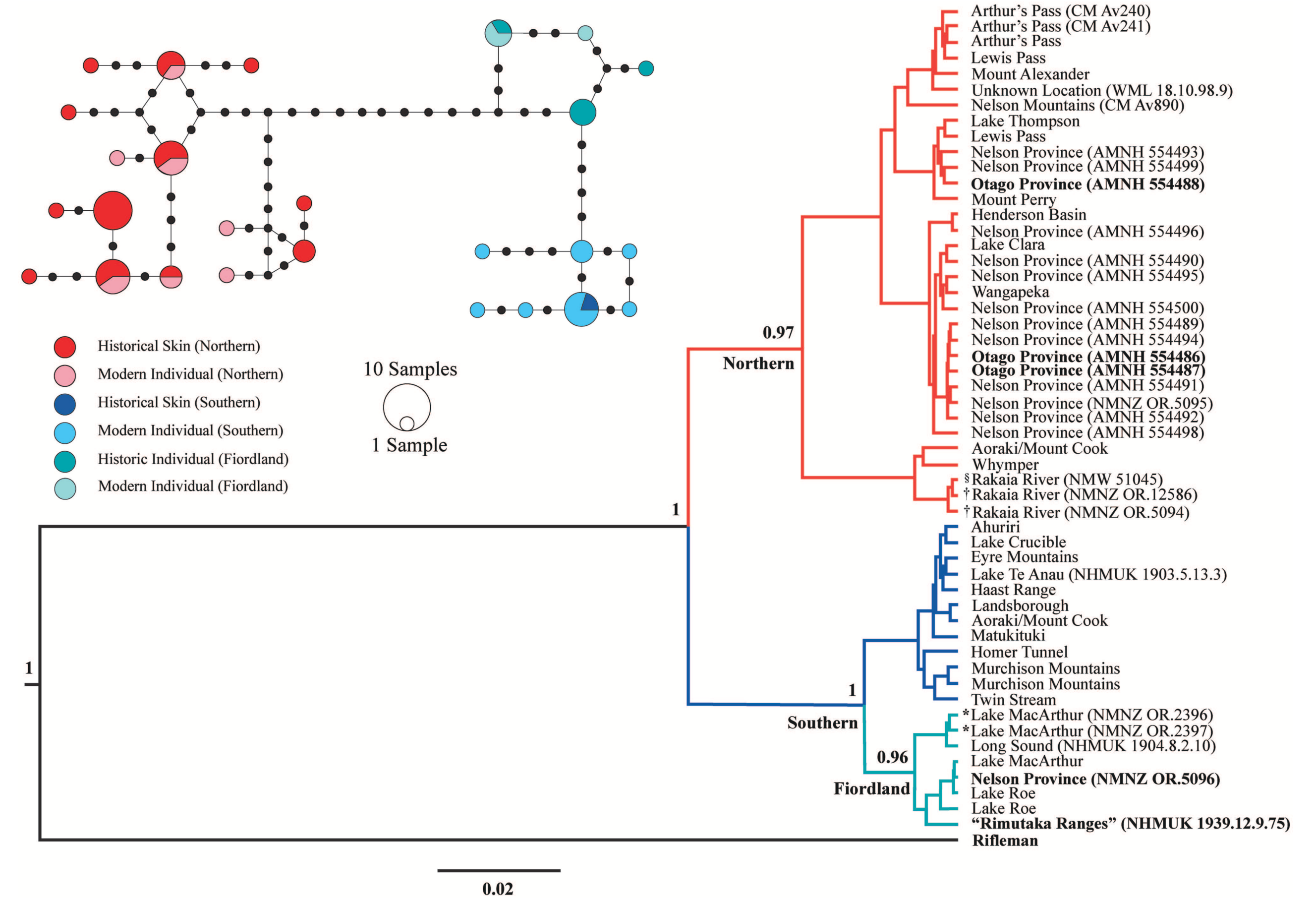

Historical Skin (Northern)
Modern Individual (Northern)
Historical Skin (Southern)
Modern Individual (Southern)
Historic Individual (Fiordland)
Modern Individual (Fiordland)
10 Samples
1 Sample
Arthur's Pass (CM Av240)
Arthur's Pass (CM Av241)
Arthur's Pass
Lewis Pass
Mount Alexander
Unknown Location (WML 18.10.98.9)
Nelson Mountains (CM Av890)
Lake Thompson
Lewis Pass
Nelson Province (AMNH 554493)
Nelson Province (AMNH 554499)
**Otago Province (AMNH 554488)**
Mount Perry
Henderson Basin
Nelson Province (AMNH 554496)
Lake Clara
Nelson Province (AMNH 554490)
Nelson Province (AMNH 554495)
Wangapeka
Nelson Province (AMNH 554500)
Nelson Province (AMNH 554489)
Nelson Province (AMNH 554494)
**Otago Province (AMNH 554486)**
**Otago Province (AMNH 554487)**
Nelson Province (AMNH 554491)
Nelson Province (NMNZ OR.5095)
Nelson Province (AMNH 554492)
Nelson Province (AMNH 554498)
Aoraki/Mount Cook
Whymper
§Rakaia River (NMW 51045)
†Rakaia River (NMNZ OR.12586)
†Rakaia River (NMNZ OR.5094)
Ahuriri
Lake Crucible
Eyre Mountains
Lake Te Anau (NHMUK 1903.5.13.3)
Haast Range
Landsborough
Aoraki/Mount Cook
Matukituki
Homer Tunnel
Murchison Mountains
Murchison Mountains
Twin Stream
*Lake MacArthur (NMNZ OR.2396)
*Lake MacArthur (NMNZ OR.2397)
Long Sound (NHMUK 1904.8.2.10)
Lake MacArthur
**Nelson Province (NMNZ OR.5096)**
Lake Roe
Lake Roe
**"Rimutaka Ranges" (NHMUK 1939.12.9.75)**
**Rifleman**
0.97
Northern
1
1
1
Southern
0.96
Fiordland
0.02

**FIGURE 3 |** Bayesian phylogeny and median-joining haplotype network of NZ rock wren (*Xenicus gilviventris*) mitochondrial control region sequences (284 bp). Major clades mentioned within the text are identified and posterior probabilities are displayed next to main nodes. Bold labels denote museum specimens for which the recorded collection locality and genetic sequence do not match. **X. gilviventris rineyi* type series specimen, †*X. haasti* syntype, §*X. gilviventris* syntype. The scale bar depicts the distance corresponding to 0.02 nucleotide substitutions per site. Within the haplotype network, mutational steps are illustrated via small black circles, and haplotypes are colored according to genetic lineage and specimen type (contemporary individual or museum skin).

and "H. H. Travers," alongside an approximate collection date ("4/97" i.e., April 1897) and the sex of each specimen (Thomas Trombone, personal communication). Whether this mislabeling constitutes deliberate fraud or poor record keeping is unknown. Furthermore, little is known about the discrepant specimen held within the National Museum of New Zealand (NMNZ OR.5096). Collected in 1899, it is likely that H. H. Travers also obtained this individual, as he was one of the primary collectors and sellers of birds to the National Museum of New Zealand during the late 1800s and early 1900s, and is associated with 52 other bird specimens labeled "Nelson" or "Nelson Province" in the National Museum of New Zealand collections (AJDT, unpublished data). If this hypothesis is correct, then a total of four NZ rock wren specimens collected by H. H. Travers have erroneous collection data, in addition to other Travers' bird specimens previously identified (see Boessenkool et al., 2009; Miskelly, 2012). Our results cast further doubt on the label accuracy of specimens obtained from H. H. Travers.

While we acknowledge the considerable debate over the reliability of single gene trees with reference to incomplete lineage sorting (Ballard and Whitlock, 2004; Rubinoff and Holland, 2005; Galla and Johnson, 2015), and changing tree topology due to historically high levels of genetic diversity (Rawlence et al., 2014b, 2015), the strong north-south phylogeographic structure within NZ rock wrens (3.68 ± 0.5% inter-lineage genetic divergence within Cyt B, and 13.3 ± 4.9% inter-lineage genetic divergence within the CR) is well-supported by both mitochondrial and nuclear microsatellite DNA (Weston and Robertson, 2015). While the mitochondrial DNA split between southern Fiordland individuals and the remainder of the southern lineage is shallower, with 1.2% genetic divergence within Cyt B (Weston and Robertson, 2015). Multiple substitutions, consistent between contemporary and historical NZ rock wrens, govern the placement of individuals into one of the three major clades. We believe that this strong phylogeographic structure and consistency across time periods is sufficient to rule out the impacts of incomplete lineage sorting and historical genetic diversity on tree topology.

## Were New Zealand Rock Wrens Ever Present Within the North Island?

Our genetic results demonstrate that there is currently no definitive record of NZ rock wren having inhabited the North Island and suggests that the NZ rock wren was not extirpated from the North Island following the arrival of Europeans. However, it is possible that NZ rock wrens were present within the North Island post-European arrival and were never sighted or collected. Only two specimens of laughing owl (*Ninox albifaces*) were ever collected from the North Island, and have since been lost (Worthy, 1997), while no museum skins of kakapo (*Strigops habroptilus*) are known from the North Island, although there are sightings of the species from there into the early twentieth century (Williams, 1956). Additionally, it is still unknown whether NZ rock wrens were present within the North Island prior to European arrival. Polynesian colonization of New Zealand (∼280 A.D.; Wilmshurst et al., 2008) also marked the arrival of kiore, the pacific rat (*Rattus exulans*), which devastated the invertebrate and small vertebrate faunas of New Zealand (Holdaway, 1989; Towns and Daugherty, 1994; Gibbs, 2009). The current distribution of NZ rock wrens may reflect where populations survived human-mediated extinction pressures. Within the New Zealand archipelago, the North Island is known to have suffered the greatest loss of avifauna (Holdaway et al., 2001; Rawlence et al., 2019). The kea (*Nestor notabilis*), another endemic avian species for which alpine areas constitute an important habitat, was extirpated from the North Island subsequent to the arrival of Polynesians, likely due to predation by humans and kiore (Tennyson et al., 2014). Ancient DNA analysis, or a reappraisal of the morphology and osteology of subfossil *Xenicus* remains is necessary to confirm whether NZ rock wrens occurred within the North Island prior to human arrival.

## Species Translocations

Most conservation translocations of New Zealand species involve the movement of individuals between areas within their historical or prehistoric range, to ensure the persistence of a species Miskelly and Powlesland, 2013. However, some species have been introduced to areas outside their range for a variety of reasons, including the establishment of insurance populations during pest eradication schemes [e.g., Codfish Island fernbird (*Bowdleria punctata wilsoni*) (McClelland, 2002)] or the replacement of extinct relatives (e.g., replacement of the South Island snipe (*C. iredalei*) by the Snares Island snipe (*C. huegeli*) (Miskelly et al., 2012) (reviewed by Miskelly and Powlesland, 2013). According to our genetic results, any translocation of NZ rock wrens to the North Island may constitute a novel introduction of the species to a new area, rather than a species re-introduction or a species restoration. Ongoing management of NZ rock wrens includes translocation to predator-free sites (Willans and Weston, 2005; Weston, 2006), including a successful translocation to Secretary Island in southern Fiordland (Reid and Edge Hill, 2017). Presently, translocation of NZ rock wrens to the North Island does not satisfy the objective of securing a population in a predator-free site to ensure species persistence.

## New Zealand Rock Wren Taxonomy

The northern and southern NZ rock wren (*X. gilviventris* Pelzeln, 1867) lineages are currently considered separate evolutionarily significant units (Robertson et al., 2016). A subspecies, *X. gilviventris rineyi* Falla, 1953, was described from southern Fiordland (see also Riney, 1953), but never widely accepted (e.g., Kinsky, 1970). It is currently treated as a junior synonym (Gill et al., 2010). Falla (1953) compared southern Fiordland NZ rock wren individuals to museum specimens from the northern South Island and considered those from southern Fiordland to be slightly smaller- and to have brighter plumage than their northern counterparts. Riney (1953) suggested that southern Fiordland individuals lacked the bobbing behavior prevalent in other NZ rock wren populations. However, Weston (2014) observed no significant behavioral or morphological differences between contemporary southern Fiordland individuals and other NZ rock wren populations but this was before the exact demarcation of the southern subpopulations was defined using genetics. Further morphological analysis may expose physical differences between genetically distinct populations. Genetic sequences from the holotype and a paratype of *X. gilviventris rineyi* (NMNZ OR.2396 and NMNZ OR.2397), collected in 1953, closely matched (**Figures 2**, **3**) recently sampled individuals from the type locality (Weston and Robertson, 2015).

Genetic analyses of *X. haasti* (NMNZ OR.5094 and NMNZ OR.12586), and *X. gilviventris* (NMW 51045) syntypes determined that all three specimens belong to the northern NZ rock wren lineage (**Figures 2**, **3**), adding support to the hypothesis that these specimens were collected from the headwaters of the Rakaia River. The exact type locality of *X. gilviventris* Pelzeln, 1867 and *Xenicus haasti* Buller, 1869 has until now never been determined. The name *X. gilviventris* Pelzeln, 1867 was based on a single specimen included with 72 birds sent by Julius Haast to Georg Frauenfeld in Vienna (Naturhistorisches Museum Wien, unpublished archive). Frauenfeld was chief zoologist aboard the Frigate *Novara* and Haast had probably met him when he first arrived in New Zealand. August Pelzeln was also aboard the *Novara* and was tasked by Frauenfeld to describe Haast's specimens. In March and early April of 1866, Julius Haast and Frederick Fuller, his taxidermist, visited the headwaters of the Rakaia River. Over a period of 40 days they collected "160 skins of birds … several of them either new to science or at least very rare, and desirable objects for the completion of our own collection" (Haast, 1866). Haast describes the habits of a wren that are presumably the NZ rock wren from the vicinity of Mein's Knob:

> "Another very interesting inhabitant of this district is a large greenish-brown Wren with a drab-colored breast (Certhiparus?), which lives exclusively amongst the large talusses of debris high on the mountain sides. This bird, instead of flying away when frightened or when thrown at with stones, or even when shot at,

> hides itself among the angular debris of which these huge talusses are composed. We tried several times, in vain, by removing some of the blocks and surrounding it to catch one of them alive. It reminded me strongly of the habits and movements of the lizards which live in the same region in similar localities."

Pelzeln's collection of specimens received from Haast contained 39 species all of which were either mentioned in Haast's account of his trip to the Rakaia (Haast, 1866) or are typical of the area. Pelzeln described two new species, *X. gilviventris* Pelzeln, 1867 and a putative taxon of alpine bellbird *Anthornis ruficeps* (Pelzeln, 1867) [= *Anthornis melanura melanura* (Pelzeln, 1867)], from Haast's collections. It appears that Walter Buller also received NZ rock wren specimens from this same collection at approximately the same time, i.e., "collected in the Southern Alps in the 1860s by J.F. Haast" (Bartle and Tennyson, 2009). Buller's collection that contains the types of his *Xenicus haasti* Buller, 1869, also included a syntype of his 'new' parrot *Platycercus alpinus* Buller, 1869 [= *Cyanoramphus malherbi* Souancé, 1857] (Bartle and Tennyson, 2009). Buller "received several specimens" of this parrot from Haast "from the forests of the Southern Alps" (Buller, 1869), so these were probably also part of the same collection.

The type localities and dates of collection for all four taxa can thus be restricted to "Headwaters of the Rakaia River, March or April 1866. Collected by J. Haast and F. Fuller."

## CONCLUDING REMARKS

The genetic data presented within our study illustrates the applied use of aDNA to enhance the utility of museum collections and inform both taxonomy and conservation management; yet highlights the need to determine the veracity of specimen data. Museum specimens should be treated with caution when they are used to formulate conservation management plans that depend upon the accuracy of their associated data. Steps should be taken to discern the accuracy of such details, especially for specimens reportedly obtained by collectors known for poor record keeping or potential data falsification. Preferably, these steps should incorporate a multidisciplinary approach that includes the use of museum records and collectors' notes, alongside biological information (e.g., DNA sequence data). This approach helps to ensure that museum specimens remain a valuable resource for future scientific research. Additionally, future genetic research on museum specimens should also utilize the power of high throughput DNA sequencing, depending upon the nature of the research question(s) addressed. This technology allows for greater detection of post-mortem DNA damage and exogenous contamination, aiding in the authentication of aDNA sequence data. Authentication of aDNA sequencing data continues to form an essential aspect of the aDNA field, and helps ensure that the results of aDNA studies are robust and reliable.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found within GenBank (https://www.ncbi.nlm.nih.gov/genbank/). Accession Numbers: MN402619–MN402676.

## ETHICS STATEMENT

Ethical review and approval was not required for this study because no living animals were collected or examined. DNA samples were taken solely from museum specimens.

## AUTHOR CONTRIBUTIONS

BR, KW, and NR conceived the study. AV and LS performed the laboratory work. AV analyzed the data and wrote the manuscript with contributions from all authors. LS created the figures. AT and RS examined archival material and provided comment on historical avian collectors and NZ rock wren taxonomy. All authors reviewed the manuscript.

## FUNDING

Funding for the presented research was provided to BR by the New Zealand Department of Conservation, and to NR by the Royal Society of New Zealand Marsden Fund (16-UOO-096).

## ACKNOWLEDGMENTS

We are grateful to Hein van Grouw and Alex Bond (Natural History Museum at Tring), Tony Parker (World Museum Liverpool), Hans-Martin Berg (Naturhistorisches Museum Wien), and Thomas Trombone and Paul Sweet (American Museum of Natural History) for access to museum collections and supporting information/comments on the NZ rock wren specimens held by their respective institutions. Thank you to Rosa Wilkie, Andrea van Turnhout, and Nikita-Maree Norris for their assistance with laboratory work as part of the GENE312 course at the University of Otago.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2019.00496/full#supplementary-material

## REFERENCES

Ballard, J. W. O., and Whitlock, M. C. (2004). The incomplete natural history of mitochondria. *Mol. Ecol.* 13, 729–744. doi: 10.1046/j.1365-294X.2003.02063.x

Barbanera, F., Moretti, B., Guerrini, M., Al-Sheikhly, O. F., and Forcina, G. (2016). Investigation of ancient DNA to enhance natural history museum collections: misidentification of smooth-coated otter (*Lutrogale perspicillata*) specimens across multiple museums. *Belg. J. Zool.* 146, 101–112.

Barker, F. K., Cibois, A., Schikler, P., Feinstein, J., and Cracraft, J. (2004). Phylogeny and diversification of the largest avian radiation. *Proc. Natl. Acad. Sci. U.S.A.* 101, 11040–11045. doi: 10.1073/pnas.0401892101

Bartle, J. A., and Tennyson, A. J. D. (2009). History of Walter Buller's collections of New Zealand birds. *Tuhinga Rec. Mus. N. Z. Te Papa Tongarewa* 20, 81–136.

Boessenkool, S., Star, B., Scofield, R. P., Seddon, P. J., and Waters, J. M. (2009). Lost in translation or deliberate falsification? Genetic analyses reveal erroneous museum data for historic penguin specimens. *Proc. R. Soc. B Biol. Sci.* 277, 1057–1064. doi: 10.1098/rspb.2009.1837

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., et al. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 10:e1003537. doi: 10.1371/journal.pcbi.1003537

Brotherton, P., Endicott, P., Sanchez, J. J., Beaumont, M., Barnett, R., Austin, J., et al. (2007). Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Res.* 35, 5717–5728. doi: 10.1093/nar/gkm588

Buller, W. (1869). On some new species of New-Zealand birds. *Ibis* 5, 37–43. doi: 10.1111/j.1474-919X.1869.tb07092.x

Cooper, A., and Poinar, H. N. (2000). Ancient DNA: do it right or not at all. *Science* 289, 1139–1139. doi: 10.1126/science.289.5482.1139b

Cooper, A., Rhymer, J., James, H. F., Olson, S. L., McIntosh, C. E., Sorenson, M. D., et al. (1996). Ancient DNA and island endemics. *Nature* 381:484. doi: 10.1038/381484a0

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., et al. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15758–15763. doi: 10.1073/pnas.1314445110

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9:772. doi: 10.1038/nmeth.2109

Falla, R. A. (1953). Description of a new form of New Zealand wren. *Notornis* 5, 142–143.

Fulton, T. L., and Shapiro, B. (2019). Setting up an ancient DNA laboratory. *Methods Mol Biol.* 1963, 1–13. doi: 10.1007/978-1-4939-9176-1_1

Galla, S. J., and Johnson, J. A. (2015). Differential introgression and effective size of marker type influence phylogenetic inference of a recently divergent avian group (Phasianidae: Tympanuchus). *Mol. Phylogenet. Evol.* 84, 1–13. doi: 10.1016/j.ympev.2014.12.012

Gibbs, G. W. (2009). The end of an 80-million year experiment: a review of evidence describing the impact of introduced rodents on New Zealand's 'mammal-free'invertebrate fauna. *Biol. Invasions* 11, 1587–1593. doi: 10.1007/s10530-008-9408-x

Gilbert, M. T. P., Tomsho, L. P., Rendulic, S., Packard, M., Drautz, D. I., Sher, A., et al. (2007). Whole-genome shotgun sequencing of Mitochondria from ancient hair shafts. *Science* 317, 1927–1930. doi: 10.1126/science.1146971

Gill, B. J., Bell, B. D., Chambers, G. K., Medway, D. G., Palma, R. L., Scofield, R. P., et al. (2010). *Checklist of the birds of New Zealand, Norfolk and Macquarie Islands, and the Ross Dependency, Antarctica.* Wellington: Te Papa Press.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., and Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends Ecol. Evol.* 19, 497–503. doi: 10.1016/j.tree.2004.07.006

Grealy, A., Rawlence, N., and Bunce, M. (2017). Time to spread your wings: a review of the avian ancient DNA field. *Genes* 8:184. doi: 10.3390/genes8070184

Grosser, S., Abdelkrim, J., Wing, J., Robertson, B. C., and Gemmell, N. J. (2017). Strong isolation by distance argues for separate population management of endangered blue duck (*Hymenolaimus malacorhynchos*). *Conserv. Genet.* 18, 327–341. doi: 10.1007/s10592-016-0908-4

Haast, J. (1866). *Report on the Headwaters of the River Rakaia.* Christchurch: Press Co.

Hackett, S. J., Kimball, R. T., Reddy, S., Bowie, R. C. K., Braun, E. L., Braun, M. J., et al. (2008). A phylogenomic study of birds reveals their evolutionary history. *Science* 320, 1763–1768. doi: 10.1126/science.1157704

Harrison, G., McLenachan, P., Phillips, M., Slack, K. E., Cooper, A., and Penny, D. (2004). Four new avian mitochondrial genomes help get to basic evolutionary questions in the late Cretaceous. *Mol. Biol. Evol.* 21, 974–983. doi: 10.1093/molbev/msh065

Hofreiter, M., Jaenicke, V., Serre, D., Haeseler, A. V., and Pääbo, S. (2001). DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* 29, 4793–4799. doi: 10.1093/nar/29.23.4793

Holdaway, R. N. (1989). New Zealand's pre-human avifauna and its vulnerability. *N. Z. J. Ecol.* 12, 11–25.

Holdaway, R. N., Worthy, T. H., and Tennyson, A. J. D. (2001). A working list of breeding bird species of the New Zealand region at first human contact. *NZ. J. Zool.* 28, 119–187. doi: 10.1080/03014223.2001.9518262

IUCN (2013). *Guidelines for Reintroductions and Other Conservation Translocations. Version 1.0*. Gland: IUCN Species Survival Commission.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., et al. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346, 1320–1331. doi: 10.1126/science.1253451

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Kinsky, F. C. (1970). *Annotated Checklist of the Birds of New Zealand.* Wellington: A.H. and A.W. Reed.

Leigh, J. W., and Bryant, D. (2015). popart: full-feature software for haplotype network construction. *Methods Ecol. Evol.* 6, 1110–1116. doi: 10.1111/2041-210X.12410

Leonard, J. A. (2008). Ancient DNA applications for wildlife conservation. *Mol. Ecol.* 17, 4186–4196. doi: 10.1111/j.1365-294X.2008.03891.x

Little, L., King, C. M., and O'Donnell, C. F. J. (2017). Behaviour of stoats (*Mustela erminea*) raiding the nests of rock wrens (*Xenicus gilviventris*) in alpine New Zealand. *Notornis* 64, 124–135.

McClelland, P. J. (2002). "Eradication of Pacific rats (*Rattus exulans*) from Whenua Hou Nature Reserve (Codfish Island), Putauhinu and Rarotoka Islands, New Zealand," in *Turning the Tide: The Eradication of Invasive Species*, eds C. R. Veitch and M. N. Clout (Gland; Cambridge: IUCN SSC Invasive Specialist Group), 173–181.

Michelsen-Heath, S., and Gaze, P. (2007). Changes in abundance and distribution of the rock wren (*Xenicus gilviventris*) in the South Island, New Zealand. *Notornis* 54, 71–78.

Miller, H. C., and Lambert, D. M. (2006). A molecular phylogeny of New Zealand's Petroica (Aves: Petroicidae) species based on mitochondrial DNA sequences. *Mol. Phylogenet. Evol.* 40, 844–855. doi: 10.1016/j.ympev.2006.04.012

Miskelly, C. M. (2012). Discovery and extinction of the South Island Snipe (*Coenocorypha iredalei*) on islands around Stewart Island. *Notornis* 59, 15–31.

Miskelly, C. M., Charteris, M. R., and Fraser, J. R. (2012). Successful translocation of Snares Island snipe (*Coenocorypha huegeli*) to replace the extinct South Island snipe (*C. iredalei*). *Notornis* 59, 32–38.

Miskelly, C. M., and Powlesland, R. G. (2013). Conservation translocations of New Zealand birds, 1863–2012. *Notornis* 60, 3–28.

Mitchell, K. J., Wood, J. R., Llamas, B., McLenachan, P. A., Kardailsky, O., Scofield, R. P., et al. (2016). Ancient mitochondrial genomes clarify the evolutionary history of New Zealand's enigmatic acanthisittid wrens. *Mol. Phylogenet. Evol.* 102, 295–304. doi: 10.1016/j.ympev.2016.05.038

Murphy, S. A., Flux, I. A., and Double, M. C. (2006). Recent evolutionary history of New Zealand's North and South Island Kokako (*Callaeas cinerea*) inferred from mitochondrial DNA sequences. *Emu-Austral Ornithol.* 106, 41–48. doi: 10.1071/MU05007

O'Donnell, C. F. J., Weston, K. A., and Monks, J. M. (2017). Impacts of introduced mammalian predators on New Zealands alpine fauna. *N. Z. J. Ecol.* 41, 1–22. doi: 10.20417/nzjecol.41.18

Olson, S. L., and James, H. F. (1982). Fossil Birds from the Hawaiian Islands: evidence for wholesale extinction by man before western contact. *Science* 217, 633–635. doi: 10.1126/science.217.4560.633

Pelzeln, A. v. (1867). Ueber eine von Herrn Julius Haast erhaltene Sendung von Vogelbälgen aus Neu Seeland. *Verhandlungen der Zoologisch-Botanischen Gesellschaft in Wien* 17, 315–318.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Rawlence, N. J., Kennedy, M., Anderson, C. N. K., Prost, S., Till, C. E., Smith, I. W. G., et al. (2015). Geographically contrasting biodiversity reductions in a widespread New Zealand seabird. *Mol. Ecol.* 24, 4605–4616. doi: 10.1111/mec.13338

Rawlence, N. J., Kennedy, M., Waters, J. M., and Scofield, R. P. (2014a). Morphological and ancient DNA analyses reveal inaccurate labels on two of Buller's bird specimens. *J. R. Soc. N. Z.* 44, 163–169. doi: 10.1080/03036758.2014.972962

Rawlence, N. J., Tennyson, A. J. D., Cole, T. L., Verry, A. J. F., and Scofield, R. P. (2019). Evidence for breeding of Megadyptes penguins in the North Island at the time of human arrival. *NZ. J. Zool.* 46, 165–173. doi: 10.1080/03014223.2018.1523202

Rawlence, N. J., Till, C. E., Scofield, R. P., Tennyson, A. J. D., Collins, C. J., Lalas, C., et al. (2014b). Strong phylogeographic structure in a sedentary seabird, the Stewart Island Shag (*Leucocarbo chalconotus*). *PLoS ONE* 9:e90769. doi: 10.1371/journal.pone.0090769

Reid, R., and Edge Hill, K. A. (2017). *Conserving Fiordland's Biodiversity 1987-2015. The Challenges, the Achievements, the Knowledge*. Department of Conservation.

Riney, T. (1953). Notes on habitat and behaviour of the rock wren subspecies *Xenicus gilviventris* rineyi Falla. *Notornis* 5, 186–188.

Robertson, H. A., Baird, K., Dowding, J. E., Elliott, G. P., Hitchmough, R. A., Miskelly, C. M., et al. (2016). "Conservation status of New Zealand birds, 2016," in *New Zealand Threat Classification Series* (Wellington: Department of Conservation), 1–27.

Rubinoff, D., and Holland, B. S. (2005). Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference. *Syst. Biol.* 54, 952–961. doi: 10.1080/10635150500234674

Russel, P. M., Brewer, B. J., Klaere, S., and Bouckaert, R. R. (2018). Model selection and parameter inference in phylogenetics using Nested Sampling. *Syst. Biol.* 68, 219–233. doi: 10.1093/sysbio/syy050

Shaffer, H. B., Fisher, R. N., and Davidson, C. (1998). The role of natural history collections in documenting species declines. *Trends Ecol. Evol.* 13, 27–30. doi: 10.1016/S0169-5347(97)01177-4

Shepherd, L. D., Tennyson, A. J. D., and Lambert, D. M. (2013). Using ancient DNA to enhance museum collections: a case study of rare kiwi (*Apteryx* spp.) specimens. *J. R. Soc. N. Z.* 43, 119–127. doi: 10.1080/03036758.2012.732585

Stidolph, R. H. D. (1926). Bird-life around Wellington, N.Z. *Emu - Austral Ornithol.* 25, 204–207. doi: 10.1071/MU925204

Tennyson, A. J., and Bartle, J. S. (2008). Catalogue of type specimens of birds in the Museum of New Zealand Te Papa Tongarewa. *Tuhinga Rec. Mus. N. Z. Te Papa Tongarewa* 19, 185–207.

Tennyson, A. J. D., Easton, L. J., and Wood, J. R. (2014). Kea (*Nestor notabilis*)–another North Island human-caused extinction. *Notornis* 61, 174–176.

Thomas, J. E., Carvalho, G. R., Haile, J., Martin, M. D., Castruita, J. A. S., Niemann, J., et al. (2017). An 'Aukward' tale: a genetic approach to discover the whereabouts of the Last Great Auks. *Genes* 8:164. doi: 10.3390/genes8060164

Thomas, W. K., Pääbo, S., Villablanca, F. X., and Wilson, A. C. (1990). Spatial and temporal continuity of kangaroo rat populations shown by sequencing mitochondrial DNA from museum specimens. *J. Mol. Evol.* 31, 101–112. doi: 10.1007/BF02109479

Towns, D. R., and Daugherty, C. H. (1994). Patterns of range contractions and extinctions in the New Zealand herpetofauna following human colonisation. *N. Z. J. Zool.* 21, 325–339. doi: 10.1080/03014223.1994.9518003

Trewick, S. A. (1996). Morphology and evolution of two takahe: flightless rails of New Zealand. *J. Zool.* 238, 221–237. doi: 10.1111/j.1469-7998.1996.tb05391.x

Wandeler, P., Hoeck, P. E. A., and Keller, L. F. (2007). Back to the future: museum specimens in population genetics. *Trends Ecol. Evol.* 22, 634–642. doi: 10.1016/j.tree.2007.08.017

Weston, K. A. (2006). *Post-translocation Monitoring of Rock Wren Xenicus gilviventris on Anchor Island.* Internal Report, Department of Conservation, Te Anau.

Weston, K. A. (2014). *Conservation genetics of alpine rock wren (Xenicus gilviventris)* (Ph.D. Thesis). University of Otago, Dunedin, New Zealand.

Weston, K. A., O'Donnell, C. F. J., van dam-Bates, P., and Monks, J. M. (2018). Control of invasive predators improves breeding success of an endangered alpine passerine. *Ibis* 160, 892–899. doi: 10.1111/ibi.12617

Weston, K. A., and Robertson, B. C. (2015). Population structure within an alpine archipelago: strong signature of past climate change in the New Zealand rock wren (*Xenicus gilviventris*). *Mol. Ecol.* 24, 4778–4794. doi: 10.1111/mec.13349

Willans, M., and Weston, K. A. (2005). *Translocation of Rock Wren Xenicus gilviventris From Murchison Mountains to Anchor Island, Dusky Sound, Fiordland*. Te Anau: Department of Conservation Internal Report.

Williams, G. R. (1956). The Kakapo (Strigops Habrotilus, Gray): a review and reappraisal of a near-extinct species. *Notornis* 7, 29–56.

Wilmshurst, J. M., Anderson, A. J., Higham, T. F. G., and Worthy, T. H. (2008). Dating the late prehistoric dispersal of Polynesians to New Zealand using the commensal Pacific rat. *Proc. Natl. Acad. Sci. U.S.A.* 105, 7676–7680. doi: 10.1073/pnas.0801507105

Winker, K. (2000). Obtaining, preserving, and preparing bird specimens. *J. Field Ornithol.* 71, 250–297. doi: 10.1648/0273-8570-71.2.250

Winters, M., Barta, J. L., Monroe, C., and Kemp, B. M. (2011). To clone or not to clone: method analysis for retrieving consensus sequences in ancient DNA samples. *PLoS ONE* 6:e21247. doi: 10.1371/journal.pone.0021247

Worthy, T. (1997). A survey of historical laughing owl (*Sceloglaux albifacies*) specimens in museum collections. *Notornis* 44, 241–252.

Worthy, T. H., and Holdaway, R. N. (2002). *The Lost World of the Moa: Prehistoric Life of New Zealand*. Bloomington: Indiana University Press.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 31 January 2020
doi: 10.3389/fevo.2020.00011

# Uncovering Signatures of DNA Methylation in Ancient Plant Remains From Patterns of Post-mortem DNA Damage

*Stefanie Wagner[1,2], Christophe Plomion[3] and Ludovic Orlando[1,4]**

[1] CNRS UMR 5288, Faculté de Médecine de Purpan, Laboratoire d'Anthropobiologie et d'Imagerie de Synthèse, Toulouse, France, [2] INRAE – Centre National de Ressources Génomique Végétales, Castanet-Tolosan, France, [3] INRAE, Univ. Bordeaux, BIOGECO, Cestas, France, [4] Globe Institute, University of Copenhagen, Copenhagen, Denmark

The ultra-short DNA molecules still preserved in archeological remains can provide invaluable genetic information about past individuals, species, and communities within the half to 1-million-year time range. The sequence data are, however, generally affected by post-mortem DNA damage and include specific patterns of nucleotide mis-incorporations, which can help data authentication. Recent work in ancient mammals has shown that such patterns can also help assess past levels of DNA methylation in CpG contexts. Despite pioneering work in barley and sorghum, ancient epigenetic marks have received limited attention in plants and it remains unknown whether ancient epigenetic signatures can be retrieved in any of the three main sequence contexts (CG, CHG, and CHH). To address this question, we extended a statistical methylation score originally proposed to trace cytosine methylation in mammal sequence data to accommodate the three methylation contexts common in plants. We applied this score to a range of tissues (wood, cobs, and grains) and species (oak, maize, and barley), spanning both desiccated and waterlogged archeological samples. Ancient sequence data obtained for USER-treated DNA extracts yielded methylation scores on par with DNA methylation levels of modern organellar and nuclear genomes. At the quantitative level, scores were (1) positively correlated to post-mortem cytosine deamination, and (2) replicated relative contributions of CG, CHG, and CHH contexts to DNA methylation assessed by bisulfite DNA sequencing of modern plant tissues. This demonstrates that genuine DNA methylation signatures can be characterized in ancient plant remains, which opens new avenues for investigating the plant evolutionary response to farming, pollution, epidemics, and changing environmental conditions.

**Keywords: ancient DNA, DNA methylation, genomics, post-mortem DNA damage, oak, maize, barley**

OPEN ACCESS

***Edited by:***
*Michael David Martin,*
*Norwegian University of Science and Technology, Norway*

***Reviewed by:***
*Logan Kistler,*
*Smithsonian Institution, United States*
*Sarah Palmer,*
*Aberystwyth University, United Kingdom*

****Correspondence:***
*Ludovic Orlando*
*ludovic.orlando@univ-tlse3.fr*

***Specialty section:***
*This article was submitted to Paleoecology, a section of the journal Frontiers in Ecology and Evolution*

***Received:*** *06 November 2019*
***Accepted:*** *16 January 2020*
***Published:*** *31 January 2020*

***Citation:***
*Wagner S, Plomion C and Orlando L (2020) Uncovering Signatures of DNA Methylation in Ancient Plant Remains From Patterns of Post-mortem DNA Damage. Front. Ecol. Evol. 8:11. doi: 10.3389/fevo.2020.00011*

## INTRODUCTION

Ancient DNA research has witnessed a true revolution over the last decade, shifting from the sequencing of single genomes from ancient anatomically modern humans (Rasmussen et al., 2010) and extinct hominins (Green et al., 2010; Reich et al., 2010) to now multiple thousands of genomes, including from non-human animals and plants (Brunson and Reich, 2019) and their pathogens (Spyrou et al., 2018, 2019). This has provided invaluable insights into the human past (Prüfer et al., 2014; Vernot and Akey, 2014; Narasimhan et al., 2019) and at the same time ancient genome time series have started revealing biological changes underlying plant and animal domestication, uncovering genomic turnovers that were unexpected from patterns of genetic variation present in modern genomes (Fonseca et al., 2015; Gaunitz et al., 2018; Frantz et al., 2019; Verdugo et al., 2019).

In addition to providing snapshots of DNA variation in the past, ancient DNA data can also provide insights into ancient epigenetic marks, leveraging degradation reactions that naturally affect DNA molecules after death (Hanghøj and Orlando, 2018). This could potentially inform on evolutionary changes in expression regulatory networks (Gokhman et al., 2014; Pedersen et al., 2014), age pyramids in ancient population (Hanghøj et al., 2016), as well as social and stress exposure. While accurate ancient chromatin compaction maps can be inferred from differential DNA decay within and outside of nucleosomes (Pedersen et al., 2014; Hanghøj et al., 2016), DNA cytosine methylation represents the epigenetic mark that received most of the scholar attention thus far in ancient DNA research. Early work was based on bisulfite sequencing (Llamas et al., 2012; Smith et al., 2015) but this approach requires amounts of DNA material generally not available in archeological specimens. Alternative experimental methodologies, including enrichment of DNA extracts through methyl binding domains (Smith et al., 2014; Seguin-Orlando et al., 2015), have thus been developed. They, however, also showed limitations owing to the ultra-fragmented and degraded nature of ancient DNA molecules (Seguin-Orlando et al., 2015). Therefore, ancient DNA methylation marks are generally not directly measured, but statistically inferred from patterns of nucleotide mis-incorporations along the genome.

The most powerful approach leverages ancient DNA sequence data retrieved after treatment of the DNA extracts with an enzymatic mix (USER) that excises Uracil nucleotidic bases (U) present in ancient DNA templates, and cleaves the DNA at the resulting abasic site (Briggs et al., 2010). Post-mortem deamination rates are known to be inflated toward the termini of ancient DNA molecules, due to the presence of overhanging ends (Briggs et al., 2007). These degradation reactions affect C residues regardless of their methylation state, transforming those unmethylated C residues into Uracils, and those methylated into Thymines (Pedersen et al., 2014). In the absence of USER treatment, U residues are maintained in the pool of ancient DNA molecules entering into DNA library preparation and sequencing. As a result, C → T mis-incorporations during sequencing will take place at all C residues affected by post-mortem deamination, be methylated (and transformed into T residues) or not (transformed into U residues, and sequenced as T residues). Following USER treatment, however, ancient DNA templates are cleaved at U residues, which almost excludes the incorporation of templates containing U residues into DNA libraries. Here, C → T mis-incorporations are introduced only at those methylated C that were affected by post-mortem deamination (and transformed into T residues). Simple scores, linearly related to the average counts of CpG → TpG mutations, have thus been proposed to infer regional levels of DNA methylation in ancient genome (Gokhman et al., 2014; Pedersen et al., 2014). Therefore, scores calculated from sequence data generated following USER-treatment should be lower than those measured in the absence of USER-treatment. This framework has been recently extended, including statistical models of DNA damage, to account for sequencing errors and sequence variation within the sequenced individual (Hanghøj et al., 2019).

Statistical inference of ancient DNA methylation has so far been only applied to mammals (mainly hominins, horses, and aurochsen; Hanghøj et al., 2016) and no studies have explored whether the same principles could help to obtain DNA methylation profiles in ancient plant remains, despite a growing wealth of sequence data in ancient crops, such as maize (Fonseca et al., 2015; Ramos-Madrigal et al., 2016; Vallebueno-Estrada et al., 2016; Swarts et al., 2017; Kistler et al., 2018) and barley (Mascher et al., 2016), and other cultivated (e.g., sorghum, Smith et al., 2019; grape, Ramos-Madrigal et al., 2019) or uncultivated (e.g., oak, Wagner et al., 2018) plants. This is likely due to the complex methylation and demethylation machinery present in higher plants, which in contrast to mammals, affects three main different sequence contexts: CG (or CpG), CHG, and CHH [where H refers to Adenine (A), C, or T], with relative methylation rates generally following the sequence CG > CHG > CHH (Feng et al., 2010; Niederhuth et al., 2016; Takuno et al., 2016; Bartels et al., 2018).

In this study, we extended the methylation score originally proposed by Pedersen et al. (2014) to accommodate the three sequence contexts underlying DNA methylation in plants and assessed, for the first time, whether DNA methylation information can be statistically inferred from ancient DNA data retrieved from a range of tissues (seeds, cobs, and wood) and species (barley, maize, and oak). The retrieval of signatures both qualitatively and quantitatively in line with expectations from plant DNA methylation profiles in modern plant tissues supports the validity of the approach implemented. This opens new avenues for future ancient DNA research in plants combining both genome and epigenetic time series to retrace the evolutionary response of plants to farming, pollution, epidemics, and changing environmental conditions, including global warming and increasing droughts.

## MATERIALS AND METHODS

### Preliminary Note

In the absence of DNA methylation maps for ancient plants, we reasoned that the validity of putative ancient DNA methylation signatures could be gauged through their ability to replicate genome-wide patterns characterized in modern plants. This follows the rationale first applied to mammals and contrasting C deamination rates within and outside CpGs, (almost) unmethylated mitochondrial DNA vs. nuclear DNA, hypomethylated CpG islands vs. hypermethylated CG-rich promoters, and more (Pedersen et al., 2014; Hanghøj et al., 2016). More specifically, we extended the Ms statistics of normalized C → T mutation counts to other contexts than CG (i.e., CHG and CHH) (**Figure 1**) and tested two predictions. First, we

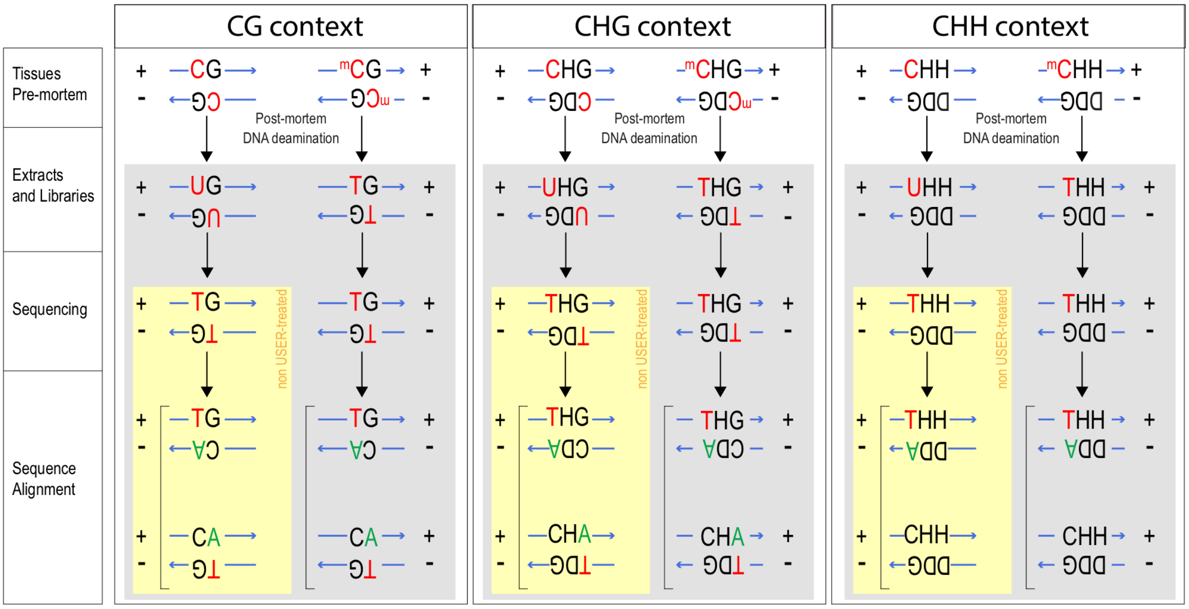


**FIGURE 1 |** Tracking methylation in CG, CHG, and CHH contexts. After death, unmethylated Cytosines (C) become deaminated into Uracil (U), whereas methylated Cytosines ($^m$C) become Thymines (T). In the absence of USER treatment, Us enter into library preparation and sequencing. C → T mis-incorporation will be observed for deaminated residues from methylated (C → T) and unmethylated Cytosine residues (C → U, sequenced as T; yellow box). If USER treatment is applied, U is excised prior to sequencing and C → T conversions can be used as a proxy for those Cs that were methylated pre-mortem. Base mis-incorporations highlighted in red and green. Ms ratios were calculated for raw extracts and USER treated extracts for the three plant methylation contexts CG, CHG, and CHH (where H refers to A, C, or T and D to T, G, or A).

tested whether overall Ms levels calculated from nuclear DNA were significantly greater than those estimated from chloroplast DNA, a locus known for an almost complete lack of DNA methylation (Cokus et al., 2008; Feng et al., 2010). Second, we tested the expected sequence of higher DNA methylation rates in CG contexts relative to CHG and CHH contexts (Feng et al., 2010; Takuno et al., 2016; Bartels et al., 2018). Both predictions are expected to be valid and specific to sequence data generated following USER treatment of raw ancient DNA extracts, should Ms scores reflect ancient DNA methylation signatures. The putative DNA methylation signatures retrieved should also be greater at read ends where post-mortem C deamination is faster than in the central region of sequence reads, owing to the overhanging nature of ancient DNA molecules (Briggs et al., 2007).

## Calculation of Ms Ratios as Methylation Scores in CG, CHG, and CHH Contexts

Post-mortem C deamination rates were calculated within CG contexts following Pedersen et al. (2014). The calculation was extended to CHG contexts, by scanning CAG, CCG, and CTG individually and to CHH contexts by scanning for CAA, CAC, CAT, CCA, CCC, CCT, CTA, CTC, and CTT. Overall, we therefore summed C → T mis-incorporation rates observed at the first position of a total of 12 sequence contexts. For reads aligned against the negative strand of the reference genome, we focused instead on G → A mutations affecting the last position of those contexts. This is so as C deamination on the negative strand replaces the methylated C residue into a T residue, which is then sequenced and read as an A residue when aligned to the positive strand (**Figure 1**). Let note + sequences aligned to the positive strand, and – those aligned against the negative strand, keeping orientation so as to match BAM alignments (i.e., sequence mapping against the negative strand are displayed in the same orientation as the positive strand). A total count of those ancient C residues that were methylated and present in a CG context is thus provided by the cumulated sum of CG+ → TG+ and CG− → CA− mis-incorporations, where CG+ and CG− refer to the reference genome sequence at a given position, and TG+ and CA− to that sequenced in the ancient specimen. Methylation rates are estimated by normalizing this numerator denominator of all counts where the sequence data show and lack evidence of DNA deamination, i.e. CG+ → CG+, CG+ → TG+, GC− → GT−, and GC− → GC−, thus providing so-called Ms ratios (Pedersen et al., 2014). Now extending to CAG contexts, the numerator can easily be written as the sum of CAG+ → TAG+ and CAG− → CAA− counts, while the denominator becomes the sum of CAG+ → CAG+, CAG+ → TAG+, CAG− → CAA−, and CAG− → CAG− counts. **Figure 1** provides a breakdown of the different counts considered in all three main sequence contexts. Read alignments including indels were disregarded to avoid potential inflation of substitution counts on their flanking nucleotides in case of mis-alignment.

Note that a number of processes other than just DNA methylation, such as sequencing errors and the presence of sequence polymorphism in the individual sequenced, contribute to the counts considered. Similarly, in the presence of high post-mortem C deamination rates, the conversion of methylated C residues into T residues is expected to be inflated. Therefore, the Ms ratio calculated do not scale across individuals and can only be used to compare putative methylation levels within (rather than between) individuals.

### Samples, Sequencing Data, and Mapping

Previously published Illumina sequence data were retrieved from public data repositories for 12 oak, 6 maize, and 2 barley ancient samples representing a range of preservation conditions, DNA deamination rates, species, and tissues (**Table 1** and **Supplementary Table S1**; Mascher et al., 2016; Swarts et al., 2017; Wagner et al., 2018). Raw sequence data were processed using PALEOMIX (Schubert et al., 2014), which automatically handles adapter trimming, alignment, duplicate removal, and filtering of low-quality read alignments. We used the PALEOMIX parameters specified in Wagner et al. (2018), and the haploid version of the *Quercus robur* nuclear (750 Mb) and chloroplast[1] reference genomes for oak data (Plomion et al., 2016, 2018; Leroy et al., 2017). For maize and barley, the softmasked version of the respective reference genomes (including nuclear and organelle DNA) were used for read alignments. Those references are available on the Ensembl Plants Database[2] (maize: B73 reference genome v4; barley: IBSC reference genome v2) (Schnable et al., 2009; International Barley Genome Sequencing Consortium et al., 2012; Kersey et al., 2018) and represent a nuclear genome size of 2.4 and 5.3 Gb, respectively. Mapping statistics are provided in **Supplementary Tables S2**, **S3**. For sequence data generated in the absence of USER treatment, ancient DNA damage patterns and quantification were carried out through PALEOMIX, using the methodology implemented in MapDamage (Jónsson et al., 2013).

## RESULTS

We calculated Ms ratios for the 20 ancient specimens where paired sequence data were available for both USER-treated and raw DNA extracts (**Figure 2**). We confirmed that the Ms ratios obtained from USER-treated data of all three species examined were consistently lower than those obtained in the absence of USER treatment. This hold true for both chloroplast DNA and nuclear DNA, although it was most pronounced for the former (1.2–21-fold vs. 1.1–4-fold reduction). This is in line with the (almost) absence of C methylation along the chloroplast DNA, which considerably limits the amount of deaminated methylated C residues available for sequencing. The same was true for the mitochondrial genome, which could be calculated for maize and barley, which also showed values of Ms ratios post-USER treatment lower to those calculated on nuclear DNA data. This demonstrates that Ms ratios as calculated in this study can recapitulate the known methylation difference between the chloroplast and nuclear genomes.

We next calculated Ms ratios within the main three sequence methylation contexts present in plants (CG, CHG, and CHH)

[1] https://w3.pierroton.inra.fr/QuercusPortal/index.php?p=GENOMIC_SEQWe

[2] https://plants.ensembl.org

**TABLE 1 |** Summary of samples analyzed in this study.

| Sample ID | Taxon | Excavation site | State | Age (cal. years BP) | Tissue | Preservation | References |
|---|---|---|---|---|---|---|---|
| 233 | *Quercus* sp. | Sutz Lattrigen | Switzerland | 4700 | Wood | Waterlogged | Wagner et al., 2018 |
| 235 | *Quercus* sp. | Sutz Lattrigen | Switzerland | 4700 | Wood | Waterlogged | Wagner et al., 2018 |
| 321 | *Quercus* sp. | Spica | Slovenia | 4500 | Wood | Waterlogged | Wagner et al., 2018 |
| 354 | *Quercus* sp. | Strojanove vode | Slovenia | 5600 | Wood | Waterlogged | Wagner et al., 2018 |
| 382 | *Quercus* sp. | Hamburg Harburg | Germany | 550 | Wood | Waterlogged | Wagner et al., 2018 |
| 384 | *Quercus* sp. | Hamburg Harburg | Germany | 550 | Wood | Waterlogged | Wagner et al., 2018 |
| 423 | *Quercus* sp. | Greifswald | Germany | 750 | Wood | Waterlogged | Wagner et al., 2018 |
| 44 | *Quercus* sp. | Erstein | France | 3700 | Wood | Waterlogged | Wagner et al., 2018 |
| 442 | *Quercus* sp. | Eschenz Orkopf | Switzerland | 3700 | Wood | Waterlogged | Wagner et al., 2018 |
| 447 | *Quercus* sp. | Eschenz Orkopf | Switzerland | 5200 | Wood | Waterlogged | Wagner et al., 2018 |
| 46 | *Quercus* sp. | Erstein | France | 3700 | Wood | Waterlogged | Wagner et al., 2018 |
| 466 | *Quercus* sp. | Champeaux | France | 1400 | Wood | Waterlogged | Wagner et al., 2018 |
| JK3009E1 | *Hordeum vulgare* subsp. *vulgare* | Joram Cave | Israel | 3939–3775 | Grain | Desiccated | Mascher et al., 2016 |
| JK3010E1 | *Hordeum vulgare* subsp. *vulgare* | Joram Cave | Israel | 3940–3773 | Grain | Desiccated | Mascher et al., 2016 |
| JK1691 | *Zea mays* | Turkey Pen Shelter | Utah | 1785 | Cob | Desiccated | Swarts et al., 2017 |
| JK1696 | *Zea mays* | Turkey Pen Shelter | Utah | 1910 ± 20 | Cob | Desiccated | Swarts et al., 2017 |
| JK1697 | *Zea mays* | Turkey Pen Shelter | Utah | 1900 ± 20 | Cob | Desiccated | Swarts et al., 2017 |
| JK1698 | *Zea mays* | Turkey Pen Shelter | Utah | 1869 ± 20 | Cob | Desiccated | Swarts et al., 2017 |
| JK1699 | *Zea mays* | Turkey Pen Shelter | Utah | 2026 ± 20 | Cob | Desiccated | Swarts et al., 2017 |
| JK1700 | *Zea mays* | Turkey Pen Shelter | Utah | 1903 ± 20 | Cob | Desiccated | Swarts et al., 2017 |

*For all samples, we used paired sequence data generated based upon USER-treated and raw extracts. See* ***Supplementary Table S1*** *for accession numbers and details.*

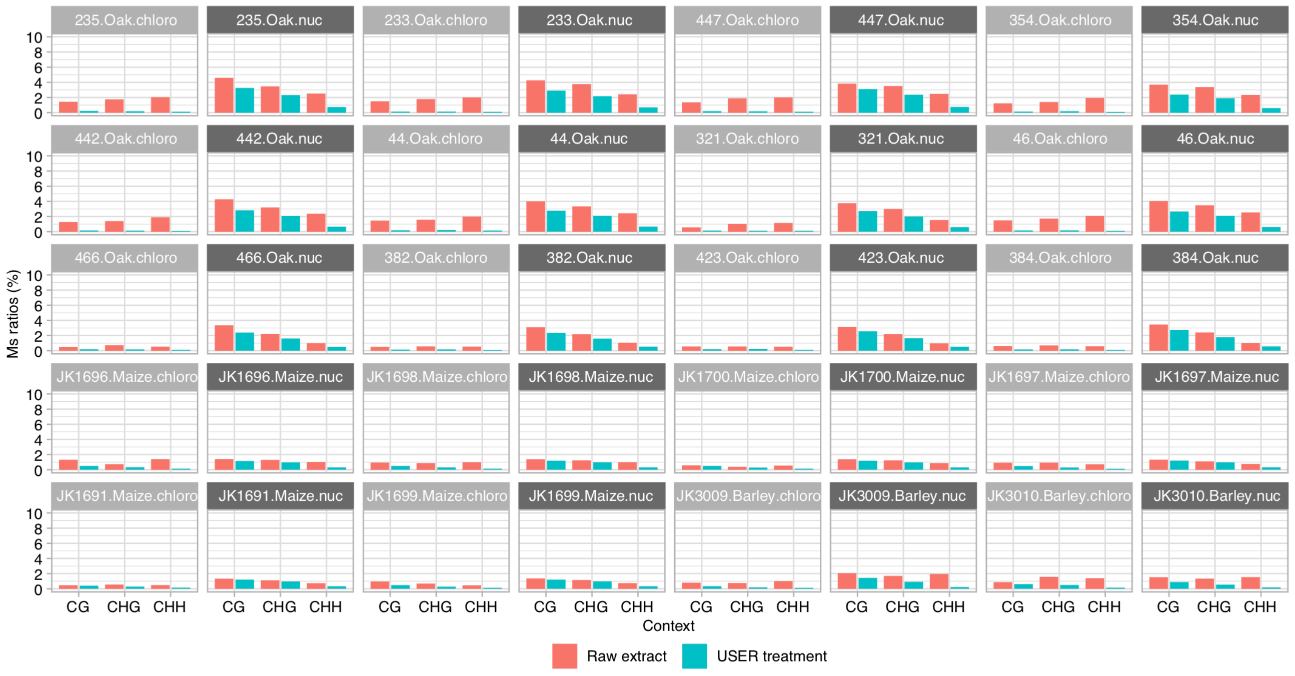


**FIGURE 2 |** Ms ratios for CG, CHG, and CHH contexts. Results are shown for paired sequence data of USER-treated and raw extracts of ancient oak, maize, and barley samples, for chloroplast and nuclear DNA, respectively. Scores are reported for full read length. For Ms ratios computed for outer and inner read regions, see **Supplementary Figure S1**.

in order to test whether we could recapitulate the generally predominant contribution of CG methylation over both other contexts and the generally minimal contribution of CHH contexts (Feng et al., 2010; Takuno et al., 2016; Bartels et al., 2018). In modern oak, bisulfite sequencing data obtained from buds of European oak trees (*Quercus petraea* and *Q. robur*) showed that CG contexts contribute to a fraction of 54.4% of DNA methylation marks in the nuclear genome, while CHG contexts contribute to 39.5% and CHH to 6.1% (unpublished, data kindly provided by Stéphane Maury, ANR EPITREE[3]). In all 12 ancient oak wood samples investigated here, we found that the sequence context providing the highest Ms ratios calculated on USER-treated data (thus contributing the most to DNA methylation) was consistently the CG context, while CHG and CHH provided the second most and least important values, respectively (**Figures 2**, **3**). Normalizing individual Ms ratios by the sum of all three Ms ratios in each sample calculated provided relative estimates of the respective contribution of each sequence context to DNA methylation (**Figure 3**). We found Ms ratios supporting 48.6–54.1% methylation within CG contexts, and 35.0–38.9 and 10.9–12.5% within CHG and CHH contexts, respectively, in line with bisulfite DNA sequencing measures on modern oak buds. This suggests that the calculation of Ms ratios proposed in this study can retrieve genuine, quantitative estimates of DNA methylation from ancient plant DNA sequence data. For maize, previous work on unfertilized ears showed that CG contexts contributed to 52.0% of the DNA methylation marks present in the nuclear genome, while CHG and CHH contexts represented a fraction of 44.7 and 3.3%, respectively (Gent et al., 2013). The estimates retrieved within CG (47.3–48.0%) and CHG (38.6–39.2%) contexts from ancient maize cobs were slightly inferior, due to an increased contribution of CHH contexts (12.7–13.5%, **Figure 3**). This may pertain to marginal differences in the DNA methylation profiles of the tissues compared. Similarly, the estimates retrieved for the relative contribution of CHH contexts to nuclear DNA methylation marks was slightly higher (8.9 and 12.1%) than that directly measured through bisulfite sequencing of barley (1.1%) on mixed samples of seedling, leaf, and apex tissues (unpublished, Benjamin Stich, personal communication) (54.0 and 55.7 vs. 58.2%; and 33.8 and 35.4 vs. 40.7%).

The capacity of Ms ratios to track ancient C methylation marks does not only result from the removal of any possible contribution of unmethylated C residues to sequencing mis-incorporations. It also comes from the deamination reactions that transform methylated C residues into Thymines after death. Therefore, Ms ratios are expected to be maximized in situations where post-mortem C deamination is maximized, for example at read ends. We confirmed this expectation by using sequence data obtained following USER-treatment and contrasting Ms ratios within the first and last 10 read positions on the one hand, and on the remaining positions on the other hand (**Supplementary Figure S1**). Additionally, we found that post-mortem C deamination rates within overhanging ends, as estimated using mapDamage2

[3] https://www6.inra.fr/epitree-project/Le-projet-EPITREE

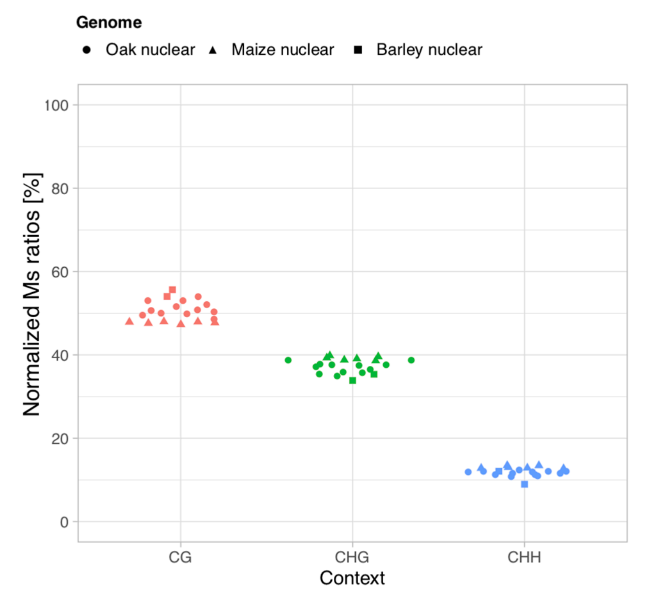


**FIGURE 3 |** Relative contributions of CG, CHG, and CHH contexts to DNA methylation. Normalized Ms ratios are shown for nuclear DNA analyzed from sequencing data generated based upon USER-treated extracts of 12 ancient oak, 6 maize, and 2 barley samples.

(Jónsson et al., 2013), were positively correlated with Ms ratios. This demonstrates that the Ms ratios calculated in this study provide a measure of the intensity of C deamination reactions.

Altogether, Ms ratios both respond to increased C deamination post-mortem reactions and recapitulate known methylation features between organellar and nuclear DNA as well as between different methylation sequence contexts (CG, CHG, and CHH). This was thus far only described for CG contexts in animals. The extension of Ms ratio calculation to tri-nucleotide context proposed here for the first time, thus provide genuine proxies for inferring plant DNA methylation levels in the past. The ancient sequence data for the three species investigated here is not sufficient to generate precise genome-wide DNA methylation maps and identify differentially methylated regions in ancient and modern samples. The data we used allowed to explore the relative contributions of each individual sequence context to CHG (i.e., CAG, CCG, and CTG) and CHH (i.e., CAA, CAC, CAG, CCA, CCC, CCG, CTA, CTC, and CTG) methylation (**Figure 4**). We considered that Ms ratios obtained within CG contexts on USER-treated data represented an arbitrary but convenient score of 100%. We then normalized the Ms ratios of each other sequence context examined relative to this value to assess their respective contributions to DNA methylation. We found that all three sequence motifs underlying CHG contexts contributed evenly to DNA methylation (**Figure 4**). The same was true for the nine sequence motifs underlying CHH contexts, suggesting that the cellular enzymatic machinery responsible for DNA methylation shows no strong preference for any particular sequence motif.

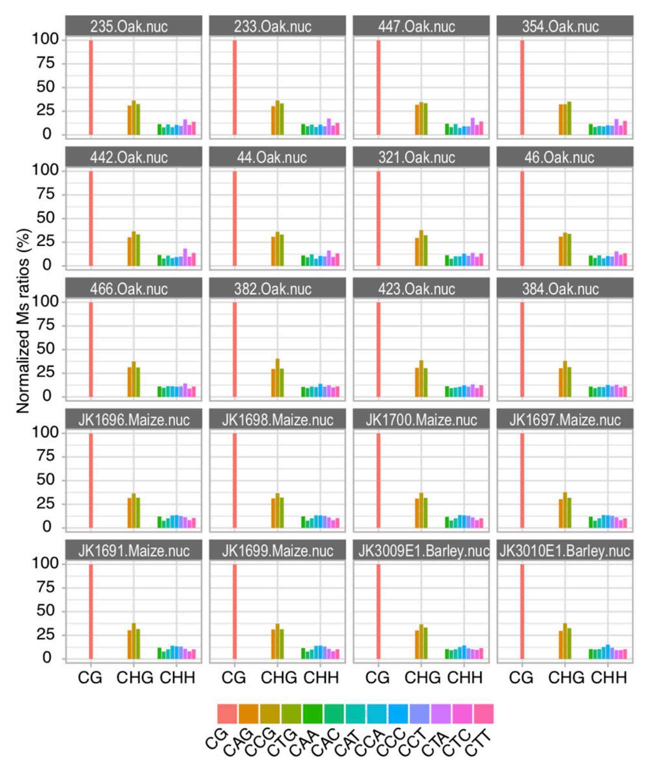


**FIGURE 4 |** Relative contribution of individual tri-nucleotide sequences to methylation in CHG and CHH contexts. Normalized Ms ratios scored for full read length are shown for nuclear DNA analyzed from sequencing data generated based upon USER-treated extracts.

## DISCUSSION

A variety of ancient DNA studies have demonstrated that epigenetic marks can be revealed in the fossilized tissues from ancient individuals, either through experimental assays directly targeting methylated C residues or indirect statistical analyses leveraging patterns of sequencing errors predominantly affecting these sites [see Hanghøj and Orlando (2018) for a review]. To the best of our knowledge, only two of previous studies were applied to ancient plant DNA material (Smith et al., 2014, 2019). In the first, the authors made use of proteins showing strong affinity to methylated CG to enrich ancient barley DNA extracts for methylated fragments. They also applied standard bisulfite sequencing to quantify genome-wide average of DNA methylation levels, a measure that can reflect the exposure of host cells to viruses (e.g., Boyko et al., 2007). In the second study, the authors have applied a software package originally designed to track methylation at CpG sites in ancient mammal genomes (Hanghøj et al., 2016) to archeological sorghum DNA samples from Egypt spanning a full time series (1,800–100 years BP) (Smith et al., 2019). No information was gathered outside the CG context despite other sequence contexts that contribute to DNA methylation in plants known to be characteristic and highly dynamic in response to intrinsic and extrinsic stress factors (Feng et al., 2010; Takuno et al., 2016; Bartels et al., 2018).

In this study, we demonstrate for the first time that the methodology originally proposed to retrieve DNA methylation information from ancient mammal tissues within CpG contexts (Pedersen et al., 2014) can be extended to tri-nucleotide contexts to retrieve DNA methylation signatures within CG, CHG, and CHH contexts. Several lines of evidence support the validity of the approach proposed here. At the qualitative level, Ms ratios calculated on sequence data generated following USER-treatment of raw ancient DNA extracts replicate the known levels of DNA methylation of organellar and nuclear genomes. At the quantitative level, such Ms ratios are (1) positively correlated to levels of post-mortem C deamination rates, (2) inferior to those estimated on the sequence data produced in the absence of USER-treatment, and (3) replicate the relative contributions of CG, CHG, and CHH contexts to DNA methylation. The latter suggests similar kinetics of cytosine DNA deamination after death in all three contexts. However, direct experimental evidence remains necessary to assess whether methylated C residues show similar deamination when present on CG, CHG, and CHH sequence contexts. Finding support for differential decay would impact the method proposed here, as it would exhibit non-uniform sensitivity for inferring DNA methylation levels across the genome. Direct comparisons directly contrasting specific genomic regions (e.g., promoters), without controlling for similar base compositional effects (e.g., similar fractions of CG, CHG, and/or CGG contexts), would then be flawed.

Ancient DNA studies focusing on plant remains have made significant contributions to the fields of anthropology and biology (Palmer et al., 2012; Allaby et al., 2015; Brown et al., 2015; Estrada et al., 2018; Brunson and Reich, 2019; Pont et al., 2019), with important projects reconstructing the genomic history of maize, grape, and barley domestication (Fonseca et al., 2015; Mascher et al., 2016; Ramos-Madrigal et al., 2016, 2019; Vallebueno-Estrada et al., 2016; Swarts et al., 2017; Kistler et al., 2018). The findings reported here suggest that the ever-increasing amounts of sequencing data underlying such projects will soon open for an evaluation of the epigenetic impact of domestication. By extension, when other taxa and other archeological and environmental contexts will be included, the production of larger ancient sequence datasets will help researchers characterize the epigenetic response of plant species to many other selective pressures, including climate change, pollution, and more (Allaby et al., 2015; Gutaker and Burbano, 2017). Unless candidate loci of interest or predefined sets of CG, CHG, and CHH loci are specifically targeted (e.g., through in solution DNA capture, Ramos-Madrigal et al., 2019), reaching this goal will, however, require massive sequencing efforts. It is crucial to keep in mind that (1) ancient DNA analyses are destructive, (2) archeological material represent a finite resource, and that (3) C deamination inflate sequencing errors and can, thus, impact downstream analyses. In our opinion, focusing sequencing efforts on DNA libraries prepared from ancient DNA extracts treated with the USER enzyme mix will mitigate these issues and will reduce analytical costs, as both genomic and epigenetic information could be revealed from the same sequence data. As we observed that the correlation between the Ms ratios and expected methylation levels at different markers was lost in the absence of USER treatment (**Figure 2**), we expect that the power to retrieve both types of information from the sequence data obtained on DNA libraries prepared from raw DNA extracts will be limited.

Previous work has shown that signatures that could inform about gene expression, such as nucleosome positioning along genes and their phasing across cells, could be obtained from the sequencing data underlying ancient mammal genomes (Hanghøj et al., 2016). Whether such signatures could also be collected in plants remains unknown and will require further research. Future work should also focus on improving the methodology proposed here, following what recently done for mammals where statistical models, such as the one implemented in the DamMet package (Hanghøj et al., 2019), have been developed to account for the contribution of other factors than DNA methylation to C → T mis-incorporations. This should reduce the impact of sequencing errors and non-reference variants present in the individual sequenced, and thus improve the accuracy of DNA methylation inference. Extending DamMet to tri-nucleotide contexts will, however, be challenging as the number of underlying genotypes could become rapidly intractable and machine learning approaches may prove more powerful to disentangle nucleotide mis-incorporations pertaining to post-mortem deamination reactions affecting methylated C residues from those resulting from all other processes contributing to sequencing errors. Once the methodology will have been refined, and following the amount of sequence data that were found necessary for retrieving accurate DNA methylation estimates in mammals, we anticipate that ancient genomes characterized at a minimal sequencing depth of 20–25× may be used for in depth studies on ancient plant methylomes. Interestingly, such datasets have already started to be produced in plants (e.g., Mascher et al., 2016: Sample JK3014, average read depth 20×).

In this study, we extended a statistical procedure to uncover ancient DNA methylation signatures in the three sequence contexts present in plants. We applied our procedure on previously published ancient DNA data generated using NGS for oak waterlogged oak wood (*Quercus* sp.), desiccated maize cobs (*Zea mays*), and desiccated barley grains (*Hordeum vulgare* subsp. *vulgare*) and obtained signatures in line with modern plant methylomes. This opens new avenues for designing evolutionary studies on the rich plant material preserved in herbaria and in the archeological record.

## DATA AVAILABILITY STATEMENT

All the sequencing data analyzed in this study have been previously published and are available for download from the European Nucleotide Archive under Accession Nos. PRJEB21188 (https://www.ebi.ac.uk/ena/data/search?query=PRJEB21188), PRJEB20529 (https://www.ebi.ac.uk/ena/data/search?query=PRJEB20529), and PRJEB12197 (https://www.ebi.ac.uk/ena/data/search?query=+PRJEB12197).

## AUTHOR CONTRIBUTIONS

LO conceived and designed the experiments, analyzed, and interpreted the data. LO wrote the manuscript with inputs from SW and CP. SW prepared the figures.

## FUNDING

This work was supported by the Initiative d'Excellence Chaires d'Attractivité, Université de Toulouse (OURASI), and the Villum Fonden miGENEPI research project. CP has received funding from the ANR (epitree project, ANR-17-CE32-0009-01, 2018–21). LO has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program (Grant Agreement No. 681605-PEGASUS).

## ACKNOWLEDGMENTS

We thank Antoine Kremer and the Advanced Grant Project TREEPEACE (No. FP7-339728) for supporting this work and Stéphane Maury and Benjamin Stich for sharing unpublished modern plant methylome data.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.00011/full#supplementary-material

## REFERENCES

Allaby, R. G., Gutaker, R., Clarke, A. C., Pearson, N., Ware, R., Palmer, S. A., et al. (2015). Using archaeogenomic and computational approaches to unravel the history of local adaptation in crops. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370:20130377. doi: 10.1098/rstb.2013.0377

Bartels, A., Han, Q., Nair, P., Stacey, L., Gaynier, H., and Mosley, M. (2018). Dynamic DNA methylation in plant growth and development. *Int. J. Mol. Sci* 19:2144. doi: 10.3390/ijms19072144

Boyko, A., Kathiria, P., Zemp, F. J., Yao, Y., Pogribny, I., and Kovalchuk, I. (2007). Transgenerational changes in the genome stability and methylation in pathogen-infected plants: virus-induced plant genome instability. *Nucleic Acids Res.* 35, 1714–1725. doi: 10.1093/nar/gkm029

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., et al. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U.S.A.* 104, 14616–14621. doi: 10.1073/pnas.0704665104

Briggs, A. W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., and Pääbo, S. (2010). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* 38:e87. doi: 10.1093/nar/gkp1163

Brown, T., Cappellini, E., Kistler, L., Lister, D., Oliveira, H., Wales, N., et al. (2015). Recent advances in ancient DNA research and their implications for archaeobotany. *Veg. Hist. Archaeobot.* 24, 207–214. doi: 10.1007/s00334-014-0489-4

Brunson, K., and Reich, D. (2019). The promise of paleogenomics beyond our own species. *Trends Genet.* 35, 319–329. doi: 10.1016/j.tig.2019.02.006

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., et al. (2008). Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* 452, 215–219. doi: 10.1038/nature06745

Estrada, O., Breen, J., Richards, S. M., and Cooper, A. (2018). Ancient plant DNA in the genomic era. *Nat. Plants* 4, 394–396. doi: 10.1038/s41477-018-0187-9

Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G., et al. (2010). Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl. Acad. Sci. U.S.A.* 107, 8689–8694. doi: 10.1073/pnas.1002720107

Fonseca, R. R., Smith, B. D., Wales, N., Cappellini, E., Skoglund, P., and Fumagalli, M. (2015). The origin and evolution of maize in the Southwestern United States. *Nat. Plants* 1:14003. doi: 10.1038/nplants.2014.3

Frantz, L. A. F., Haile, J., Lin, A. T., Scheu, A., Geörg, C., Benecke, N., et al. (2019). Ancient pigs reveal a near-complete genomic turnover following their introduction to Europe. *Proc. Natl. Acad. Sci. U.S.A.* 116, 17231–17238. doi: 10.1073/pnas.1901169116

Gaunitz, C., Fages, A., Hanghøj, K., Albrechtsen, A., Khan, N., Schubert, M., et al. (2018). Ancient genomes revisit the ancestry of domestic and Przewalski's horses. *Science* 360, 111–114. doi: 10.1126/science.aao3297

Gent, J. I., Ellis, N. A., Guo, L., Harkess, A. E., Yao, Y., Zhang, X., et al. (2013). CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Res.* 23, 628–637. doi: 10.1101/gr.146985.112

Gokhman, D., Lavi, E., Prüfer, K., Fraga, M. F., Riancho, J. A., Kelso, J., et al. (2014). Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. *Science* 344, 523–527. doi: 10.1126/science.1250368

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., et al. (2010). A draft sequence of the Neandertal genome. *Science* 328, 710–722. doi: 10.1126/science.1188021

Gutaker, R. M., and Burbano, H. A. (2017). Reinforcing plant evolutionary genomics using ancient DNA. *Curr. Opin. Plant Biol.* 36, 38–45. doi: 10.1016/j.pbi.2017.01.002

Hanghøj, K., and Orlando, L. (2018). "Ancient epigenomics," in *Paleogenomics: Genome-Scale Analysis of Ancient DNA*, eds C. Lindqvist, and O. P. Rajora, (Cham: Springer International Publishing), 75–111. doi: 10.1007/13836_2018_18

Hanghøj, K., Renaud, G., Albrechtsen, A., and Orlando, L. (2019). DamMet: ancient methylome mapping accounting for errors, true variants, and post-mortem DNA damage. *Gigascience* 8:giz025. doi: 10.1093/gigascience/giz025

Hanghøj, K., Seguin-Orlando, A., Schubert, M., Madsen, T., Pedersen, J. S., Willerslev, E., et al. (2016). Fast, accurate and automatic ancient nucleosome and methylation maps with epiPALEOMIX. *Mol. Biol. Evol.* 33, 3284–3298. doi: 10.1093/molbev/msw184

International Barley Genome Sequencing Consortium, Mayer, K. F. X., Waugh, R., Brown, J. W. S., Schulman, A., and Langridge, P. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature* 491, 711–716. doi: 10.1038/nature11543

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684. doi: 10.1093/bioinformatics/btt193

Kersey, P. J., Allen, J. E., Allot, A., Barba, M., Boddu, S., Bolt, B. J., et al. (2018). Ensembl genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.* 46, D802–D808. doi: 10.1093/nar/gkx1011

Kistler, L., Maezumi, S. Y., Gregorio de Souza, J., Przelomska, N. A. S., Malaquias Costa, F., Smith, O., et al. (2018). Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. *Science* 362, 1309–1313. doi: 10.1126/science.aav0207

Leroy, T., Roux, C., Villate, L., Bodénès, C., Romiguier, J., Paiva, J. A. P., et al. (2017). Extensive recent secondary contacts between four European white oak species. *New Phytol.* 214, 865–878. doi: 10.1111/nph.14413

Llamas, B., Holland, M. L., Chen, K., Cropley, J. E., Cooper, A., and Suter, C. M. (2012). High-resolution analysis of cytosine methylation in ancient DNA. *PLoS One* 7:e30226. doi: 10.1371/journal.pone.0030226

Mascher, M., Schuenemann, V. J., Davidovich, U., Marom, N., Himmelbach, A., Hübner, S., et al. (2016). Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley. *Nat. Genet.* 48, 1089–1093. doi: 10.1038/ng.3611

Narasimhan, V. M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., and Mallick, S. (2019). The formation of human populations in South and Central Asia. *Science* 365:eaat7487. doi: 10.1126/science.aat7487

Niederhuth, C. E., Bewick, A. J., Ji, L., Alabady, M. S., Kim, K. D., Li, Q., et al. (2016). Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* 17:194.

Palmer, S. A., Smith, O., and Allaby, R. G. (2012). The blossoming of plant archaeogenetics. *Ann. Anat.* 194, 146–156. doi: 10.1016/j.aanat.2011.03.012

Pedersen, J. S., Valen, E., Velazquez, A. M. V., Parker, B. J., Rasmussen, M., Lindgreen, S., et al. (2014). Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Res.* 24, 454–466. doi: 10.1101/gr.163592.113

Plomion, C., Aury, J.-M., Amselem, J., Alaeitabar, T., Barbe, V., Belser, C., et al. (2016). Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. *Mol. Ecol. Resour.* 16, 254–265. doi: 10.1111/1755-0998.12425

Plomion, C., Aury, J.-M., Amselem, J., Leroy, T., Murat, F., Duplessis, S., et al. (2018). Oak genome reveals facets of long lifespan. *Nat. Plants* 4, 440–452. doi: 10.1038/s41477-018-0172-3

Pont, C., Wagner, S., Kremer, A., Orlando, L., Plomion, C., and Salse, J. (2019). Paleogenomics: reconstruction of plant evolutionary trajectories from modern and ancient DNA. *Genome Biol.* 20:29. doi: 10.1186/s13059-019-1627-1

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505, 43–49. doi: 10.1038/nature12886

Ramos-Madrigal, J., Runge, A. K. W., Bouby, L., Lacombe, T., Samaniego Castruita, J. A., Adam-Blondon, A.-F., et al. (2019). Palaeogenomic insights into the origins of French grapevine diversity. *Nat. Plants* 5, 595–603. doi: 10.1038/s41477-019-0437-5

Ramos-Madrigal, J., Smith, B. D., Moreno-Mayar, J. V., Gopalakrishnan, S., Ross-Ibarra, J., Gilbert, M. T. P., et al. (2016). Genome sequence of a 5,310-year-old maize cob provides insights into the early stages of maize domestication. *Curr. Biol.* 26, 3195–3201. doi: 10.1016/j.cub.2016.09.036

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., et al. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* 463, 757–762. doi: 10.1038/nature08835

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., et al. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468, 1053–1060. doi: 10.1038/nature09710

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115. doi: 10.1126/science.1178534

Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* 9, 1056–1082. doi: 10.1038/nprot.2014.063

Seguin-Orlando, A., Gamba, C., Sarkissian, C. D., Ermini, L., Louvel, G., Boulygina, E., et al. (2015). Pros and cons of methylation-based enrichment methods for ancient DNA. *Sci. Rep.* 5:11826. doi: 10.1038/srep11826

Smith, O., Clapham, A. J., Rose, P., Liu, Y., Wang, J., and Allaby, R. G. (2014). Genomic methylation patterns in archaeological barley show de-methylation as a time-dependent diagenetic process. *Sci. Rep.* 4:5559. doi: 10.1038/srep05559

Smith, O., Nicholson, W. V., Kistler, L., Mace, E., Clapham, A., Rose, P., et al. (2019). A domestication history of dynamic adaptation and genomic deterioration in *Sorghum*. *Nat Plants* 5, 369–379. doi: 10.1038/s41477-019-0397-9

Smith, R. W. A., Monroe, C., and Bolnick, D. A. (2015). Detection of cytosine methylation in ancient DNA from five native american populations using bisulfite sequencing. *PLoS One* 10:e0125344. doi: 10.1371/journal.pone.0125344

Spyrou, M. A., Keller, M., Tukhbatova, R. I., Scheib, C. L., Nelson, E. A., Andrades Valtueña, A., et al. (2019). Phylogeography of the second plague pandemic revealed through analysis of historical *Yersinia pestis* genomes. *Nat. Commun.* 10:4470. doi: 10.1038/s41467-019-12154-0

Spyrou, M. A., Tukhbatova, R. I., Wang, C.-C., Valtueña, A. A., Lankapalli, A. K., Kondrashin, V. V., et al. (2018). Analysis of 3800-year-old *Yersinia pestis* genomes suggests Bronze Age origin for bubonic plague. *Nat. Commun.* 9:2234. doi: 10.1038/s41467-018-04550-9

Swarts, K., Gutaker, R. M., Benz, B., Blake, M., Bukowski, R., Holland, J., et al. (2017). Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. *Science* 357, 512–515. doi: 10.1126/science.aam9425

Takuno, S., Ran, J.-H., and Gaut, B. S. (2016). Evolutionary patterns of genic DNA methylation vary across land plants. *Nat. Plants* 2:15222. doi: 10.1038/nplants.2015.222

Vallebueno-Estrada, M., Rodríguez-Arévalo, I., Rougon-Cardoso, A., Martínez González, J., García Cook, A., Montiel, R., et al. (2016). The earliest maize from San Marcos Tehuacán is a partial domesticate with genomic evidence of inbreeding. *Proc. Natl. Acad. Sci. U.S.A.* 113, 14151–14156. doi: 10.1073/pnas.1609701113

Verdugo, M. P., Mullin, V. E., Scheu, A., Mattiangeli, V., Daly, K. G., Maisano Delser, P., et al. (2019). Ancient cattle genomics, origins, and rapid turnover in the Fertile Crescent. *Science* 365, 173–176. doi: 10.1126/science.aav1002

Vernot, B., and Akey, J. M. (2014). Resurrecting surviving Neandertal lineages from modern human genomes. *Science* 343, 1017–1021. doi: 10.1126/science.1245938

Wagner, S., Lagane, F., Seguin-Orlando, A., Schubert, M., Leroy, T., Guichoux, E., et al. (2018). High-throughput DNA sequencing of ancient wood. *Mol. Ecol.* 27, 1138–1154. doi: 10.1111/mec.14514

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

ORIGINAL RESEARCH
published: 03 March 2020
doi: 10.3389/fevo.2020.00037

# Unveiling the Ecological Applications of Ancient DNA From Mollusk Shells

**Clio Der Sarkissian[1]*, Per Möller[2], Courtney A. Hofman[3,4], Peter Ilsøe[5], Torben C. Rick[6], Tom Schiøtte[7], Martin Vinther Sørensen[7], Love Dalén[8,9] and Ludovic Orlando[1,5]**

*[1] Laboratoire d'Anthropologie et d'Imagerie de Synthèse, UMR 5288, CNRS, Université Paul Sabatier, Toulouse, France, [2] Department of Geology, Quaternary Sciences, Lund University, Lund, Sweden, [3] Laboratories for Molecular Anthropology and Microbiome Research, Stephenson Research and Technology Center, Norman, OK, United States, [4] Department of Anthropology, The University of Oklahoma, Norman, OK, United States, [5] The GLOBE Institute, University of Copenhagen, Copenhagen, Denmark, [6] Department of Anthropology, National Museum of Natural History, Smithsonian Institution, Washington, DC, United States, [7] Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark, [8] Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden, [9] Centre for Palaeogenetics, Stockholm, Sweden*

The shells of marine mollusks represent promising metagenomic archives of the past, adding to bones, teeth, hairs, and environmental samples most commonly examined in ancient DNA research. Seminal work has established that DNA recovery from marine mollusk shells depends on their microstructure, preservation and disease state, and that authentic ancient DNA could be retrieved from specimens as old as 7,000 years. Here, we significantly push the temporal limit for shell DNA recovery to ≥100,000 years with the successful genetic characterization of one *Portlandia arctica* and one *Mytilus* mussel sample collected within a dated permafrost layer from the Taimyr Peninsula, Russia. We expand the analysis of ancient DNA in carbonate shells to a larger number of genera (*Arctica*, *Cernuella*, *Crassostrea*, *Dreissena*, *Haliotis*, *Lymnaea*, *Margaritifera*, *Pecten*, *Ruditapes*, *Venerupis*) from marine, freshwater and terrestrial environments. We demonstrate that DNA from ancient shells can provide sufficient resolution for taxonomic, phylogenetic and/or population assignment. Our results confirm mollusk shells as long-term DNA reservoirs, opening new avenues for the investigation of environmental changes, commercial species management, biological invasion, and extinction. This is especially timely in light of modern threats to biodiversity and ecosystems.

**Keywords: ancient DNA, mollusk shell, high-throughput DNA sequencing, taxonomic assignment, climate change, invasion, extinction**

OPEN ACCESS

***Edited by:***
*Nic Rawlence,
University of Otago, New Zealand*

***Reviewed by:***
*Kieren James Mitchell,
The University of Adelaide, Australia
Catherine Collins,
University of Otago, New Zealand*

****Correspondence:***
*Clio Der Sarkissian
clio.dersarkissian@univ-tlse3.fr*

***Specialty section:***
*This article was submitted to Paleoecology, a section of the journal Frontiers in Ecology and Evolution*

***Received:*** *03 December 2019*
***Accepted:*** *07 February 2020*
***Published:*** *03 March 2020*

***Citation:***
*Der Sarkissian C, Möller P, Hofman CA, Ilsøe P, Rick TC, Schiøtte T, Sørensen MV, Dalén L and Orlando L (2020) Unveiling the Ecological Applications of Ancient DNA From Mollusk Shells. Front. Ecol. Evol. 8:37. doi: 10.3389/fevo.2020.00037*

## INTRODUCTION

Applications of ancient DNA (aDNA) to ecological studies are plentiful, especially now that high-throughput DNA sequencing (HTS) technologies make it possible to generate genome-wide data at the scale of populations for the last million years (Orlando et al., 2013; Der Sarkissian et al., 2015; Leonardi et al., 2016; Nielsen et al., 2017; Fages et al., 2019; Narasimhan et al., 2019). By allowing high-resolution genomic history reconstructions, aDNA has the power to reveal how changes in the environment (e.g., climate, pollution levels), or interactions with other species and populations (e.g., hybridization, admixture, invasions), have potentially affected past organisms or populations through adaptation, distribution shifts, or extinction (e.g., Green et al., 2010; Lorenzen et al., 2011; Orlando and Cooper, 2014; Palkopoulou et al., 2015). Although most ancient DNA studies have focused on mammal bones and teeth (Green and Speller, 2017), new knowledge in ecology and evolution could be gained by applying ancient genomics to shell-producing mollusks (Coutellec, 2017). These animals indeed grow their carbonate shells in an incremental fashion, simultaneously recording biological and environmental data (Fortunato, 2015; Steinhardt et al., 2016; Butler et al., 2019), which can be tracked through a multitude of proxies, including morphology, isotopes (e.g., $\partial^{18}O_{shell}$, $\partial^{13}C_{shell}$), trace elements and metal composition (e.g., Ba, Cd, Cu, Mo, Pb, U, Zn). The rich fossil record of mollusk shells, thus, provides an equally rich environmental archive of seawater paleo-temperature and salinity (Surge et al., 2003; Chauvaud et al., 2005; Hiebenthal et al., 2012; Vokhshoori and McCarthy, 2014; Reynolds et al., 2016; Black et al., 2017), pollution levels (Vander Putten et al., 2000; Liehr et al., 2005; Pérez-Mayol et al., 2014), stress (Hiebenthal et al., 2012; Trivellini et al., 2018), infection history (Paillard et al., 2004; Trinkler et al., 2010), as well as food availability and primary productivity (Lartaud et al., 2010; Sadler et al., 2012).

Despite recent advances in ancient genomics and the ecological relevance of mollusks, no mollusk genomic time transect has ever been produced. The potential of applying HTS to ancient marine mollusk shells has, however, been assessed for the first time in a recent study that revealed their metagenomic content (Der Sarkissian et al., 2017). This study identified species with aragonitic shell microstructures, such as *Mytilus* mussels, *Haliotis* abalones, *Arctica islandica* quahogs, and *Venerupis*/*Ruditapes* clams, as good candidates for mollusk DNA content and/or preservation, with the longest DNA survival observed for a *Mytilus* specimen dated to ∼7,000 years Before Present (year BP; Der Sarkissian et al., 2017). This timeframe allows for the examination of many important questions relative to ecological changes and the genomic responses of mollusks, as it is characterized by several climatic transitions (Marcott et al., 2013), e.g., the Little Ice Age in the North Hemisphere (∼14th century to ∼1850 C.E.) or post-industrial global warming (Mann et al., 1999, 2009). During this period of intensifying human impact, mollusks have been exposed to increasing pollution levels that have resulted from rising human population densities, industrial activities, extensive farming, fishing and transportation (Jackson, 2001). They have also been commercially exploited as resources for food, nacre or pearls (Fortunato, 2015), and have been displaced over long distances intentionally for aquaculture purposes, or unintentionally (e.g., through ship transportation) (Carlton, 1999). Although successful DNA recovery may prove challenging, investigating mollusk shell DNA preservation beyond the 7,000 year BP limit holds the potential to provide valuable information about the impact of earlier climatic shifts in the Quaternary, such as the Pleistocene glacial/interglacial/interstadial oscillations, the transition into the Holocene, and subsequent increased human environmental impacts (Stephens et al., 2019). Other important areas of research include the applicability of ancient shell genomics to freshwater and terrestrial mollusks (Lydeard et al., 2004) and determining if ancient mollusk shell DNA enclose useful phylogenetic and population structure information.

To further unlock the potential of ancient mollusk shell DNA for ecological studies, we apply an aDNA HTS approach to a range of samples that extends previous assessments to significantly older specimens, to a larger number of species, as well as to mollusks of the freshwater and terrestrial niches. We then illustrate the importance of ancient mollusk DNA for understanding the impact of environmental change, extinction and biological invasion. We also explore the limitations of such investigations and evaluate the research potential of each genus examined, including genera of commercial interest.

## MATERIALS AND METHODS

### Samples

For aDNA analyses, we selected three fossils identified as mollusk shells, sampled in a river-cut sediment succession along the lower reaches of the Bol'shaya Balaknya River (site BBR17 in Möller et al., 2019b, with close-by subsites 17A and 17B), situated in the southern portion of the Taimyr Peninsula, Russian Federation (N73° 37,084′; E105° 38,178′; **Supplementary Table S1**). This location is south of the limits of repeated Kara Sea-based glaciations during the last glacial cycle (the Zyryanka, Marine Isotope Stage (MIS) 5d-2) (Svendsen et al., 2004; Möller et al., 2015, 2019b). The samples were all found in a distinct marine silty clay designated sediment unit A, overlaid by fluvial sand and gravel (sediment unit B) (see sediment log in Möller et al., 2019b). The marine unit A sediment is rich in bivalve and gastropod shells (**Supplementary Table S2**), including subarctic species such as *Buccinum undatum*, *Mytilus edulis*, and *Macoma balthica*, not present today in the Kara Sea (to the north) and the Laptev Sea (to the east), thus suggesting interglacial conditions at sediment deposition (Möller et al., 2019b). Three shells (*Portlandia arctica*) from the unit A sediments were dated by Electron-Spin Resonance (ESR) to 101–105 kyr (± 9–12 kyr; averaged here to ≥100 kyr). The above-lying unit B fluvial sediments were dated to their deposition by Optically Stimulated Luminescence to 42–43 kyr (± 3–4 kyr), which is within the Middle Zyryanka (MIS 3), while redeposited mollusks (*P. arctica*) within the unit B sediments, eroded from underlying marine unit A, gave ESR ages of 122 and 123 kyr (± 15 kyr). Möller et al., 2019b thus concluded that the unit A marine sediment can be placed with high confidence to their deposition within the Karginsky interglacial (MIS 5e), equivalent to the North West European Eemian. Site BBR 17 is located ∼50 km from the ocean today, but paleoenvironmental reconstructions showed marine inundation of the Taimyr lowlands in front of the retreating ice margin at the transition between the Taz glaciation (MIS 6) and into the Karginsky interglacial (MIS 5e), with sea level at its maximum reaching levels in excess of 80 m above present sea level in this area (Möller et al., 2019b). The shells subjected to aDNA analyses were sampled from frozen marine unit A sediment (permafrost) at site BBR 17A and have been kept frozen since the time of collection in 2010 to optimize DNA preservation post-excavation.

An additional 39 mollusk shell specimens from nine genera (*Arctica*, *Cernuella*, *Crassostrea*, *Dreissena*, *Haliotis*, *Lottia*, *Lymnaea*, *Margaritifera, Mytilus*) were obtained from museum and laboratory collections. When available, genus- or species-level identification based on morphological examination, as well as collection location and date were retrieved from the information curated together with the specimens (**Supplementary Table S1**). Samples were selected to represent genera from marine, freshwater and terrestrial niches, with particular relevance to important ecological questions, such as those related to economical exploitation, environmental changes, biological invasions and extinctions, as well as to optimize comparison with available genomic resources.

## DNA Extraction, Shotgun DNA Library Construction and High-Throughput Sequencing

We carried out shell DNA extraction, shotgun DNA library construction and amplification setup following (Der Sarkissian et al., 2017) in conditions strictly limiting DNA contamination at the ancient DNA facilities of the Centre for GeoGenetics, University of Copenhagen, Denmark. In order to monitor DNA contamination during pre-PCR laboratory work, every step was simultaneously performed on non-template Extraction Blank Controls (EBCs; **Supplementary Table S1**).

Fragments of the shells' ventral margin were first decontaminated for 10 min under agitation in one volume of 1% sodium hypochlorite, before being rinsed three times in three volumes of distilled water, air-dried, and reduced to powder with a mortar and pestle (Der Sarkissian et al., 2017). We then performed DNA extraction based on the method developed in Yang et al., 1998; Gamba et al., 2014, 2016. Shell powder (112–2,990 mg) was incubated overnight at 37°C in 15 mL of digestion buffer (0.45 mM EDTA, 0.5% N-laurylsarcosyl, 0.25 mg/mL proteinase K) under constant mixing. Centrifugation at 3,000 RPM for 2 min allowed separating remaining solids from the liquid phase, which was then concentrated to a volume of ~200 μL using an Amicon Ultra-15 30 kDa centrifugal filter unit (Merck Millipore) at 3,000 RPM for 50–60 min. The retentate was subjected to the MinElute PCR Purification kit (Qiagen), and purified DNA was obtained in a final elution volume of 60 μL EB buffer in presence of Tween (0.05% final concentration, hereafter, EB-Tween).

Double-indexed blunt-ended DNA libraries were constructed following a protocol modified from Orlando et al., 2013; Seguin-Orlando et al., 2013, where adapters carried an identifying 7 bp-index and were paired in a combination unique to each library (Rohland et al., 2015). We used the NEBNext Quick DNA Library Prep Master Mix Set for 454 (New England Biolabs) with a starting volume of 42.6 μL shell DNA extract in reaction volumes of 50 μL for end-repair (12°C for 20 min then 37°C for 15 min) and ligation (20°C for 20 min; 0.5 μM Illumina adapter final concentration), and 25 μL for fill-in (37°C for 20 min then 80°C for 20 min). After end-repair and ligation, the reaction mixes were purified using the MinElute kit with elution volumes of 30 and 20 μL EB-Tween, respectively.

For each library, we used real-time PCR (qPCR) to estimate the optimal number of amplification cycles to obtain DNA amounts compatible with Illumina sequencing, while minimizing library clonality. We performed two independent qPCRs per library in 20 μL of the following reaction mix: 1 μL 1/20 diluted DNA library (in EB buffer), 2.5 units AccuPrime DNA polymerase, 1× AccuPrime mix (Thermo Fisher Scientific), 1 mg/mL BSA, 0.2 μM primer inPE1.0 (5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCC TACACGACGCTCTTCCGATCT-3′) and 0.2 μM of an Illumina 6 bp-indexed ("I") primer (5′-CAAGCAGAAGACGGCA TACGAGATIIIIIIGTGACTGGAGTTCAGACGTGTGCTCTTC CG-3′), and 0.8 μL of 1:4:2000 ROX:SybR:DMSO DNA dye mix (Thermo Fisher Scientific). qPCRs were performed on a LightCycler 480 Real-Time PCR System instrument (Roche Applied Science) with the following conditions: activation at 95°C for 5 min; 40 cycles of: denaturation at 95°C for 15 s, annealing at 60°C for 30 s, elongation at 68°C for 30 s; final elongation at 68°C for 5 min.

We analyzed qPCR data with the LightCycler Software 4.0 and the second derivative maximum method to estimate Ct values used as a relative measure of DNA concentration for each library. We found that libraries built from EBCs differed in Ct values by 4–6 additional cycles from those constructed from ancient shells. As the corresponding 16–64-fold increase in the amount of DNA in ancient samples relative to EBCs was not large enough to confidently rule out a significant impact of laboratory-derived DNA contamination, we subjected EBC libraries to the same PCR amplification and sequencing protocols as ancient shell libraries (see below).

Each double-indexed library was PCR-amplified using one Illumina IS4 primer and one unique 6 bp-indexed primer, thus, resulting in the introduction of a third, external index. The same conditions as for qPCR were applied, apart from non-diluted DNA and total reaction volumes of 5 and 25 μL, respectively, and from the absence of ROX:SybR:DMSO. We derived the optimal number of cycles to apply from the qPCR results on a per-library basis (14–29 for ancient samples and 34–35 for EBCs; **Supplementary Table S1**). In order to limit clonality, all libraries were first amplified for 12 cycles, before MinElute-purified reaction mixes (25 μL EB elution volume) were re-amplified as above for the required remaining number of cycles and in four independent reaction volumes (25 μL each) per library. Amplified libraries were purified using the Agencourt AMPure XP system (Beckman Coulter) and 25 μL of bead solution. Libraries eluted in 25 μL EB were quantified on a 2200 TapeStation Instrument (High Sensitivity D1000 Screen Tape; Agilent).

All triple-indexed libraries were pooled together in equimolar proportions and sequenced in paired-end mode on an Illumina HiSeq4000 platform at the Danish National High Throughput DNA Sequencing Centre. Raw sequencing data were deposited on the European Nucleotide Archive (ENA) public database (project PRJEB35671). Previously published ancient mollusk DNA data from Der Sarkissian et al. (2017) were retrieved from ENA (project PRJEB20113) and reanalyzed here.

## Post-sequencing DNA Read Processing

Post-sequencing, we de-multiplexed DNA reads and only retained those displaying the expected, unique combination of indexes, i.e., unlikely to represent chimeric sequences. We used PALEOMIX v1.2.13 (Schubert et al., 2014) to trim adapters and collapse overlapping pair-end mates with AdapterRemoval2 v2.3.0 (Schubert et al., 2016) as described in Fages et al. (2019), i.e., applying default parameters and quality/length filters –minlength 25 –trimns –trimqualities –minquality 2.

## Mapping to Mitochondrial and Nuclear Reference Sequences

Initial molecular taxonomic assignment was performed by mapping shotgun DNA reads to 145,457 reference sequences of the mitochondrial *cytochrome c oxidase subunit I* gene 5′-extremity barcode region (COI-5P) compiled for the Mollusca phylum from the Barcode of Life Database (BOLD[1]). Building on the results, reads were then mapped to reference sequences for mitochondrial complete genomes, nuclear genome or transcript assemblies, when available for the identified genus/species (**Supplementary Table S3**). Some of the genera studied show Doubly Uniparental Inheritance, where the mitochondrial genome of the mother (or F-genome) is transmitted to both the somatic and gonadic cells of the female offspring, but only to the somatic cells of the male offspring, the gonadic cells of which get their mitochondrial genome from the father (M-genome). As a consequence, we used both F- and M-genomes as reference sequences for mapping (**Supplementary Table S3**). The end of the circular reference mitochondrial genome sequences was extended using the first 30 bp of the sequence in order to take the circularity of the mitochondrial genomes into account. We used the BWA v0.5.9 *aln* alignment command (Li and Durbin, 2009) in PALEOMIX (Schubert et al., 2014) with default parameters and disabled seeding as recommended in Schubert et al., 2012. We kept all reads mapping to the BOLD Mollusca database, whereas only reads showing mapping qualities ≥30 against the mitochondrial and nuclear DNA reference sequences were retained. We removed duplicated non-collapsed reads with MarkDuplicates in Picard Tools version 1.119[2] and duplicated collapsed reads with the PALEOMIX FilterUniqueBAM Python script (Schubert et al., 2014). Average depth-of-coverage (coverage) was calculated using PALEOMIX and considering only unique high-quality reads (after quality filtering and duplicate removal). When multiple reference sequences were available for a given genus, we retained the one yielding the highest coverage for each sample.

## Complete Mitochondrial Genome Assembly and Maximum Likelihood Phylogenetic Reconstructions

For samples showing maximal coverage ≥3× when mapping against complete mitochondrial genome reference sequences, we built complete mitochondrial genome consensus sequences from which we constructed maximum-likelihood phylogenies. This was achieved using the perl script wrapper.pl (for circular genomes) as well as the C++ programs bam2prof, endoCaller and log2fasta within the schmutzi pipeline (Renaud et al., 2015), taking into consideration rates of cytosine deamination at the 5′- and 3′- read ends, and only retaining bases with quality ≥30. For each sample, we called a first consensus from the rescaled BAM file, then re-mapped all shotgun reads to the newly built consensus with PALEOMIX as described above, and called a final consensus from the resulting rescaled BAM with schmutzi.

We then constructed alignments from the consensus sequences for ancient samples obtained here and in Der Sarkissian et al. (2017), as well as all sequences previously published for the genus/species of interest. We extracted the sequences of 12 genes (*CYTB*, *COX2*, *NADH1*, *NADH4*, *COX3*, *NADH2*, *NADH3*, *COX1*, *ATP6*, *NADH4L*, *NADH5*, *NADH6*) and aligned them independently with PRANK (Löytynoja and Goldman, 2005). For the analyses of *Mytilus* sp. sequences, we added the sequences of the 12S ribosomal RNA gene (12S rRNA), and 23 transfer RNAs (tRNA). Maximum-likelihood trees were reconstructed using the program PhyML 3.0 (Guindon et al., 2010) with an approximate likelihood-ratio test (aLRT) and a Shimodaira-Hasegawa-like procedure for branch support estimation. Using the SMS heuristic procedure for model selection implemented in PhyML 3.0 (Lefort et al., 2017), we assessed the best substitution models for the concatenated alignments: the Generalized Time-Reversible model, with invariant sites and 4 gamma site-rate categories for the *Mytilus* DNA alignment, and the Variable Time substitution model with invariant sites, a gamma site-rate distribution and alignment-based estimation of amino-acid frequency equilibrium for the mollusk protein alignment.

## Population Structure Inference Based on Mitochondrial DNA

For all samples positively identified through comparison with the BOLD marker database and showing an average coverage ≥3×, a COI-5P consensus sequence was built as described for complete mitochondrial genomes. The obtained consensus was then compared to all the sequences available in the BOLD database for the corresponding genus/species using Median Joining Network (epsilon = 0) in POPART v1.7 (Leigh and Bryant, 2015) and the vouchers' geographical origin as curated in the BOLD database metadata. When available, published mitochondrial sequences from modern populations for species of interest were retrieved for comparison to ancient sequences using the same procedure. These comprised *Mytilus trossulus* (Breton et al., 2006; Smietanka et al., 2010, 2013; Zbawicka et al., 2014b; Śmietanka and Burzyński, 2017), *A. islandica* (Glöckner et al., 2013), *Ruditapes decussatus* (Cordero et al., 2014; Sanna et al., 2017), *R. philippinarum* (Cordero et al., 2017), and *Crassostrea angulata* (Grade et al., 2016).

## Population Structure Inference Based on Nuclear DNA Data

The ≥100 kyr Tx101A *Mytilus* sample from the Taimyr Peninsula could be compared to modern *Mytilus* populations for which genomic data are publicly available (Fraïsse et al., 2016, 2017). This comparative dataset was previously obtained through capture-based enrichment sequencing targeting 4.3 Mb DNA sequences assembled from *M. edulis* BAC clones and

[1] http://www.boldsystems.org

[2] http://broadinstitute.github.io/picard/

*M. galloprovincialis* cDNAs in 1,269 contigs (Fraïsse et al., 2016, 2017). It is composed of 80 individuals from ten populations: *M. trossulus* from the Baltic Sea ($N = 8$) and the North American Atlantic coast ($N = 7$), *M. edulis* from the North American Atlantic coast ($N = 11$), as well as from the Wadden Sea ($N = 8$) and the Bay of Biscay ($N = 8$; respectively, outside and inside the Atlantic hybrid zone with *M. galloprovincialis*), *M. galloprovincialis* from Portugal ($N = 6$) and Brittany ($N = 8$; respectively, outside and inside the Atlantic hybrid zone with *M. edulis*), as well from the Occidental ($N = 8$) and Oriental Mediterranean basins ($N = 8$), and *M. platensis* from the Kerguelen Islands ($N = 8$). In order to avoid computational analysis biases, the raw data published in Fraïsse et al. (2016, 2017) were re-mapped to the contig reference sequence at the same time as raw data for the Tx101A ancient sample and EBCs. The same BWA mapping parameters were used as described above for both ancient and modern samples, except that the options for mismatch penalty (-M 2), gap open penalty (-O 3) and minimum seed length (-k 10) were added as in Fraïsse et al. (2016, 2017), and that seeding was disabled for Tx101A and EBCs.

Low depth-of-coverage for Tx101A when mapped against the *Mytilus* contigs precluding accurate genotype calling, subsequent analyses where carried out using genotype likelihood in ANGSD v.0.929 (Korneliussen et al., 2014). Genotype likelihoods were estimated with the following options and filters: -doMajorMinor 1 -doMaf 1 -remove_bads 1 -uniqueOnly 1 -minMapQ 30 -C 50 -minQ 30 -baq 1 -skipTriallelic 1 -SNP_pval 1e-6 -doHWE 1 -maxHWEpval 0.05. All analyses were performed either considering all substitution types, or excluding transitions (-rmTrans 1). We used PCAngsd v.0.98 (Meisner and Albrechtsen, 2018) and the option -minMaf 0.05 to perform Principal Component Analysis (PCA) and determine the optimal number of principal components to describe the *Mytilus* population structure using the minimum average partial (MAP) test. Admixture proportions were estimated using NgsAdmix (Skotte et al., 2013) for a number of clusters equal to the previously determined optimal number of principal components +1 (Meisner and Albrechtsen, 2018), i.e., four, here.

## Post-mortem DNA Damage Characterization

We used mapDamage v2.0.1 (Jónsson et al., 2013) to characterize ancient shell DNA in terms of fragmentation patterns, nucleotide mis-incorporation and fragment size distributions. On the basis of 100,000 Markov Chains Monte Carlo iterations, posterior distributions were obtained for the following damage parameters: rates of deamination in double strands ($\delta_D$) and single strands ($\delta_S$), probability of reads not terminating in overhangs ($\lambda$, transformed into $1/\lambda-1$, a proxy for the length of overhanging regions). Quality scores of bases likely to represent damaged read positions were then rescaled using default parameters.

## Assessing the Potential of Ancient Mollusk DNA for Ecological Studies

We provided a summarizing visualization of the potential for ecological studies of each mollusk genus investigated. In genus-specific radar graphs, we represented: their molecular preservation potential (oldest authenticated mollusk DNA data, DNA recovery rate), the availability or unavailability (encoded as 1 and 0, respectively) of comparative datasets (BOLD barcodes, complete mitochondrial genomes, nuclear transcripts and genome assemblies), and their relevance to ecological questions with regards to the research themes “Environment,” “Commercial” exploitation, and biological “Invasion” (assessed on the basis of bibliography and encoded as 1, relevant and 0, not relevant), and “Extinction.” An “extinction score” was given for each vulnerability category defined by The IUCN Red List of Threatened Species (2019): i.e., 1 = “least concerned,” 2 = “near threatened,” 3 = “vulnerable,” 4 = “endangered,” 5 = “critically endangered,” 6 = “extinct in the wild,” 7 = “extinct,” and a genus extinction score was calculated as the average score across species listed for each genus.

# RESULTS

## Mitochondrial DNA-Based Taxonomic Identification of ≥100 kyr Mollusk Shells

We tested for long-term preservation of aDNA in marine mollusk shells by constructing shotgun DNA libraries using extracts from three mollusk shell samples (Tx100, Tx101A, Tx103) dated to ≥100 kyr. For the three libraries, qPCR showed copy-numbers compatible with HTS, which yielded 653,177–19,314,607 collapsed and non-collapsed pair-end mate reads after trimming and quality filtering (**Table 1** and **Supplementary Table S1**).

Taxonomic identification was achieved by mapping against the BOLD Mollusca mitochondrial COI-5P marker reference sequences. No hit was obtained for Tx100, precluding any classification for this specimen. On the basis of mitochondrial DNA, Tx101A could be identified as *M. trossulus* since 98.6% of the mapped nucleotides aligned to markers of this species, plus 0.9 and 0.5% aligning to markers of the *Mytilus* genus and the Mytilida order (**Supplementary Table S4**). Tx101A had been previously identified as *M. edulis* on the basis of shell morphology in Möller et al. (2019a,b), but *M. trossulus* and *M. edulis* can be difficult to distinguish due to their similarity in shape and size (Innes and Bates, 1999; Riginos and Cunningham, 2005). The morphological assignment of the Tx103 sample to *P. arctica* could be confirmed as 27.3% of the mapped nucleotides align to markers of this species, and 72.7% to the Nuculanida order (**Supplementary Table S4**).

We attempted to gain more confidence in our taxonomic assignment, and retrieve more specific information about genomic affinities of Tx101A and Tx103 with modern individuals (e.g., at the population level), by comparing their COI-5P sequences with *M. trossulus* and *P. arctica* sequences available in the BOLD database. Median Joining Network analyses, however, proved uninformative due to the limited phylogenetic resolution of the COI-5P marker, the scarcity of geographic information provided for BOLD vouchers (Tx101A), and the lack of comparative *P. arctica* entries (Tx103) (**Supplementary Figure S1**).

**TABLE 1 |** Sample information and sequencing statistics.

| Sample name | Organism | Geographical location | Date/Age | Niche | #Reads | #Retained |
|---|---|---|---|---|---|---|
| Tx100 | unidentified | Taimyr Peninsula, Russia | ≥100 kyr[1] | marine | 6,747,556 | 653,177 |
| Tx101A | *Mytilus* sp. | | | | 18,369,198 | 12,706,997 |
| Tx103 | *Portlandia arctica* | | | | 36,524,098 | 19,314,607 |
| quahog1B | *Arctica islandica* | Unknown | 1722 C.E. | marine | 30,050,698 | 14,592,438 |
| onuA | | Önundarfjordur, Iceland | 1893 C.E. | | 13,074,594 | 6,446,403 |
| bakA | | Bakkaflöt, Iceland | 1900 C.E. | | 56,559,606 | 30,000,220 |
| ves4A | | Westman Islands, Iceland | 1900 C.E. | | 17,287,016 | 7,073,491 |
| flaA | | Flatey, Iceland | 1903 C.E. | | 4,591,784 | 2,477,750 |
| ran1 | | Vidarvik, Iceland | 1903 C.E. | | 49,501,816 | 27,291,420 |
| hel2A | | Hellebaek, Denmark | 1910 C.E. | | 34,880,594 | 19,210,163 |
| bai2 | | Bair- Ísafjör∂̃ur, Iceland | 1933 C.E. | | 28,668,834 | 13,621,176 |
| ore2A | | Øresund, Denmark | 1946 C.E. | | 47,382,554 | 26,878,824 |
| cop1 | *Dreissena polymorpha* | Copenhagen, Denmark | 1925 C.E. | freshwater | 11,970,824 | 11,112,715 |
| tys1 | | Tystrup Lake, Denmark | 1941 C.E. | | 15,009,888 | 7,731,530 |
| cal4 | *Haliotis cracherodii* | Pacific, United States | 1946 C.E. | marine | 19,342,648 | 12,741,676 |
| cal1 | *Haliotis rufescens* | Pacific, United States | Unknown | marine | 19,291,324 | 11,078,316 |
| nyk1 | *Lymnaea stagnalis* | Nykøbing, Denmark | 1906 C.E. | freshwater | 29,277,948 | 14,300,382 |

*"kyr," kilo years; "C.E.," Common Era; [1] Möller et al., 2019b.*

## Mitochondrial and Nuclear Genomic Affinities of a ≥100 kyr Year-Old Mussel

In order to achieve a more robust identification and a characterization of the Tx101A specimen with more resolution, we performed phylogenetic and population structure analyses based on HTS read alignments to all complete mitochondrial genomes published to date, as well as to genome-wide contigs for *Mytilus*. In absence of such public reference resources for *P. arctica*, or any close relative taxon, further genomic characterization of the Tx103 specimen was not achievable. The analyses provided further evidence for the mitochondrial identification of Tx101A as *M. trossulus*. The maximal coverage of the mitochondrial genome was, indeed, obtained when aligning HTS reads to the sequences of *M. trossulus* F-genomes or recently masculinized F-genomes (24.7–25.7× maximum for GenBank accession number GU936625; **Supplementary Table S5**). Other *Mytilus* mitochondrial genomes led to 10–9,267-fold lower coverage (**Supplementary Table S5**). This pattern was robust to relaxing or strengthening mapping stringency (-n, BWA edit distance option) (**Supplementary Figure S2**).

A Maximum-likelihood tree reconstructed from concatenated mitochondrial genome sequence partitions confirmed the phylogenetic placement of Tx101A within the clade of *M. trossulus* (recently masculinized) F-genomes, where it occupies a basal position (**Figure 1A**). Patterns of coverage are consistent with the phylogenetic position of Tx101A, i.e., the more distant to *M. trossulus* the mapping reference genome, the lower the coverage (**Figure 1B** and **Supplementary Table S5**). When only a 1,171 bp concatenated sequence partition of mitochondrial genes and transfer RNAs from Tx101A was compared to data from 187 present-day *M. trossulus* individuals (Breton et al., 2006; Smietanka et al., 2010, 2013; Zbawicka et al., 2014b; Śmietanka and Burzyński, 2017), the outlying position of Tx101A was confirmed. Closest relatives equidistant to the Tx101A haplotype were found in the broadly distributed present-day populations of the Baltic, the eastern and western coasts of the North Atlantic and the North Pacific (**Figure 1C**).

As hybridization can occur amongst members of the *M. edulis* complex (*M. trossulus*, *M. edulis*, *M. galloprovincialis*), we examined the nuclear genomic background of Tx101A in order to validate its taxonomic identification and/or estimate potential admixture proportions. PCA of 11,882 variable sites (considering transversions only; 29,440 variable sites considering all substitutions in **Supplementary Figure S3**) in 80 individuals and along the optimal number of three components (as determined by the MAP test in PCAngsd) confirmed that Tx101A belongs to the *M. trossulus* population cluster, clearly segregating from the other clusters of the *M. edulis* complex and of the *M. platensis* outgroup (**Figure 1D** and **Supplementary Figure S3**). This result was confirmed by clustering analyses considering four ancestries, in which a genome-wide affinity of Tx101A for *M. trossulus* modern populations and no admixture with another *M. edulis* complex or outgroup population were detected (considering transitions alone; **Figure 1E** and **Supplementary Figure S4**).

## Authentication of the DNA Data Obtained From the ≥100 kyr Old Mollusk Shells

Considering that the age of the Tx101A and Tx103 samples dramatically pushes back the limit of marine mollusk shell DNA preservation from ~7,000 years to ≥100 kyr, particular attention was paid to the line of evidence supporting data authenticity: (1) Tx101A and Tx103 were excavated at the same site, but yielded DNA reads that were identified as representative of different Bivalvia sub-classes; (2) our molecular classification confirms previous morphological identification of Tx101A as *Mytilus* and Tx103 as *P. arctica* (Möller et al., 2019b,a); (3) the detection of *M. trossulus* and *P. arctica* at

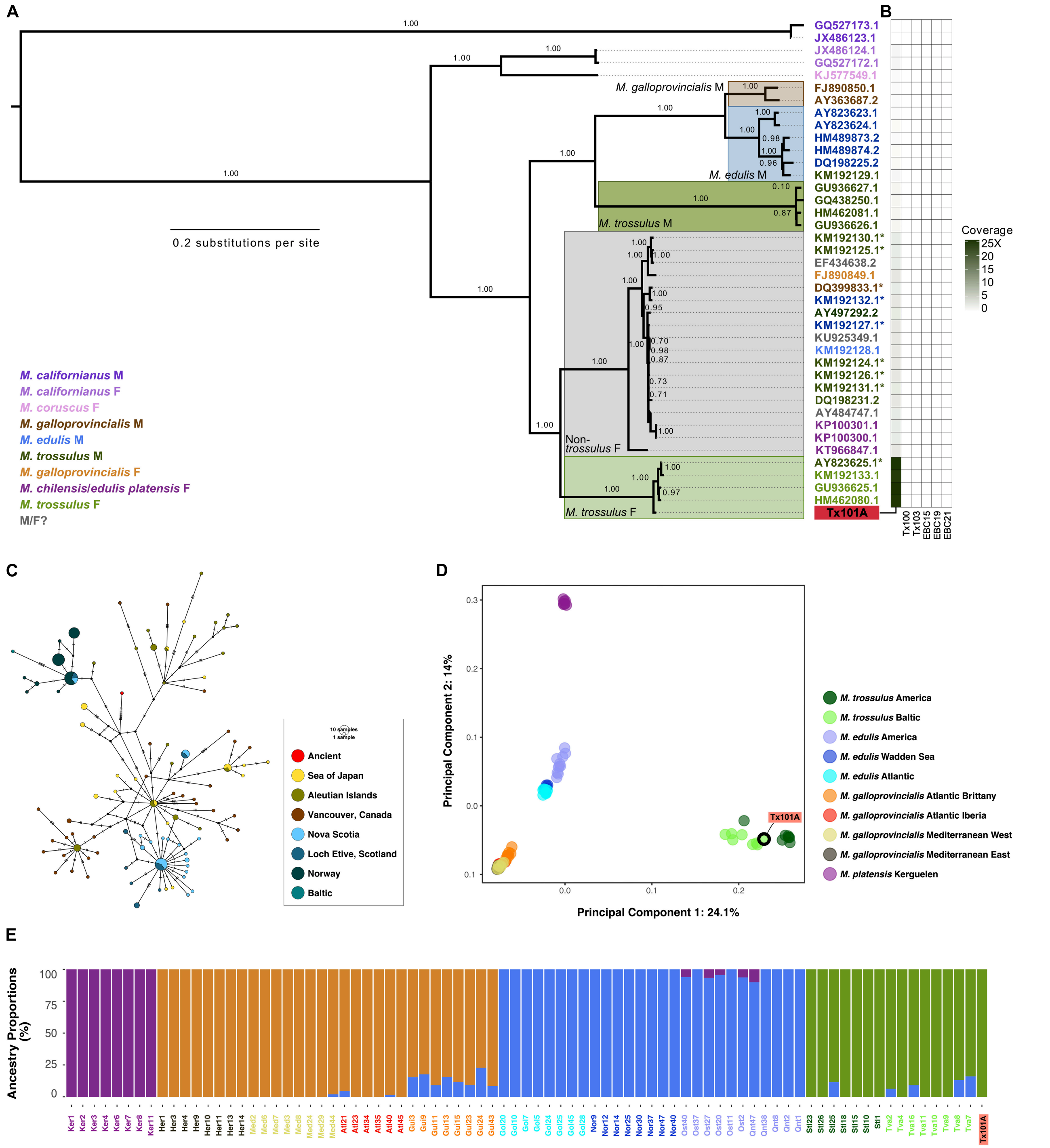


**FIGURE 1 |** Phylogenetic and population affinities of Tx101A within present-day *Mytilus*. **(A)** Maximum likelihood phylogenetic tree built from a concatenated alignment of *Mytilus* mitochondrial sequences for 12 protein-coding genes, the 12S rRNA gene and 23 tRNAs (total length 14,277 bp). The sex and species of modern individuals were determined based on morphology and nuclear DNA markers. Branch support ≥0.70 is shown as Shimodaira-Hasegawa approximate Likelihood Ratio Test. *Recently masculinized mitochondrial genomes. **(B)** Average depth-of-coverage when mapping against each *Mytilus* mitochondrial reference genome considering unique high-quality reads (≥30). **(C)** Median-Joining network built from a 1,171 bp concatenated alignment of mitochondrial genes (*COX3*, *ND2*) and transfer RNA (Ser, Met) from 187 present-day *M. trossulus* individuals. **(D)** Principal Component Analyses of the top two principal components in a dataset comprising 11,882 variable sites (considering transversions only) for 80 individuals. The proportion of the variance explained by each component is indicated for each axis. **(E)** Ancestry proportions estimated considering four ancestral populations (transversions only). Individual color code is the same as in **(D)**.

the BBR 17A site was ecologically coherent with these being considered as (sub-) arctic species of the littoral or deeper marine environments (Möller et al., 2019b); (4) none of the HTS datasets obtained for the extraction blank controls (EBCs) led to the identification of *M. trossulus* or *P. arctica* DNA sequences. Compared to those observed for Tx101A and Tx103, significantly lower coverage were attained when mapping EBC reads against reference sequences for: *M. trossulus*/*P. arctica* BOLD COI-5P (no hit), *Mytilus* complete mitochondrial genomes (≥2,335-fold coverage decrease), the *M. galloprovincialis* genome assembly (≥16,563-fold coverage decrease), *Mytilus* sp. genome-wide contigs (≥5,704-fold coverage decrease); (5) for both Tx101A and Tx103, high-quality HTS reads aligned to the various reference sequences displayed molecular signatures characteristic of *post-mortem* DNA fragmentation and cytosine deamination-induced mis-incorporation. When considering the COI-5P alignment, Tx103 showed reads of ∼66.6 bp average length and a 9.4% C-to-T mis-incorporation rate at the 5′-read extremities. As for Tx101A, when considering mitochondrial genome/nuclear contig alignments, HTS reads exhibited ∼82.8/78.5 bp reads, maximum 3.9/6.4% C-to-T mis-incorporation rates, post-mortem deamination rates in overhangs 4.9/5.2-larger than in double stranded DNA, and overhanging ends of 3.0/2.0 bp (1/λ – 1), as expected. In addition, a 10–bp periodicity in the size distribution was observed in the nuclear DNA of Tx101A, but not in its mitochondrial DNA, a specific pattern that was previously proposed to reflect nucleosome protection in authentic and degraded ancient nuclear DNA (Pedersen et al., 2014; **Figure 2** and **Supplementary Table S6**); and (6) all phylogenetic and population structure analyses of mitochondrial and nuclear DNA were consistent with Tx101A being part of *M. trossulus*. Combined, these results provide compelling evidence that the DNA obtained from ≥100 kyr mollusk shells is authentic.

## Expanding the Exploration of Ancient DNA Recovery From Mollusk Shells

We next extended the assessment of DNA recovery from ancient mollusk shells to 39 new specimens, 14 of which led to shotgun HTS datasets containing 2,477,750–30,000,220 quality-filtered collapsed and non-collapsed pair-end mate DNA reads (**Table 1** and **Supplementary Table S1**). In subsequent analyses, these were added to 32 previously published ancient shell data (Der Sarkissian et al., 2017) and the two ≥100 kyr Taimyr samples described above to build a dataset comprising 48 specimens from 13 species and 9 genera, with 73% of the samples dated to the last 300 years of the post-industrial era (**Figure 3** and **Supplementary Table S1**). Success rates were variable across genera, with the highest rates observed for *Ruditapes*/*Venerupis* sp. (100%, *N* = 20), *Haliotis* sp. (70%, *N* = 7) and *A. islandica* (57%, *N* = 12). No successful DNA recovery was achieved for the *Lottia gigantea* (*N* = 1), *Margaritifera margaritifera* (*N* = 3), and the terrestrial *Cernuella virgata* specimens (*N* = 2), thus limiting investigations to marine (*N* = 13) and freshwater individuals (*N* = 3) in this study. We found no indication that the overall success rate depends on the amount of shell powder analyzed, as specimens yielding successful and unsuccessful DNA recovery did not show significant differences in input shell median weight (Wilcoxon signed-rank test, *p*-value = 0.3).

## Mitochondrial DNA-Based Taxonomic Identification

Mapping to the Mollusca database of BOLD COI-5P markers provided sufficient read alignments for taxonomic identification, as all 48 samples, but the LEFB *Pecten* sp. shell, could be classified at least to the genus level (**Figure 4** and **Supplementary Table S4**). *A. islandica*, *C. angulata*, *Haliotis tuberculata*, *R. decussatus*, *R. philippinarum*, and *Dreissena polymorpha* specimens were unambiguously identified at the species level, and even subspecies level for *H. tuberculata coccinea* and *H. tuberculata tuberculata*. We noted, however, 6–30% false positive rates for *H. rufescens*, *V. corrugata*, and *Lymnaea stagnalis* at the genus level. Through resequencing, we also assembled 29 additional complete mitochondrial genomes with coverage ≥2.8-fold that were used to reconstruct a maximum-likelihood tree of mitochondrial genes. The *L. stagnalis* and *V. corrugata* specimens could not be included in this analysis as no reference sequences are available for these species. The obtained tree retraces known phylogenetic relationships within gastropods and bivalves, but also confirms the taxonomic identification based on COI-5P sequences (**Figure 5**). Interestingly, with five new *R. decussatus* mitochondrial genomes assembled here from ancient specimens, we could confirm a previous hypothesis based on one modern *R. decussatus* mitochondrial genome that *R. decussatus* is more closely related to the *Paphia* sp. than to *R. philippinarum* despite being named as members of the same genus (Ghiselli et al., 2017).

## Population Affinities Between Ancient and Present-Day Mollusks

Little information about affinities between ancient and modern populations could be retrieved from the phylogenetic tree owing to the limited number of published complete mitochondrial genomes in mollusks. Although more mitochondrial COI-5P sequences are available, analyses proved poorly informative for most genera due to the paucity of sequenced present-day individuals (*Pecten maximus*) and corresponding geographical metadata (*P. maximus*, *L. stagnalis*). Additionally, many ancient haplotypes were found similar to the most abundant and widely distributed haplotypes in modern populations partly due to the limited resolution of the COI-5P marker (*C. angulata*, *D. polymorpha*, *R. decussatus*; **Supplementary Figure S1**). The mitochondrial phylogenetic tree could, however, show a segregation within *R. decussatus* between the ∼30 year BP Adriatic Sea MURp individual from the 400–5,500 year BP Atlantic Ocean auzay1B, auzay3B, lmc1B, and lmc3B individuals (**Figure 5**). Within the other clam species *R. philippinarum*, both the mitochondrial gene tree and the COI-5P network showed a differentiation between the 1988 C.E. POSp French specimen and the 1983–2012 C.E. LAN1p, LAN2p, LAN3p, NEG, MATp, AKKp and KORp samples from France, Japan and Korea (**Figure 5** and **Supplementary Figure S1**). For *A. islandica* quahogs, we compared our 73–297 year-old dataset to 24 COI-5P sequences from the BOLD database (**Supplementary Figure S1**)

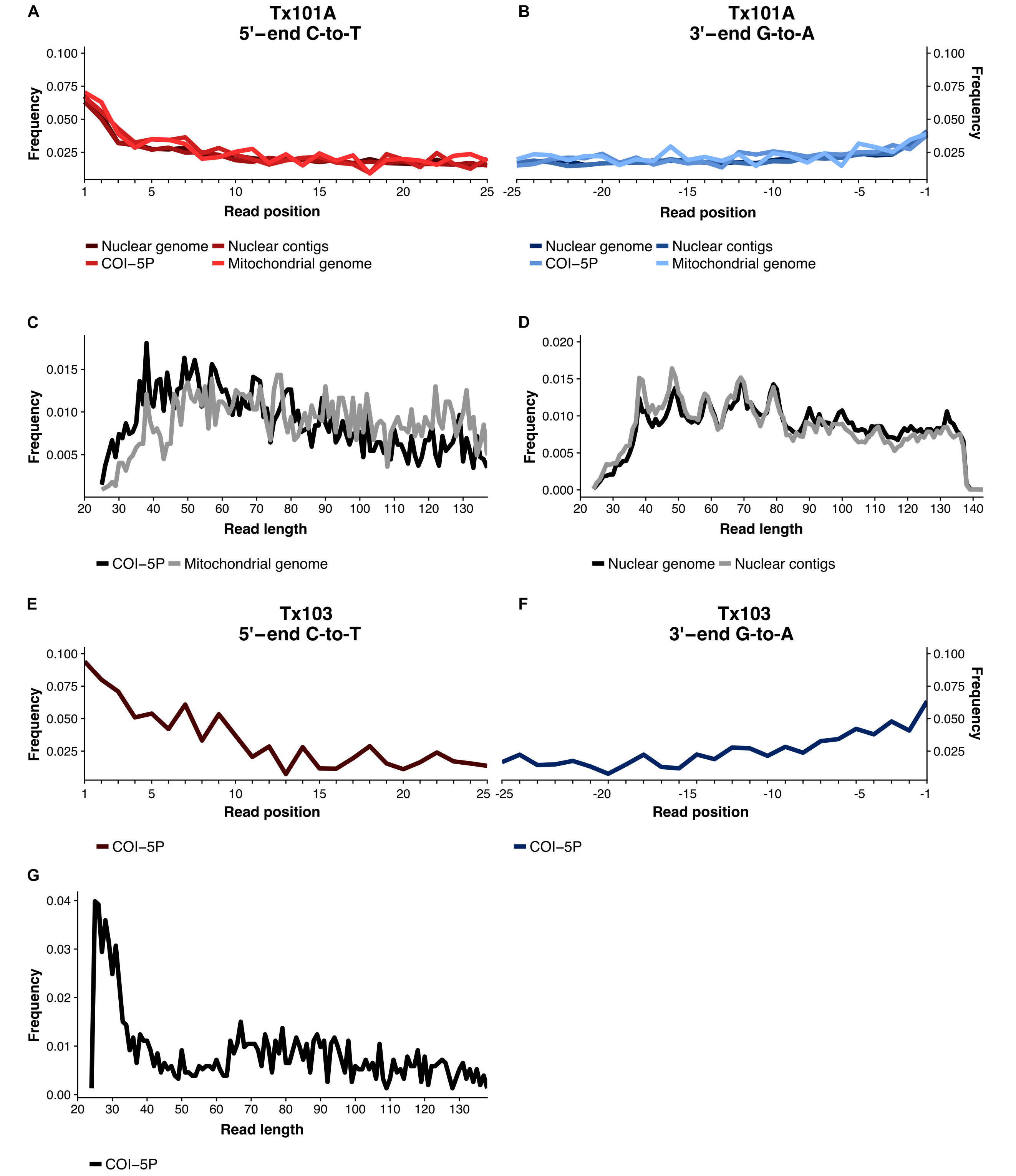


**FIGURE 2 |** Post-mortem DNA damage patterns. **(A)** For Tx101A, rates of C-to-T nucleotide mis-incorporation at 5′-read ends, **(B)** rates of G-to-A at 3′ read-ends. **(C)** Size distribution of reads mapping to mitochondrial, and **(D)** nuclear reference sequences. **(E)** For Tx103, rates of C-to-T nucleotide mis-incorporation at 5′-read ends, **(F)** Rates of G-to-A at 3′ read-ends. **(G)** Size distribution of reads mapping to mitochondrial reference sequences. Only unique high-quality (MQ≥30) reads were considered and mapDamage v.2 was run on the full alignment (100,000 iterations).

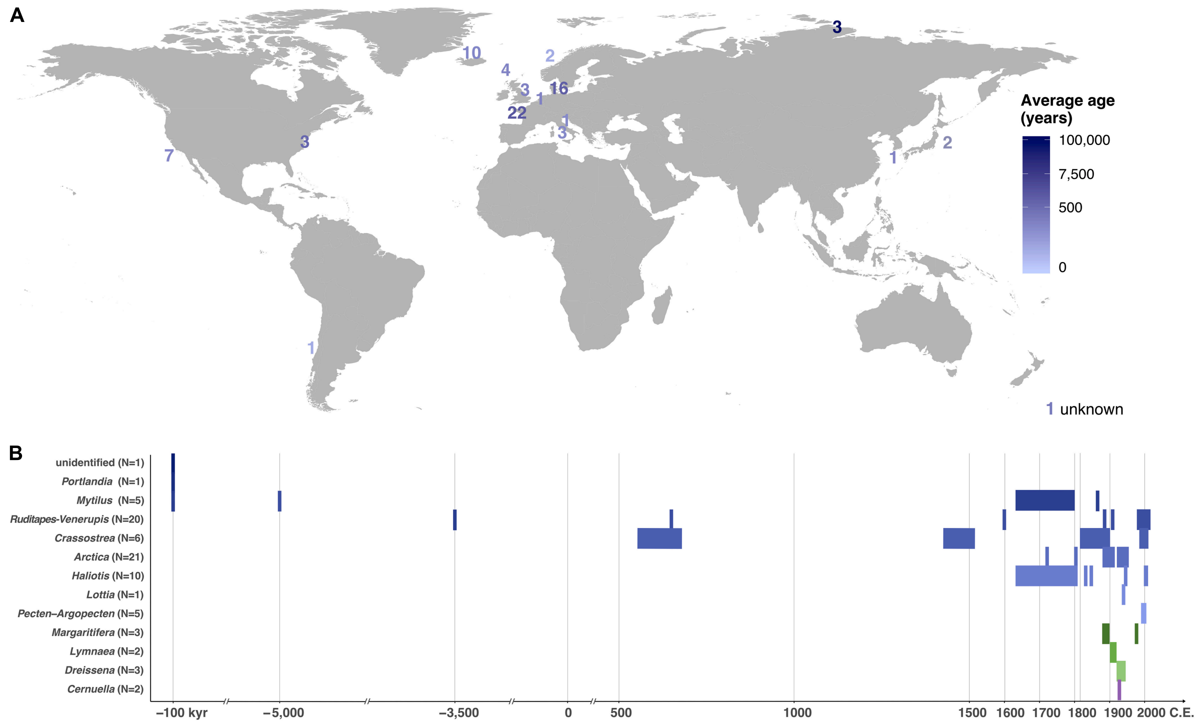


**FIGURE 3 |** Geographical location and age of the shell samples screened for DNA. **(A)** Map showing the number and average age of specimens by geographical location. **(B)** Chronology by genus (**Supplementary Table S1**). "kyr," kilo years.

as well as to 3,029-bp long mitochondrial sequences for 20 modern individuals (Glöckner et al., 2013). Modern quahogs belong to two main clades: one comprising individuals from Iceland, the Baltic and the North Seas, and one composed of individuals from Iceland and the Baltic Sea, with the Icelandic and the Baltic Sea populations displaying a larger diversity than the North Sea population (Glöckner et al., 2013). We found the same pattern in the past as our sequences from ancient quahogs fell within the two clades described previously (**Supplementary Figure S5**). More present-day data in the form of complete mitochondrial genome sequences are needed to rule out possible sampling biases in both the number and geographical origin of the sequenced individuals.

## Authentication of Ancient Mollusk DNA Data

Authenticity of the aDNA retrieved from the mollusk shells was supported by: (1) congruent molecular and morphological identification, at least at the genus level; (2) increased coverage when mapping against the complete mitochondrial genome reference sequence for the identified species (**Supplementary Figure S6**); (3) phylogenetic placement of the COI-5P barcodes and mitochondrial genes (**Figures 4**, **5** and **Supplementary Figure S1**); and (4) the presence of typical molecular signatures of mitochondrial DNA *post-mortem* damage in the form of short fragment sizes (45.1–95.2 bp), increased rates of C-to-T mis-incorporation rates at read 5′-ends (0.0–32.2%), increased mis-incorporation rates in single- versus double-strand contexts $\partial_S/\partial_D$ (2.3–70.3), and decreased values for the overhang-length proxy $1/\lambda - 1$ (0.1–4.7). Similar patterns were observed in reads mapping against nuclear reference sequences: DNA fragmentation (41.9–107.8 bp), C-to-T mis-incorporation rates at read 5′-ends (0.0–17.4%), $\partial_S/\partial_D$ (0.1–184.6), and $1/\lambda - 1$ (0.2–16.3) (**Supplementary Table S6**).

# DISCUSSION

## Preservation of ≥100 kyr DNA From Marine Mollusk Shells in Siberian Permafrost Sediments

In this work, we broaden the assessment of mollusk DNA recovery from ancient shells in terms of time scale, species, and applications for ecological studies. The successful retrieval of aDNA from ≥100 kyr shells significantly extends the temporal scale of mollusk aDNA analyses, as the oldest specimens that had previously yielded aDNA were ∼7,000 year-old *Mytilus* shells collected from shell middens in Denmark (Der Sarkissian et al., 2017). The constantly frozen Siberian permafrost marine sediments from which the ≥100 kyr shells were retrieved (Möller et al., 2019b, a) may have provided exceptionally favorable DNA preservation conditions minimizing damage induced by both water and microbial activity (as reviewed in Pedersen et al., 2015). As examples of long-term DNA preservation in similar conditions, the oldest remains for which animal aDNA could be recovered in Siberia were >63.5 kyr and ∼50 kyr mammoth hair,

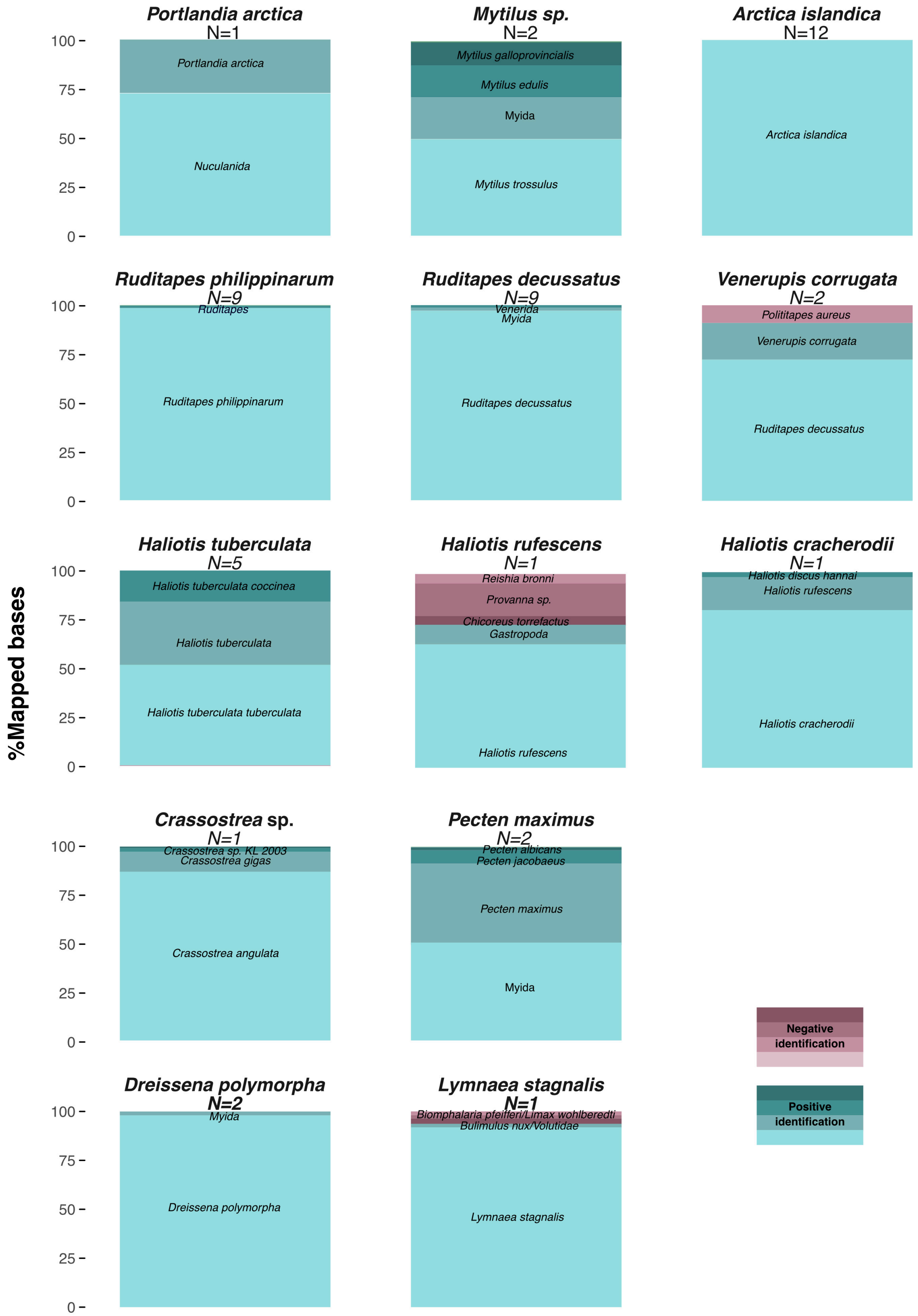


**FIGURE 4 |** COI-5P barcode-based identification of Mollusca taxa from shell DNA. Proportion of total nucleotides mapped to each reference marker combining all datasets for samples of the same genus and considering unique high quality (≥30) reads and reference markers covered by ≥150 bp (**Supplementary Table S4**).

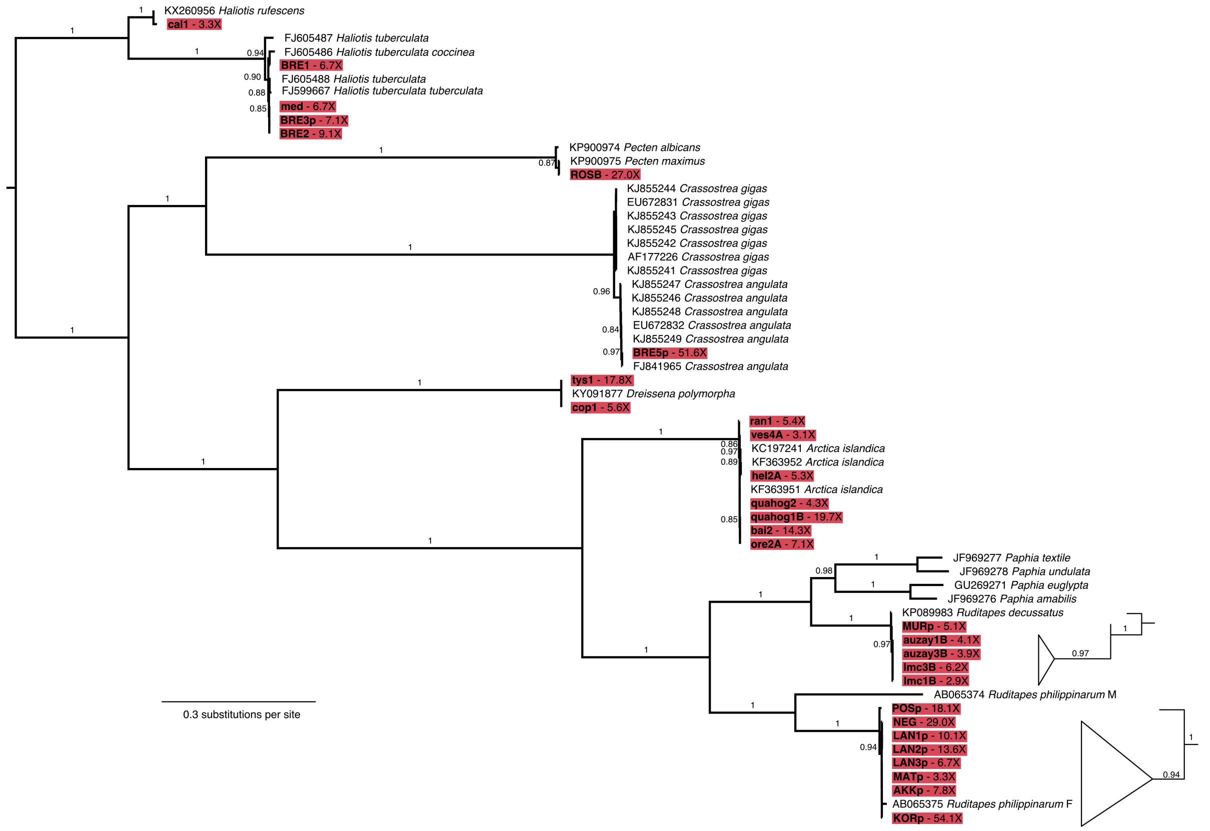

KX260956 Haliotis rufescens
cal1 - 3.3X
FJ605487 Haliotis tuberculata
FJ605486 Haliotis tuberculata coccinea
BRE1 - 6.7X
FJ605488 Haliotis tuberculata
FJ599667 Haliotis tuberculata tuberculata
med - 6.7X
BRE3p - 7.1X
BRE2 - 9.1X
KP900974 Pecten albicans
KP900975 Pecten maximus
ROSB - 27.0X
KJ855244 Crassostrea gigas
EU672831 Crassostrea gigas
KJ855243 Crassostrea gigas
KJ855245 Crassostrea gigas
KJ855242 Crassostrea gigas
AF177226 Crassostrea gigas
KJ855241 Crassostrea gigas
KJ855247 Crassostrea angulata
KJ855246 Crassostrea angulata
KJ855248 Crassostrea angulata
EU672832 Crassostrea angulata
KJ855249 Crassostrea angulata
BRE5p - 51.6X
FJ841965 Crassostrea angulata
tys1 - 17.8X
KY091877 Dreissena polymorpha
cop1 - 5.6X
ran1 - 5.4X
ves4A - 3.1X
KC197241 Arctica islandica
KF363952 Arctica islandica
hel2A - 5.3X
KF363951 Arctica islandica
quahog2 - 4.3X
quahog1B - 19.7X
bai2 - 14.3X
ore2A - 7.1X
JF969277 Paphia textile
JF969278 Paphia undulata
GU269271 Paphia euglypta
JF969276 Paphia amabilis
KP089983 Ruditapes decussatus
MURp - 5.1X
auzay1B - 4.1X
auzay3B - 3.9X
lmc3B - 6.2X
lmc1B - 2.9X
AB065374 Ruditapes philippinarum M
POSp - 18.1X
NEG - 29.0X
LAN1p - 10.1X
LAN2p - 13.6X
LAN3p - 6.7X
MATp - 3.3X
AKKp - 7.8X
AB065375 Ruditapes philippinarum F
KORp - 54.1X
0.3 substitutions per site

**FIGURE 5 |** Maximum likelihood phylogenetic tree of mollusk mitochondrial genes. Tree built from a 60-individual alignment of 12 concatenated gene sequences. Sequences assembled from shell DNA in this study are highlighted in red and their coverage is indicated in the tip label, GenBank accession numbers are provided for previously published sequences. Branch support ≥0.70 is shown as Shimodaira-Hasegawa approximate Likelihood Ratio Test. The subtrees for the *R. decussatus* and *R. philippinarum* clades were enlarged and branches collapsed in order make branching order more apparent.

and tooth/femur (Gilbert et al., 2008; Pečnerová et al., 2017), and, outside Siberia, a >560 kyr horse bone from the Yukon Territories, Canada (Orlando et al., 2013). Although earlier studies demonstrated permafrost as one of the best environments for long-term preservation of animal DNA in keratin- and calcium phosphate-based remains, we show here for the first time that permafrost also preserves calcium carbonate-based remains over significant timescales. Past marine diversity was formerly recovered from <43 kyr sediments that yielded DNA from protists, including haptophytes and foraminifera, which are other organisms producing calcium carbonate exoskeletons surviving or dissolving within sediments (Boere et al., 2009; Coolen et al., 2013; Lejzerowicz et al., 2013; Pawłowska et al., 2014; Morard et al., 2017; More et al., 2018; Armbrecht et al., 2019).

## Recovery of Degraded Ancient Mollusk DNA Molecules From 13 Marine and Freshwater Species

Compared to the previous assessment of eight marine species (Der Sarkissian et al., 2017), we demonstrate here that aDNA can be recovered from an extra five species (*P. arctica*, *H. rufescens*, *H. cracherodii*, *D. polymorpha*, and *L. stagnalis*), including two freshwater species. Considering the relevance of mollusks to ecological questions about the terrestrial niche (Lydeard et al., 2004), aDNA content of terrestrial mollusk shells is worth further investigation. In the summarizing visualization of the potential of each mollusk genus investigated for ecological studies, radar-plot maximal surface areas indicate *Mytilus*, *Haliotis*, and *Arctica* as the three genera with the most potential for ecological studies based on aDNA (**Figure 6**). Overall, we observed a DNA recovery success rate of 61.25% for the 80 shells examined (**Figure 6**), and endogenous contents of 0.09–33.26% for those genera with available reference nuclear genome assemblies (*Crassostrea*, *Dreissena*, *Haliotis*, *Lymnaea*, *Mytilus*), and we estimated that no less than 71,000 sequencing reads were required here to obtain 1× coverage of 16,358–18,653 bp mitochondrial genomes. In line with these results suggesting low numbers of shotgun-sequenceable DNA templates in ancient mollusk shell extracts, we were not successful at recovering DNA from three 42–191 year BP *M. margaritifera* shells, although positive nuclear Small Tandem Repeat (STR) PCR amplification could be achieved from fresh dry shells of the same species by Geist et al. (2008). Our negative results could be explained by the particular storage conditions of the specimens, extensive *post-mortem* DNA degradation and/or concentrations of recoverable

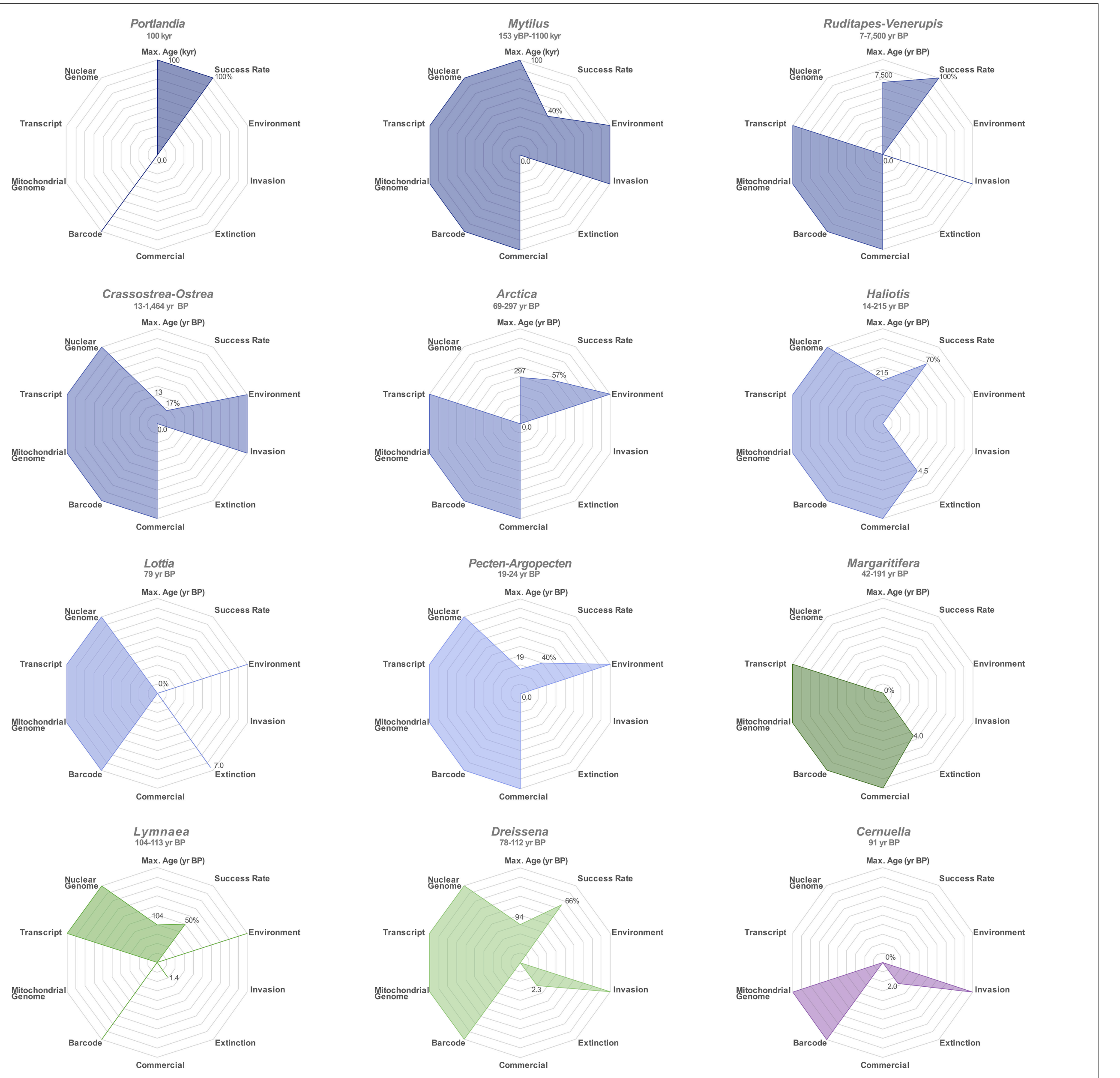


**FIGURE 6 |** Radar plot summarizing application potential of ancient shell DNA for mollusk genera. Marine genera are represented in blue shades, freshwater genera in green shades and the terrestrial genus *Cernuella* in purple. "Max. Age" refers the age of the oldest sample for which DNA could successfully be recovered; "kyr," kilo years; "year BP," years Before Present.

DNA below the sensitivity threshold of our methodology. Target-enrichment protocols may help circumventing these limitations in the future. We, however, caution that such methods will result in the loss of the vast majority of off-target DNA molecules. Their benefits in terms of information gain should be weighed against the risk of heritage loss for each (rare) ancient specimen considered. In addition, the type of capture should carefully be selected and designed to maximize the overlap with datasets generated from comparative modern or ancient populations. Importantly, neither capture nor shotgun HTS aDNA data can, at this stage, provide much information about population dynamics in cases where modern population data available for comparison are nuclear STR genotypes. These indeed typically encompass >100 bp regions (e.g., Knott et al., 2003; Cruz et al., 2005; Gardeström et al., 2008; Feldheim et al., 2011; Van Wormhoudt et al., 2011; Beldade et al., 2012; Morvezen et al., 2013; Borrell et al., 2014; d'Auriac et al., 2017; Jiang et al., 2018) that cannot reliably be typed from aDNA extracts due to their characteristic high fragmentation (here, 41.9–107.8 bp).

## Limits Associated With the Availability of Modern Mollusk DNA Data for Comparative Analyses

We observed a substantial heterogeneity among the examined genera in the availability of comparative mitochondrial sequences, complete mitochondrial genomes, nuclear transcripts and genomes (**Figure 6**). As expected for mostly non-model species, the present study highlights a general paucity of published mitochondrial and nuclear genome assemblies, population-scale data, as well as geographical metadata submitted to barcode databases (BOLD, GenBank), with some mollusk populations better described at the population genomic level, i.e., *Crassostrea* sp., *Mytilus* sp., *Haliotis* sp., and *Pecten* sp. (Zbawicka et al., 2014a, 2018; Fraïsse et al., 2016; Mathiesen et al., 2016; Wenne et al., 2016; Gutierrez et al., 2017; Harney et al., 2018; Wilson et al., 2018; El Ayari et al., 2019; Masonbrink et al., 2019; Paterno et al., 2019; Vendrami et al., 2019a,b). It is worth noting that the phylogenetically robust data reported here significantly increases the number of released complete mitochondrial genome sequences by 6-fold for *R. decussatus*, five-fold for *R. philippinarum*, three-fold for *A. islandica*, and *D. polymorpha*, and finally two-fold for *H. rufescens*, *H. tuberculata*, and *P. maximus*. Although mitochondrial DNA is very informative for genus-level classification and investigating the evolution of organellar genomes, it should be stressed that species-level taxonomic identification and population affinity inference should be confirmed by nuclear sequences for those mollusks in which interspecies hybridization is observed, e.g., *Crassostrea* and *Mytilus*. Importantly, as genomic data production is expected to increase significantly in the near future, it is crucial that appropriate metadata are made easily accessible to the research community, so that published results can be reproduced and future analyses can incorporate ever-growing datasets. Despite limitations associated with the degraded nature of mollusk shell aDNA and the occasional scarceness of comparable modern population data, phylogenetic and population structure signals could nevertheless be recovered, thus emphasizing the promising potential of mollusk shell aDNA for ecological studies.

## Potential of Ancient Mollusk DNA for Ecological Studies

Our results directly hint at taxonomic identification based on mitochondrial barcoding, which could reveal helpful in cases where morphological examination of fragmentary or undiagnostic shell remains preclude classification. Mollusk aDNA analyses are compatible with shotgun- or mitochondrial metabarcoding-based (Bush et al., 2019) reconstructions of past communities from bulk samples (e.g., Murray et al., 2013; Grealy et al., 2015; Seersholm et al., 2018) or sediment cores (e.g., Giguet-Covex et al., 2014; Pedersen et al., 2016), which could be of relevance to the study of shell middens. Temporal changes in ecosystems detected through such approaches could reveal interesting insights into the timing and impact of environmental changes, biological invasions and extinctions.

## Potential of Ancient Mollusk DNA for Studying Environmental Changes

Time series DNA data from mollusk shells hold the potential to illuminate patterns of species or population distribution and diversity. These can be compared to past global and local paleo-environments reconstructed through stable isotope, metal or trace element analyses of the very same shells in order to characterize the temporal and spatial scales of biological responses to climatic events, in particular prior to access to instrumental data or historical records (reviewed in Jones et al., 2009). Marine bivalves such as *A. islandica* and *P. maximus*, for which we reported successful aDNA recovery in this study, have previously allowed sclerochronological reconstructions of past seawater temperature and salinity changes at millennial-length annual and seasonal resolutions, respectively (Surge et al., 2003; Chauvaud et al., 2012; Vokhshoori and McCarthy, 2014; Reynolds et al., 2016; Black et al., 2017). At a more local scale, the freshwater *L. stagnalis*, for which we report shell aDNA recovery here for the first time, is a model commonly used in ecotoxicology (Amorim et al., 2019). Its abundant and widespread distribution in temperate limnic systems and its low dispersal ability can be leveraged to investigate the impact of pollutants in freshwater systems, in particular pesticides, molluscicides, algaecides or industrial (petro-) chemicals pollutants (Amorim et al., 2019). Adding temporally sampled genomic data would allow for controlling the genomic background of *L. stagnalis* population subjected to toxic stress and provide baselines prior to the use of pesticides and chemicals (Coutellec et al., 2013; Bouétard et al., 2014).

Integrating aDNA and environmental information provides new perspectives to better understand mollusk responses to global and local changes. These are expected to be complex, heterogeneous through time, space and across species. Many gaps still remain in our knowledge of the principles and mechanisms driving these responses; for example, the relative contribution of migration (routes and speed) and *in situ* tolerance in the form of phenotypic plasticity (morphology, behavior, ecophysiology, reproductive strategies) or evolutionary adaptation (speciation, selection) to the survival of species facing environmental changes (e.g., climate, pollution) of varying magnitude and rate.

## Potential of Ancient Mollusk DNA for Studying Species of Commercial Interest

*Mytilus*, *Haliotis, Crassostrea,* and *Ruditapes* species represented 73% (8,816,367 tons) of the world marine mollusk production in 2017 (FAO, 2017). Their production can be managed in large-scale hatcheries and has resulted in massive translocation across large geographical scales. For example, *Haliotis* were commercially fished in Europe in the 19th century, which led to drastic wild stock depletions and the implementation of management plans and regulations for fishing and cultivation (Huchette and Clavier, 2004). European *Crassostrea* has had a distinct history, as the Portuguese oyster *C. angulata* was first introduced unintentionally from Asia, probably in the 16th century, and cultivated until it underwent high mortalities associated with iridovirus outbreaks in the 1960–1970s (Grizel and Héral, 1991). It was then substituted by the more resistant Pacific cupped oyster sister species *C. gigas* with distributions in Asia overlapping those of native *C. angulata* (Grizel and Héral, 1991). *Haliotis* and *Crassostrea* in North America have similar histories of overfishing and stock depletion across the 19 and 20th centuries (Braje et al., 2009, 2016; Rick et al., 2016). As for the Manila clam *R. philippinarum*, it has a natural distribution on the West coast of the Pacific from the Philippines to Russia. According to historical records, *R. philippinarum* was first unintentionally introduced from Japan to North America through oyster "hitchhiking" in 1936 C.E. (Quayle, 1964), and then imported from there into Europe partly to overcome the over-exploitation, irregular yields, recruitment failures, and outbreak of bacterial infections in the endemic *R. decussatus* in 1972–1974 C.E. (Flassch and Leborgne, 1992; Sanna et al., 2017). Human-driven displacements, intense harvest, over-exploitation, and controlled reproduction in hatchery are expected to have impacted exploited mollusk species, in particular with regards to genomic patterns of diversity (e.g., inbreeding) and/or admixture (Astorga, 2014), as well as allele frequencies for loci associated with economically important phenotypic traits (e.g., survival, growth rate, body weight, and yield) (Gutierrez et al., 2018). These could be tracked through time using the same mollusk shell aDNA approach as applied here, which could also potentially reveal the existence of undocumented transfers.

Additionally, as human-mediated transfers may subject farmed mollusks to new pathogens, mollusk shell aDNA could also help track and monitor the spread of infections. For example, in *R. philippinarum*, high mortalities were observed in 1987 C.E. in Brittany from the so-called "Brown Ring Disease" (BRD) for which the bacteria *Vibrio tapetis* was identified as the etiological agent (Paillard et al., 1989, 2008; Borrego et al., 1996; Allam et al., 2000). DNA recovered from the shells of *R. philippinarum* showing substantial brown conchiolin deposits diagnostic of acute BRD (shells labeled as POSp and KORp in the present study; **Supplementary Figure S1**) was previously found in Der Sarkissian et al., 2017 to show highest affinity for the virulent *V. tapetis* strain RP2-3, whereas DNA from shells displaying weak BRD infection (LAN1p, LAN2p, and LAN3p) showed highest affinity for the less virulent *V. tapetis* strain HH6087 (Der Sarkissian et al., 2017; Dias et al., 2018). Here, phylogenetic analyses of the *R. philippinarum* COI mitochondrial gene in specimens from Landéda, Brittany, reveal a segregation between the 1988 C.E. acute BRD specimen POSp falling within the mainly Asian diversity (China, Japan), and the 1983–1988 C.E. asymptomatic or weak BRD specimens NEG, LAN1p, LAN2p, and LAN3p closely related to mainly European and American individuals (**Supplementary Figure S1**). These results are in line with a tentative scenario, in which lower virulence *V. tapetis* strains may have already been present in Brittany before the 1987 C.E. virulent BRD outbreak, which might have been triggered by a second introduction of *R. philippinarum* stocks, possibly of more recent Asian origin, containing highly virulent strains. The presence of the 2003 C.E. KORp shell in Korea showing acute BRD and a COI sequence closely related to European and American clams could possibly reflect a later transfer from Europe to Korea, where BRD was first identified by molecular methods in 2006 C.E. (Park et al., 2006), and where stocks have underwent severe reductions due to overexploitation and coastal pollution (Cordero et al., 2017).

## Potential of Ancient Mollusk DNA for Studying Biological Invasions

In a context of global trade, human activities greatly facilitate rapid species displacements over large geographical distances, whether it be intentionally (e.g., for cultivation in hatcheries or direct commercial purposes) or unintentionally (e.g., through waterways, ship and seaplane hull fouling, ballast waters, fishing) (Carlton, 1999). Thanks to a competitive advantage (e.g., long larval phases, rapid growth, early sexual maturity, high fecundity, broad salinity, and temperature tolerance), newly introduced species can proliferate from a point of introduction into new ecosystems in which they become dominant as invasive species (Valéry et al., 2008). Biological invasions are recognized as the second most important threat to biodiversity after habitat destruction due to their substantial negative ecological (e.g., food web alteration and extirpation of native species) and economic impacts [e.g., biofouling, blocking of water pipes (McCarthy et al., 1997)].

Belonging to the mollusk genera examined here, *C. gigas* and *R. philippinarum* threaten the European native species of *Ostrea edulis* and *R. decussatus*, respectively, and *M. galloprovincialis* has been listed as one of the "100 world's worst invasive alien species" (Lowe et al., 2004). *D. polymorpha* is another well-identified invasive species for which shell aDNA could be retrieved. Today, *D. polymorpha* has a worldwide distribution and its invasive status is a recognized threat (Lowe et al., 2004). It initially spread from the Ponto-Caspian Basin to Europe in the mid 1800s (or before), and then reached North America from Europe in the mid 1980s (Kinzelbach, 1986; Carlton, 1999). Although it is one of the most studied invasive species with regards to ecological, toxicological and physiological aspects, population genomic data are currently scarce. Early detection of invasive species is of foremost importance in efforts to eradicate them, or at least slow their spread down (Williams et al., 2017). Designing management plans for the control of invasions can greatly benefit from the reconstruction of their history: geographical origin, number, and timing of the introduction(s), as well as dispersal and connectivity patterns. This can be informed by phylogeography in both the native and invaded areas. The timing of invasions, however, can be difficult to estimate with precision. One reason for this is that some mollusk (cryptic) invasive and native species show similar shell morphology due to their close relationships and phenotypic plasticity (Morais and Reichard, 2018). Some of the local species may also be impacted by hybridization with invasive sister species in semi-reproductive isolation (Harrison and Larson, 2014), which is facilitated by high fecundity and dispersal potential in broadcast-spawning aquatic mollusks *Dreissena*, *Crassostrea*, and *Mytilus* species (Voroshilova et al., 2010; Fraïsse et al., 2016; Gagnaire et al., 2018). Mollusk shell DNA is a useful tool to assess the presence and genomic landscape of invaders at given points in time and space. Considering the dynamic and stochastic nature of biological invasions, shell aDNA could help evaluate their impact on native populations and ecosystems, thus showing potential for future studies in ecology and conservation.

### Potential of Ancient Mollusk DNA for Studying Extinctions

A total of 1,526 species of gastropods and bivalves are classified as threatened by The IUCN Red List of Threatened Species (2019). Among them, two species of *Haliotis* abalones of the North American Pacific coast have been classed as "endangered," i.e., *H. kamtschatkana*, and "critically endangered," i.e., *H. cracherodii* or black abalone. The latter was shown here to be amenable to aDNA analyses, similarly to another Pacific abalone species, *H. rufescens* or red abalone. Both species were found in shell middens and in artifacts at prehistoric Native American sites (Erlandson and Rick, 2008; Braje et al., 2009). *H. cracherodii* experienced critical population declines following intensive fishing in the 1850-1900s and then again after the peak in abalone farming of the 1950–1960s, which severely affected all abalone populations in the Pacific, including *H. rufescens* (Rogers-Bennett et al., 2002). Abalone population declines were aggravated by the recovery of sea otters (*Enhydra lutris nereis and E. l. kenyoni*) in the 20th century, abalone predators that had nearly gone extinct following overexploitation by the historical fur trade (Lee et al., 2016). High mortalities due to infections by the bacterium *Candidatus xenohaliotis* responsible for the Withering syndrome devastated abalone populations (Crosson et al., 2014) and contributed to the low genetic diversities and population differentiations observed in *H. cracherodii* and *H. rufescens* today (Gaffney et al., 1996; Burton and Tegner, 2000; Hamm and Burton, 2000). Conservation efforts are focused in restoring and protecting abalones in the Pacific: shore picking of *H. rufescens* is allowed only under strict rules, and fishing of *H. cracherodii* (commercial or recreational) have been suspended (Haas et al., 2019). Management and conservation plans would greatly benefit from shell aDNA analyses that could help better understand the dynamics of genomic diversity and distribution for the two species, the effects of exploitation, infection and predator recovery, and complement archeological studies aimed at helping in abalone restoration (Braje et al., 2009, 2016; Hofman et al., 2015).

## CONCLUSION

In this study, we illustrate the potential of ancient mollusk shell mitochondrial and nuclear DNA analyses using geological, archeological and museum specimens from 13 species, spanning ≥100 kyr. Ancient mollusk DNA is of growing relevance for ecological studies as present-day societies are increasingly concerned by on-going global environmental changes (e.g., warming, pollution, sea level rise, ocean acidification, eutrophication) and associated biodiversity loss (Cardinale et al., 2012) continue to increase at an alarming pace. Some environmental stressors are expected to influence directly shell-producing mollusks in their ability to calcify, and hence, their integrity and vulnerability to predators (Kroeker et al., 2013), with some of the affected species playing the role of keystone species with large impacts on ecosystems. Our work highlights the importance of mollusk shell aDNA as a powerful and novel tool for ecological studies.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the European Nucleotide Archive public database (project PRJEB35671).

## ETHICS STATEMENT

Ethical review and approval was not required for this study because no living animals were collected or examined.

## AUTHOR CONTRIBUTIONS

CD and LO conceived and coordinated the study. CD performed molecular and computational analyses, with input from LO. PM, CH, PI, TR, MS, TS, LD, and LO provided samples and reagents. CD wrote the article with input from PM, LD, and LO.

## FUNDING

This work was supported by the Danish Council for Independent Research, the Natural Sciences (FNU, 4002-00152B), the Danish National Research Foundation (DNRF94), and the Initiative d'Excellence Chaires d'attractivité, Université de Toulouse (OURASI). LO has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programe (Grant agreement no. 681605-PEGASUS). PM and LD received funding from the Swedish Natural Science Research Council (VR) (Grant agreements nos. 621-2008-3759 and 621-2008-4896, respectively), and expedition logistics, including for the Taimyr 2010 expedition, from the Swedish Polar Research Secretariat (SPRS).

## ACKNOWLEDGMENTS

We thank Martin Sikora at The GLOBE Institute, University of Copenhagen, for transporting samples. We also thank the PALEOMIX group at the University of Copenhagen and the AGES group at the Laboratory AMIS UMR 5288 (CNRS, University Toulouse 3), as well as the staff of the Danish National High-Throughput DNA Sequencing Centre and of AMIS for support.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.00037/full#supplementary-material

## REFERENCES

Allam, B., Paillard, C., Howard, A., and Le Pennec, M. (2000). Isolation of the pathogen *Vibrio tapetis* and defense parameters in brown ring diseased Manila clams *Ruditapes philippinarum* cultivated in England. *Dis. Aquat. Organ.* 41, 105–113. doi: 10.3354/dao041105

Amorim, J., Abreu, I., Rodrigues, P., Peixoto, D., Pinheiro, C., Saraiva, A., et al. (2019). *Lymnaea stagnalis* as a freshwater model invertebrate for ecotoxicological studies. *Sci. Total Environ.* 669, 11–28. doi: 10.1016/j.scitotenv.2019.03.035

Armbrecht, L. H., Coolen, M. J. L., Lejzerowicz, F., George, S. C., Negandhi, K., Suzuki, Y., et al. (2019). Ancient DNA from marine sediments: precautions and considerations for seafloor coring, sample handling and data generation. *Earth Sci. Rev.* 196:102887. doi: 10.1016/j.earscirev.2019.102887

Astorga, M. P. (2014). Genetic considerations for mollusk production in aquaculture: current state of knowledge. *Front. Genet.* 5:435. doi: 10.3389/fgene.2014.00435

Beldade, R., Bell, C. A., Raimondi, P. T., George, M. K., Miner, C. M., and Bernardi, G. (2012). Isolation and characterization of 8 novel microsatellites for the black abalone, *Haliotis cracherodii*, a marine gastropod decimated by the withering disease. *Conserv. Genet. Resour.* 4, 1071–1073. doi: 10.1007/s12686-012-9709-3

Black, H. D., Andrus, C. F. T., Lambert, W. J., Rick, T. C., and Gillikin, D. P. (2017). $\delta^{15}$N values in *Crassostrea virginica* shells provides early direct evidence for nitrogen loading to Chesapeake bay. *Sci. Rep.* 7:44241. doi: 10.1038/srep44241

Boere, A. C., Abbas, B., Rijpstra, W. I. C., Versteegh, G. J. M., Volkman, J. K., Damsté, J. S. S., et al. (2009). Late-Holocene succession of dinoflagellates in an Antarctic fjord using a multi-proxy approach: paleoenvironmental genomics, lipid biomarkers and palynomorphs. *Geobiology* 7, 265–281. doi: 10.1111/j.1472-4669.2009.00202.x

Borrego, J. J., Castro, D., Luque, A., Paillard, C., Maes, P., Garcia, M. T., et al. (1996). *Vibrio tapetis* sp. nov., the causative agent of the brown ring disease affecting cultured clams. *Int. J. Syst. Evol. Microbiol.* 46, 480–484. doi: 10.1099/00207713-46-2-480

Borrell, Y. J., Arias-Pérez, A., Freire, R., Valdés, A., Sánchez, J. A., Méndez, J., et al. (2014). Microsatellites and multiplex PCRs for assessing aquaculture practices of the grooved carpet shell *Ruditapes decussatus* in Spain. *Aquaculture* 42, 49–59. doi: 10.1016/j.aquaculture.2014.01.010

Bouétard, A., Côte, J., Besnard, A.-L., Collinet, M., and Coutellec, M.-A. (2014). Environmental versus anthropogenic effects on population adaptive divergence in the freshwater snail *Lymnaea stagnalis*. *PLoS One* 9:e106670. doi: 10.1371/journal.pone.0106670

Braje, T. J., Erlandson, J. M., Rick, T. C., Dayton, P. K., and Hatch, M. B. A. (2009). Fishing from past to present: continuity and resilience of red abalone fisheries on the Channel Islands, California. *Ecol. Appl.* 19, 906–919. doi: 10.1890/08-0135.1

Braje, T. J., Rick, T. C., Erlandson, J. M., Rogers-Bennett, L., and Catton, C. A. (2016). Historical ecology can inform restoration site selection: the case of black abalone (*Haliotis cracherodii*) along California's Channel Islands. *Aquat. Conserv. Mar. Freshw. Ecosyst* 26, 470–481. doi: 10.1002/aqc.2561

Breton, S., Burger, G., Stewart, D. T., and Blier, P. U. (2006). Comparative analysis of gender-associated complete mitochondrial genomes in marine mussels (*Mytilus* spp.). *Genetics* 172, 1107–1119. doi: 10.1534/genetics.105.047159

Burton, R. S., and Tegner, M. J. (2000). Enhancement of red abalone *Haliotis rufescens* stocks at San Miguel Island: reassessing a success story. *Mar. Ecol. Prog. Ser.* 202, 303–308. doi: 10.3354/meps202303

Bush, A., Compson, Z. G., Monk, W. A., Porter, T. M., Steeves, R., Emilson, E., et al. (2019). Studying ecosystems with DNA metabarcoding: lessons from biomonitoring of aquatic macroinvertebrates. *Front. Ecol. Evol.* 7:434. doi: 10.3389/fevo.2019.00434

Butler, P. G., Freitas, P. S., Burchell, M., and Chauvaud, L. (2019). "Archaeology and sclerochronology of marine bivalves," in *Goods and Services of Marine Bivalves*, eds A. C. Smaal, J. G. Ferreira, J. Grant, J. K. Petersen, and Ø. Strand, (Cham: Springer), 413–444. doi: 10.1007/978-3-319-96776-9_21

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., et al. (2012). Biodiversity loss and its impact on humanity. *Nature* 486, 59–67. doi: 10.1038/nature11148

Carlton, J. (1999). Molluscan invasions in marine and estuarine communities. *Malacologia* 41, 439–454.

Chauvaud, L., Lorrain, A., Dunbar, R. B., Paulet, Y.-M., Thouzeau, G., Jean, F., et al. (2005). Shell of the great scallop *Pecten maximus* as a high-frequency archive of paleoenvironmental changes. *Geochem. Geophys. Geosyst.* 6:Q08001. doi: 10.1029/2004GC000890

Chauvaud, L., Patry, Y., Jolivet, A., Cam, E., Goff, C. L., Strand, Ø., et al. (2012). Variation in size and growth of the great scallop *Pecten maximus* along a latitudinal gradient. *PLoS One* 7:e37717. doi: 10.1371/journal.pone.0037717

Coolen, M. J. L., Orsi, W. D., Balkema, C., Quince, C., Harris, K., Sylva, S. P., et al. (2013). Evolution of the plankton paleome in the Black Sea from the deglacial to anthropocene. *Proc. Natl. Acad. Sci. U.S.A.* 110, 8609–8614. doi: 10.1073/pnas.1219283110

Cordero, D., Delgado, M., Liu, B., Ruesink, J., and Saavedra, C. (2017). Population genetics of the Manila clam (*Ruditapes philippinarum*) introduced in North America and Europe. *Sci. Rep.* 7:39745. doi: 10.1038/srep39745

Cordero, D., Peña, J. B., and Saavedra, C. (2014). Phylogeographic analysis of introns and mitochondrial DNA in the clam *Ruditapes decussatus* uncovers the effects of Pleistocene glaciations and endogenous barriers to gene flow. *Mol. Phylogenet. Evol.* 71, 274–287. doi: 10.1016/j.ympev.2013.11.003

Coutellec, M.-A. (2017). Mollusc shells as metagenomic archives: The true treasure is the chest itself. *Mol. Ecol. Resour.* 17, 854–857. doi: 10.1111/1755-0998.12716

Coutellec, M.-A., Besnard, A.-L., and Caquet, T. (2013). Population genetics of *Lymnaea stagnalis* experimentally exposed to cocktails of pesticides. *Ecotoxicology* 22, 879–888. doi: 10.1007/s10646-013-1082-9

Crosson, L. M., Wight, N., VanBlaricom, G. R., Kiryu, I., Moore, J. D., and Friedman, C. S. (2014). Abalone withering syndrome: distribution, impacts, current diagnostic methods and new findings. *Dis. Aquat. Organ.* 108, 261–270. doi: 10.3354/dao02713

Cruz, P., Ibarra, A. M., Fiore-Amaral, G., Galindo-Sánchez, C. E., and Mendoza-Carrión, G. (2005). Isolation of microsatellite loci in green abalone (*Haliotis fulgens*) and cross-species amplification in two other North American red (*Haliotis rufescens*) and pink (*Haliotis corrugata*) abalones. *Mol. Ecol. Notes* 5, 857–859. doi: 10.1111/j.1471-8286.2005.01088.x

d'Auriac, M. B. A., Rinde, E., Norling, P., Lapègue, S., Staalstrøm, A., Hjermann, Ø., et al. (2017). Rapid expansion of the invasive oyster *Crassostrea gigas* at its northern distribution limit in Europe: Naturally dispersed or introduced? *PLoS One* 12:e0177481. doi: 10.1371/journal.pone.0177481

Der Sarkissian, C., Allentoft, M. E., Ávila-Arcos, M. C., Barnett, R., Campos, P. F., Cappellini, E., et al. (2015). Ancient genomics. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 370:20130387. doi: 10.1098/rstb.2013.0387

Der Sarkissian, C., Pichereau, V., Dupont, C., Ilsøe, P. C., Perrigault, M., Butler, P., et al. (2017). Ancient DNA analysis identifies marine mollusc shells as new metagenomic archives of the past. *Mol. Ecol. Resour.* 17, 835–853. doi: 10.1111/1755-0998.12679

Dias, G. M., Bidault, A., Le Chevalier, P., Choquet, G., Der Sarkissian, C., Orlando, L., et al. (2018). *Vibrio tapetis* displays an original type IV secretion system in strains pathogenic for bivalve molluscs. *Front. Microbiol.* 9:227. doi: 10.3389/fmicb.2018.00227

El Ayari, T., Trigui El Menif, N., Hamer, B., Cahill, A. E., and Bierne, N. (2019). The hidden side of a major marine biogeographic boundary: a wide mosaic hybrid zone at the Atlantic-Mediterranean divide reveals the complex interaction between natural and genetic barriers in mussels. *Heredity* 122, 770–784. doi: 10.1038/s41437-018-0174-y

Erlandson, J. M., and Rick, T. C. (2008). "Archaeology, marine ecology, and human impacts on marine environments," in *Human Impacts on Ancient Marine Ecosystems: A Global Perspective*, eds T. C. Rick, and J. M. Erlandson, (Berkeley, CA: University of California Press), 1–19.

Fages, A., Hanghøj, K., Khan, N., Gaunitz, C., Seguin-Orlando, A., Leonardi, M., et al. (2019). Tracking five millennia of horse management with extensive ancient genome time series. *Cell* 177, 1419–1435.e31. doi: 10.1016/j.cell.2019.03.049

FAO (2017). *Fishery and Aquaculture Statistics. Aquacultre Production*. Available at: http://www.fao.org/fishery/static/Yearbook/YB2017_USBcard/root/aquaculture/yearbook_aquaculture.pdf (accessed February 17, 2020).

Feldheim, K. A., Brown, J. E., Murphy, D. J., and Stepien, C. A. (2011). Microsatellite loci for dreissenid mussels (Mollusca: Bivalvia: Dreissenidae) and relatives: markers for assessing exotic and native populations. *Mol. Ecol. Resour.* 11, 725–732. doi: 10.1111/j.1755-0998.2011.03012.x

Flassch, J.-P., and Leborgne, Y. (1992). ). Introduction in Europe, from 1972 to 1980, of the Japanese Manila clam (*Tapes philippinarum*) and the effects on aquaculture production and natural settlement. *ICES Mar. Sci. Symp.* 194, 92–96.

Fortunato, H. (2015). Mollusks: tools in environmental and climate research. *Am. Malacol. Bull.* 33, 310–324. doi: 10.4003/006.033.0208

Fraïsse, C., Belkhir, K., Welch, J. J., and Bierne, N. (2016). Local interspecies introgression is the main cause of extreme levels of intraspecific differentiation in mussels. *Mol. Ecol.* 25, 269–286. doi: 10.1111/mec.13299

Fraïsse, C., Haguenauer, A., Gérard, K., Weber, A. A.-T., Bierne, N., and Chenuil, A. (2017). Fine-grained genetic-environment association in an admixed population of mussels in the small isolated Kerguelen island. *bioRxiv* [Preprint]. doi: 10.1101/239244

Gaffney, P. M., Rubin, V. P., Hedgecock, D., Powers, D. A., Morris, G., and Hereford, L. (1996). Genetic effects of artificial propagation: signals from wild and hatchery populations of red abalone in California. *Aquaculture* 143, 257–266. doi: 10.1016/0044-8486(96)01278-1

Gagnaire, P.-A., Lamy, J.-B., Cornette, F., Heurtebise, S., Dégremont, L., Flahauw, E., et al. (2018). Analysis of genome-wide differentiation between native and introduced populations of the cupped oysters *Crassostrea gigas* and *Crassostrea angulata*. *Genome Biol. Evol.* 10, 2518–2534. doi: 10.1093/gbe/evy194

Gamba, C., Hanghøj, K., Gaunitz, C., Alfarhan, A. H., Alquraishi, S. A., Al-Rasheid, K. A. S., et al. (2016). Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. *Mol. Ecol. Resour.* 16, 459–469. doi: 10.1111/1755-0998.12470

Gamba, C., Jones, E. R., Teasdale, M. D., McLaughlin, R. L., Gonzalez-Fortes, G., Mattiangeli, V., et al. (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* 5:5257. doi: 10.1038/ncomms6257

Gardeström, J., Pereyra, R. T., and André, C. (2008). Characterization of six microsatellite loci in the Baltic blue mussel *Mytilus trossulus* and cross-species amplification in North Sea *Mytilus edulis*. *Conserv. Genet.* 9, 1003–1005. doi: 10.1007/s10592-007-9432-x

Geist, J., Wunderlich, H., and Kuehn, R. (2008). Use of mollusc shells for DNA-based molecular analyses. *J. Molluscan Stud.* 74, 337–343. doi: 10.1093/mollus/eyn025

Ghiselli, F., Milani, L., Iannello, M., Procopio, E., Chang, P. L., Nuzhdin, S. V., et al. (2017). The complete mitochondrial genome of the grooved carpet shell, *Ruditapes decussatus* (Bivalvia, Veneridae). *PeerJ* 5:e3692. doi: 10.7717/peerj.3692

Giguet-Covex, C., Pansu, J., Arnaud, F., Rey, P.-J., Griggo, C., Gielly, L., et al. (2014). Long livestock farming history and human landscape shaping revealed by lake sediment DNA. *Nat. Commun.* 5:3211. doi: 10.1038/ncomms4211

Gilbert, M. T. P., Drautz, D. I., Lesk, A. M., Ho, S. Y. W., Qi, J., Ratan, A., et al. (2008). Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes. *Proc. Natl. Acad. Sci. U.S.A.* 105, 8327–8332. doi: 10.1073/pnas.0802315105

Glöckner, G., Heinze, I., Platzer, M., Held, C., and Abele, D. (2013). The mitochondrial genome of Arctica islandica; phylogeny and variation. *PLoS One* 8:e82857. doi: 10.1371/journal.pone.0082857

Grade, A., Chairi, H., Lallias, D., Power, D. M., Ruano, F., Leitão, A., et al. (2016). New insights about the introduction of the *Portuguese oyster*, *Crassostrea angulata*, into the North East Atlantic from Asia based on a highly polymorphic mitochondrial region. *Aquat. Living Resour.* 29:404. doi: 10.1051/alr/2016035

Grealy, A. C., McDowell, M. C., Scofield, P., Murray, D. C., Fusco, D. A., Haile, J., et al. (2015). A critical evaluation of how ancient DNA bulk bone metabarcoding complements traditional morphological analysis of fossil assemblages. *Quat. Sci. Rev.* 128, 37–47. doi: 10.1016/j.quascirev.2015.09.014

Green, E. J., and Speller, C. F. (2017). Novel substrates as sources of ancient DNA: prospects and hurdles. *Genes* 8:180. doi: 10.3390/genes8070180

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., et al. (2010). A draft sequence of the Neandertal genome. *Science* 328, 710–722. doi: 10.1126/science.1188021

Grizel, H., and Héral, M. (1991). Introduction into France of the Japanese oyster (*Crassostrea gigas*). *ICES J. Mar. Sci.* 47, 399–403. doi: 10.1093/icesjms/47.3.399

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Gutierrez, A. P., Matika, O., Bean, T. P., and Houston, R. D. (2018). Genomic selection for growth traits in pacific oyster (*Crassostrea gigas*): potential of low-density marker panels for breeding value prediction. *Front. Genet.* 9:391. doi: 10.3389/fgene.2018.00391

Gutierrez, A. P., Turner, F., Gharbi, K., Talbot, R., Lowe, N. R., Peñaloza, C., et al. (2017). Development of a medium density combined-species SNP array for pacific and European oysters (*Crassostrea gigas* and *Ostrea edulis*). *G3* 7, 2209–2218. doi: 10.1534/g3.117.041780

Haas, H., Braje, T. J., Edwards, M. S., Erlandson, J. M., and Whitaker, S. G. (2019). Black abalone (*Haliotis cracherodii*) population structure shifts through deep time: management implications for southern California's northern Channel Islands. *Ecol. Evol.* 9, 4720–4732. doi: 10.1002/ece3.5075

Hamm, D. E., and Burton, R. S. (2000). Population genetics of black abalone, *Haliotis cracherodii*, along the central California coast. *J. Exp. Mar. Biol. Ecol.* 254, 235–247. doi: 10.1016/S0022-0981(00)00283-5

Harney, E., Lachambre, S., Roussel, S., Huchette, S., Enez, F., Morvezen, R., et al. (2018). Transcriptome based SNP discovery and validation for parentage assignment in hatchery progeny of the European abalone *Haliotis tuberculata*. *Aquaculture* 491, 105–113. doi: 10.1016/j.aquaculture.2018.03.006

Harrison, R. G., and Larson, E. L. (2014). Hybridization, introgression, and the nature of species boundaries. *J. Hered.* 105, 795–809. doi: 10.1093/jhered/esu033

Hiebenthal, C., Philipp, E. E. R., Eisenhauer, A., and Wahl, M. (2012). Interactive effects of temperature and salinity on shell formation and general condition in Baltic Sea *Mytilus edulis* and Arctica islandica. *Aquat. Biol.* 14, 289–298. doi: 10.3354/ab00405

Hofman, C. A., Rick, T. C., Fleischer, R. C., and Maldonado, J. E. (2015). Conservation archaeogenomics: ancient DNA and biodiversity in the anthropocene. *Trends Ecol. Evol.* 30, 540–549. doi: 10.1016/j.tree.2015.06.008

Huchette, S. M. H., and Clavier, J. (2004). Status of the ormer (*Haliotis tuberculata* L.) industry in Europe. *J. Shellfish Res.* 23, 951–956.

Innes, D. J., and Bates, J. A. (1999). Morphological variation of *Mytilus edulis* and *Mytilus trossulus* in eastern Newfoundland. *Mar. Biol.* 133, 691–699. doi: 10.1007/s002270050510

Jackson, J. B. C. (2001). What was natural in the coastal oceans? *Proc. Natl. Acad. Sci. U.S.A.* 98, 5411–5418. doi: 10.1073/pnas.091092898

Jiang, L., Nie, H., Li, C., Li, D., Huo, Z., and Yan, X. (2018). The genetic diversity of wild and cultivated Manila clam (*Ruditapes philippinarum*) revealed by 29 novel microsatellite markers. *Electron. J. Biotechnol.* 34, 17–21. doi: 10.1016/j.ejbt.2018.05.003

Jones, P. D., Briffa, K. R., Osborn, T. J., Lough, J. M., van Ommen, T. D., Vinther, B. M., et al. (2009). High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects. *The Holocene* 19, 3–49. doi: 10.1177/0959683608098952

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684. doi: 10.1093/bioinformatics/btt193

Kinzelbach, R. (1986). The recent distribution of the Zebra Mussel, *Dreissena polymorpha*, in the Aegean Region and in Anatolia. *Zool. Middle East* 1, 132–138. doi: 10.1080/09397140.1986.10637537

Knott, K. E., Puurtinen, M., and Kaitala, V. (2003). Primers for nine microsatellite loci in the hermaphroditic snail *Lymnaea stagnalis*. *Mol. Ecol. Notes* 3, 333–335. doi: 10.1046/j.1471-8286.2003.00444.x

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* 15:356. doi: 10.1186/s12859-014-0356-4

Kroeker, K. J., Kordas, R. L., Crim, R., Hendriks, I. E., Ramajo, L., Singh, G. S., et al. (2013). Impacts of ocean acidification on marine organisms: quantifying sensitivities and interaction with warming. *Glob. Change Biol.* 19, 1884–1896. doi: 10.1111/gcb.12179

Lartaud, F., Emmanuel, L., de Rafaelis, M., Pouvreau, S., and Renard, M. (2010). Influence of food supply on the $\delta^{13}$C signature of mollusc shells: implications for palaeoenvironmental reconstitutions. *Geomar. Lett.* 30, 23–34. doi: 10.1007/s00367-009-0148-4

Lee, L. C., Watson, J. C., Trebilco, R., and Salomon, A. K. (2016). Indirect effects and prey behavior mediate interactions between an endangered prey and recovering predator. *Ecosphere* 7:e01604. doi: 10.1002/ecs2.1604

Lefort, V., Longueville, J.-E., and Gascuel, O. (2017). SMS: smart model selection in PhyML. *Mol. Biol. Evol.* 34, 2422–2424. doi: 10.1093/molbev/msx149

Leigh, J. W., and Bryant, D. (2015). popart: full-feature software for haplotype network construction. *Methods Ecol. Evol.* 6, 1110–1116. doi: 10.1111/2041-210X.12410

Lejzerowicz, F., Esling, P., Majewski, W., Szczuciński, W., Decelle, J., Obadia, C., et al. (2013). Ancient DNA complements microfossil record in deep-sea subsurface sediments. *Biol. Lett.* 9:20130283. doi: 10.1098/rsbl.2013.0283

Leonardi, M., Librado, P., Der Sarkissian, C., Schubert, M., Alfarhan, A. H., Alquraishi, S. A., et al. (2016). Evolutionary patterns and processes: lessons from ancient DNA. *Syst. Biol.* 66, e1–e29. doi: 10.1093/sysbio/syw059

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Liehr, G. A., Zettler, M. L., Leipe, T., and Witt, G. (2005). The ocean quahog *Arctica islandica* L.: a bioindicator for contaminated sediments. *Mar. Biol.* 147, 671–679. doi: 10.1007/s00227-005-1612-y

Lorenzen, E. D., Nogués-Bravo, D., Orlando, L., Weinstock, J., Binladen, J., Marske, K. A., et al. (2011). Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature* 479, 359–364. doi: 10.1038/nature10574

Lowe, S., Browne, M., Boudjelas, S., De Poorter, M., and The Invasive Species Specialist Group, (2004). *100 of the World's Worst Invasive Alien Species: A Selection from the Global Invasive Species Database*. The Invasive Species Specialist Group.

Löytynoja, A., and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. U.S.A.* 102, 10557–10562. doi: 10.1073/pnas.0409137102

Lydeard, C., Cowie, R. H., Ponder, W. F., Bogan, A. E., Bouchet, P., Clark, S. A., et al. (2004). The global decline of nonmarine mollusks. *Bioscience* 54, 321–330.

Mann, M. E., Bradley, R. S., and Hughes, M. K. (1999). Northern hemisphere temperatures during the past millennium: inferences, uncertainties, and limitations. *Geophys. Res. Lett.* 26, 759–762. doi: 10.1029/1999GL900070

Mann, M. E., Zhang, Z., Rutherford, S., Bradley, R. S., Hughes, M. K., Shindell, D., et al. (2009). Global signatures and dynamical origins of the little ice age and medieval climate anomaly. *Science* 326, 1256–1260. doi: 10.1126/science.1177303

Marcott, S. A., Shakun, J. D., Clark, P. U., and Mix, A. C. (2013). A reconstruction of regional and global temperature for the past 11,300 years. *Science* 339, 1198–1201. doi: 10.1126/science.1228026

Masonbrink, R. E., Purcell, C. M., Boles, S. E., Whitehead, A., Hyde, J. R., Seetharam, A. S., et al. (2019). An Annotated Genome for *Haliotis rufescens* (Red Abalone) and resequenced green, pink, pinto, black, and white abalone species. *Genome Biol. Evol.* 11, 431–438. doi: 10.1093/gbe/evz006

Mathiesen, S. S., Thyrring, J., Hemmer-Hansen, J., Berge, J., Sukhotin, A., Leopold, P., et al. (2016). Genetic diversity and connectivity within *Mytilus* spp. in the subarctic and Arctic. *Evol. Appl.* 10, 39–55. doi: 10.1111/eva.12415

McCarthy, T. K., Fitzgerald, J., and O'Connor, W. (1997). The occurrence of the Zebra mussel *Dreissena polymorpha* (Pallas, 1771), an introduced biofouling freshwater bivalve in Ireland. *Ir. Nat. J.* 25, 413–416.

Meisner, J., and Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics* 210, 719–731. doi: 10.1534/genetics.118.301336

Möller, P., Alexanderson, H., Funder, S., and Hjort, C. (2015). The Taimyr Peninsula and the Severnaya Zemlya archipelago, Arctic Russia: a synthesis of glacial history and palaeo-environmental change during the Last Glacial cycle (MIS 5e–2). *Quat. Sci. Rev.* 107, 149–181. doi: 10.1016/j.quascirev.2014.10.018

Möller, P., Benediktsson, Í. Ö., Anjar, J., Bennike, O., Bernhardson, M., Funder, S., et al. (2019a). Data set on sedimentology, palaeoecology and chronology of middle to late pleistocene deposits on the Taimyr Peninsula, Arctic Russia. *Data Brief* 25:104267. doi: 10.1016/j.dib.2019.104267

Möller, P., Benediktsson, Í. Ö., Anjar, J., Bennike, O., Bernhardson, M., Funder, S., et al. (2019b). Glacial history and palaeo-environmental change of southern Taimyr Peninsula, Arctic Russia, during the Middle and Late Pleistocene. *Earth Sci. Rev.* 196:102832. doi: 10.1016/j.earscirev.2019.04.004

Morais, P., and Reichard, M. (2018). Cryptic invasions: a review. *Sci. Total Environ.* 613–614, 1438–1448. doi: 10.1016/j.scitotenv.2017.06.133

Morard, R., Lejzerowicz, F., Darling, K. F., Lecroq-Bennet, B., Winther Pedersen, M., Orlando, L., et al. (2017). Planktonic foraminifera-derived environmental DNA extracted from abyssal sediments preserves patterns of plankton macroecology. *Biogeosciences* 14, 2741–2754. doi: 10.5194/bg-14-2741-2017

More, K. D., Orsi, W. D., Galy, V., Giosan, L., He, L., Grice, K., et al. (2018). A 43 kyr record of protist communities and their response to oxygen minimum zone variability in the Northeastern Arabian Sea. *Earth Planet. Sci. Lett.* 496, 248–256. doi: 10.1016/j.epsl.2018.05.045

Morvezen, R., Cornette, F., Charrier, G., Guinand, B., Lapegue, S., Boudry, P., et al. (2013). Multiplex PCR sets of novel microsatellite loci for the great scallop *Pecten maximus* and their application in parentage assignment. *Aquat. Living Resour.* 26, 207–213. doi: 10.1051/alr/2013052

Murray, D. C., Haile, J., Dortch, J., White, N. E., Haouchar, D., Bellgard, M. I., et al. (2013). Scrapheap Challenge: A novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages. *Sci. Rep.* 3:3371. doi: 10.1038/srep03371

Narasimhan, V. M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., et al. (2019). The formation of human populations in South and Central Asia. *Science* 365:eaat7487. doi: 10.1126/science.aat7487

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., and Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature* 541, 302–310. doi: 10.1038/nature21347

Orlando, L., and Cooper, A. (2014). Using ancient DNA to understand evolutionary and ecological processes. *Annu. Rev. Ecol. Evol. Syst.* 45, 573–598. doi: 10.1146/annurev-ecolsys-120213-091712

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., et al. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* 499, 74–78. doi: 10.1038/nature12323

Paillard, C., Korsnes, K., Le Chevalier, P., Le Boulay, C., Harkestad, L., Eriksen, A. G., et al. (2008). *Vibrio tapetis*-like strain isolated from introduced Manila clams *Ruditapes philippinarum* showing symptoms of brown ring disease in Norway. *Dis. Aquat. Organ.* 81, 153–161. doi: 10.3354/dao01950

Paillard, C., Percelay, L., Le Pennec, M., and Le Picard, D. (1989). Origine pathogène de l'"anneau brun" chez *Tapes philippinarum* (Mollusque, bivalve). *C. R. Acad. Sci.* 309, 235–241.

Paillard, C., Roux, F. L., and Borrego, J. J. (2004). Bacterial disease in marine bivalves, a review of recent studies: trends and evolution. *Aquat. Living Resour.* 17, 477–498. doi: 10.1051/alr:2004054

Palkopoulou, E., Mallick, S., Skoglund, P., Enk, J., Rohland, N., Li, H., et al. (2015). Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Curr. Biol.* 25, 1395–1400. doi: 10.1016/j.cub.2015.04.007

Park, K.-I., Paillard, C., Le Chevalier, P., and Choi, K.-S. (2006). Report on the occurrence of brown ring disease (BRD) in Manila clam, *Ruditapes philippinarum*, on the west coast of Korea. *Aquaculture* 255, 610–613. doi: 10.1016/j.aquaculture.2005.12.011

Paterno, M., Bat, L., Souissi, J. B., Boscari, E., Chassanite, A., Congiu, L., et al. (2019). A genome-wide approach to the phylogeography of the mussel *Mytilus galloprovincialis* in the Adriatic and the Black Seas. *Front. Mar. Sci.* 6:566. doi: 10.3389/fmars.2019.00566

Pawłowska, J., Lejzerowicz, F., Esling, P., Szczuciński, W., Zajączkowski, M., and Pawlowski, J. (2014). Ancient DNA sheds new light on the Svalbard foraminiferal fossil record of the last millennium. *Geobiology* 12, 277–288. doi: 10.1111/gbi.12087

Pečnerová, P., Díez-del-Molino, D., Dussex, N., Feuerborn, T., von Seth, J., van der Plicht, J., et al. (2017). Genome-based sexing provides clues about behavior and social structure in the woolly mammoth. *Curr. Biol.* 27, 3505–3510.e3. doi: 10.1016/j.cub.2017.09.064

Pedersen, J. S., Valen, E., Velazquez, A. M. V., Parker, B. J., Rasmussen, M., Lindgreen, S., et al. (2014). Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Res.* 24, 454–466. doi: 10.1101/gr.163592.113

Pedersen, M. W., Overballe-Petersen, S., Ermini, L., Sarkissian, C. D., Haile, J., Hellstrom, M., et al. (2015). Ancient and modern environmental DNA. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370:20130383. doi: 10.1098/rstb.2013.0383

Pedersen, M. W., Ruter, A., Schweger, C., Friebe, H., Staff, R. A., Kjeldsen, K. K., et al. (2016). Postglacial viability and colonization in North America's ice-free corridor. *Nature* 537, 45–49. doi: 10.1038/nature19085

Pérez-Mayol, S., Blasco, J., Tornero, V., Morales-Nin, B., Massanet, A., and Tovar-Sánchez, A. (2014). Are the shells of *Scrobicularia plana* useful for monitoring trace metal pollution events? *J. Environ. Biol. Acad. Environ. Biol. India* 35, 9–17.

Quayle, D. B. (1964). Distribution of introduced marine Mollusca in British Columbia waters. *J. Fish. Res. Board Can.* 21, 1155–1181. doi: 10.1139/f64-102

Renaud, G., Slon, V., Duggan, A. T., and Kelso, J. (2015). Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* 16:224. doi: 10.1186/s13059-015-0776-0

Reynolds, D. J., Scourse, J. D., Halloran, P. R., Nederbragt, A. J., Wanamaker, A. D., Butler, P. G., et al. (2016). Annually resolved North Atlantic marine climate over the last millennium. *Nat. Commun.* 7:13502. doi: 10.1038/ncomms13502

Rick, T. C., Reeder-Myers, L. A., Hofman, C. A., Breitburg, D., Lockwood, R., Henkes, G., et al. (2016). Millennial-scale sustainability of the Chesapeake Bay Native American oyster fishery. *Proc. Natl. Acad. Sci. U.S.A.* 113, 6568–6573. doi: 10.1073/pnas.1600019113

Riginos, C., and Cunningham, C. W. (2005). INVITED REVIEW: Local adaptation and species segregation in two mussel (*Mytilus edulis* × *Mytilus trossulus*) hybrid zones. *Mol. Ecol.* 14, 381–400. doi: 10.1111/j.1365-294X.2004.02379.x

Rogers-Bennett, L., Haaker, P. L., Huff, T. O., Dayton, P. K., and Hall, R. (2002). Estimating baseline abundances of abalone in California for restoration. *CalCOFI Rep.* 43, 97–111.

Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., and Reich, D. (2015). Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370:20130624. doi: 10.1098/rstb.2013.0624

Sadler, J., Carré, M., Azzoug, M., Schauer, A. J., Ledesma, J., Cardenas, F., et al. (2012). Reconstructing past upwelling intensity and the seasonal dynamics of primary productivity along the Peruvian coastline from mollusk shell stable isotopes. *Geochem. Geophys. Geosyst.* 13:Q01015. doi: 10.1029/2011GC003595

Sanna, D., Lai, T., Cossu, P., Scarpa, F., Dedola, G. L., Cristo, B., et al. (2017). Cytochrome c oxidase subunit I variability in *Ruditapes decussatus* (Veneridae) from the western Mediterranean. *Eur. Zool. J.* 84, 554–565. doi: 10.1080/24750263.2017.1395914

Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* 9, 1056–1082. doi: 10.1038/nprot.2014.063

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., Al-Rasheid, K. A., Willerslev, E., et al. (2012). Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* 13:178. doi: 10.1186/1471-2164-13-178

Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* 9:88. doi: 10.1186/s13104-016-1900-2

Seersholm, F. V., Cole, T. L., Grealy, A., Rawlence, N. J., Greig, K., Knapp, M., et al. (2018). Subsistence practices, past biodiversity, and anthropogenic impacts revealed by New Zealand-wide ancient DNA survey. *Proc. Natl. Acad. Sci. U.S.A.* 115, 7771–7776. doi: 10.1073/pnas.1803573115

Seguin-Orlando, A., Schubert, M., Clary, J., Stagegaard, J., Alberdi, M. T., Prado, J. L., et al. (2013). Ligation bias in illumina next-generation DNA libraries: implications for sequencing ancient genomes. *PLoS One* 8:e78575. doi: 10.1371/journal.pone.0078575

Skotte, L., Korneliussen, T. S., and Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics* 195, 693–702. doi: 10.1534/genetics.113.154138

Śmietanka, B., and Burzyński, A. (2017). Disruption of doubly uniparental inheritance of mitochondrial DNA associated with hybridization area of European *Mytilus edulis* and *Mytilus trossulus* in Norway. *Mar. Biol.* 164:209. doi: 10.1007/s00227-017-3235-5

Smietanka, B., Burzyński, A., and Wenne, R. (2010). Comparative genomics of marine mussels (*Mytilus* spp.) gender associated mtDNA: rapidly evolving atp8. *J. Mol. Evol.* 71, 385–400. doi: 10.1007/s00239-010-9393-4

Smietanka, B., Zbawicka, M., Sańko, T., Wenne, R., and Burzyński, A. (2013). Molecular population genetics of male and female mitochondrial genomes in subarctic *Mytilus trossulus*. *Mar. Biol.* 160, 1709–1721. doi: 10.1007/s00227-013-2223-7

Steinhardt, J., Butler, P. G., Carroll, M. L., and Hartley, J. (2016). The application of long-lived bivalve sclerochronology in environmental baseline monitoring. *Front. Mar. Sci.* 3:176. doi: 10.3389/fmars.2016.00176

Stephens, L., Fuller, D., Boivin, N., Rick, T., Gauthier, N., Kay, A., et al. (2019). Archaeological assessment reveals Earth's early transformation through land use. *Science* 365, 897–902. doi: 10.1126/science.aax1192

Surge, D. M., Lohmann, K. C., and Goodfriend, G. A. (2003). Reconstructing estuarine conditions: oyster shells as recorders of environmental change, Southwest Florida. *Estuar. Coast. Shelf Sci.* 57, 737–756. doi: 10.1016/S0272-7714(02)00370-0

Svendsen, J. I., Alexanderson, H., Astakhov, V. I., Demidov, I., Dowdeswell, J. A., Funder, S., et al. (2004). Late Quaternary ice sheet history of northern Eurasia. *Quat. Sci. Rev.* 23, 1229–1271. doi: 10.1016/j.quascirev.2003.12.008

The IUCN Red List of Threatened Species (2019). *IUCN Red List Threat, Species*. Available at: https://www.iucnredlist.org/en (accessed October 13, 2019).

Trinkler, N., Labonne, M., Marin, F., Jolivet, A., Bohn, M., Poulain, C., et al. (2010). Clam shell repair from the brown ring disease: a study of the organic matrix using confocal Raman micro-spectrometry and WDS microprobe. *Anal. Bioanal. Chem.* 396, 555–567. doi: 10.1007/s00216-009-3114-0

Trivellini, M. M., Van der Molen, S. V., and Márquez, F. (2018). Fluctuating asymmetry in the shell shape of the Atlantic Patagonian mussel *Mytilus platensis* generated by habitat-specific constraints. *Hydrobiologia* 822, 189–201. doi: 10.1007/s10750-018-3679-8

Valéry, L., Fritz, H., Lefeuvre, J.-C., and Simberloff, D. (2008). In search of a real definition of the biological invasion phenomenon itself. *Biol. Invasions* 10, 1345–1351. doi: 10.1007/s10530-007-9209-7

Van Wormhoudt, A., Roussel, V., Courtois, G., and Huchette, S. (2011). Mitochondrial DNA introgression in the European abalone *Haliotis tuberculata tuberculata*: evidence for experimental mtDNA paternal inheritance and a natural hybrid sequence. *Mar. Biotechnol.* 13, 563–574. doi: 10.1007/s10126-010-9327-6

Vander Putten, E., Dehairs, F., Keppens, E., and Baeyens, W. (2000). High resolution distribution of trace elements in the calcite shell layer of modern *Mytilus edulis*: environmental and biological controls. *Geochim. Cosmochim. Acta* 64, 997–1011. doi: 10.1016/S0016-7037(99)00380-4

Vendrami, D. L. J., Houston, R. D., Gharbi, K., Telesca, L., Gutierrez, A. P., Gurney-Smith, H., et al. (2019a). Detailed insights into pan-European population structure and inbreeding in wild and hatchery Pacific oysters (*Crassostrea gigas*) revealed by genome-wide SNP data. *Evol. Appl.* 12, 519–534. doi: 10.1111/eva.12736

Vendrami, D. L. J., Noia, M. D., Telesca, L., Handal, W., Charrier, G., Boudry, P., et al. (2019b). RAD sequencing sheds new light on the genetic structure and local adaptation of European scallops and resolves their demographic histories. *Sci. Rep.* 9:7455. doi: 10.1038/s41598-019-43939-4

Vokhshoori, N. L., and McCarthy, M. D. (2014). Compound-Specific $\delta^{15}$N amino acid measurements in littoral mussels in the California upwelling ecosystem: a new approach to generating baseline $\delta^{15}$N Isoscapes for coastal ecosystems. *PLoS One* 9:e98087. doi: 10.1371/journal.pone.0098087

Voroshilova, I. S., Artamonova, V. S., Makhrov, A. A., and Slyn'ko, Y. V. (2010). Natural hybridization of two mussel species *Dreissena polymorpha* (Pallas, 1771) and *Dreissena bugensis* (Andrusov, 1897). *Biol. Bull.* 37, 542–547. doi: 10.1134/S1062359010050158

Wenne, R., Bach, L., Zbawicka, M., Strand, J., and McDonald, J. H. (2016). A first report on coexistence and hybridization of *Mytilus trossulus* and *M. edulis* mussels in Greenland. *Polar Biol.* 39, 343–355. doi: 10.1007/s00300-015-1785-x

Williams, M. R., Stedtfeld, R. D., Engle, C., Salach, P., Fakher, U., Stedtfeld, T., et al. (2017). Isothermal amplification of environmental DNA (eDNA) for direct field-based monitoring and laboratory confirmation of *Dreissena* sp. *PLoS One* 12:e0186462. doi: 10.1371/journal.pone.0186462

Wilson, J., Matejusova, I., McIntosh, R. E., Carboni, S., and Bekaert, M. (2018). New diagnostic SNP molecular markers for the *Mytilus* species complex. *PLoS One* 13:e0200654. doi: 10.1371/journal.pone.0200654

Yang, D. Y., Eng, B., Waye, J. S., Dudar, J. C., and Saunders, S. R. (1998). Improved DNA extraction from ancient bones using silica-based spin columns. *Am. J. Phys. Anthropol.* 105, 539–543. doi: 10.1002/(sici)1096-8644(199804)105:4<539::aid-ajpa10>3.0.co;2-1

Zbawicka, M., Sańko, T., Strand, J., and Wenne, R. (2014a). New SNP markers reveal largely concordant clinal variation across the hybrid zone between *Mytilus* spp. in the Baltic Sea. *Aquat. Biol.* 21, 25–36. doi: 10.3354/ab00566

Zbawicka, M., Trucco, M. I., and Wenne, R. (2018). Single nucleotide polymorphisms in native South American Atlantic coast populations of smooth shelled mussels: hybridization with invasive European *Mytilus galloprovincialis*. *Genet. Sel. Evol.* 50:5. doi: 10.1186/s12711-018-0376-z

Zbawicka, M., Wenne, R., and Burzyński, A. (2014b). Mitogenomics of recombinant mitochondrial genomes of Baltic Sea *Mytilus* mussels. *Mol. Genet. Genomics* 289, 1275–1287. doi: 10.1007/s00438-014-0888-3

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Ecology and Evolution

PERSPECTIVE
published: 25 March 2020
doi: 10.3389/fevo.2020.00074

# Ancient Plant DNA as a Window Into the Cultural Heritage and Biodiversity of Our Food System

*Natalia A. S. Przelomska*[1,2]*, *Chelsey G. Armstrong*[3] *and Logan Kistler*[1]*

[1] Department of Anthropology, Smithsonian Institution, National Museum of Natural History, Washington, DC, United States, [2] Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, Richmond, United Kingdom, [3] Department of Anthropology, University of British Columbia, Vancouver, BC, Canada

Since the beginning of the ancient DNA revolution in the 1980s, archeological plant remains and herbarium specimens have been analyzed with molecular techniques to probe the evolutionary interface of plants and humans. In tandem with archeobotany, ethnobiology, and other methods, ancient DNA offers tremendous insights into the co-evolution of people and plants, and the modern genomic era offers increasingly nuanced perspectives on plant use through time. Meanwhile, our global food system faces threats linked with declining biodiversity, an uncertain climate future, and vulnerable crop–wild relatives. Ancient plant DNA does not yield easy answers to these complex challenges, but we discuss how it can play an important role in ongoing conversations about resilience, sustainability, and sovereignty in our food system.

**Keywords: archaeogenomics, ethnobiology, agrobiodiversity, food security, domestication, archaeobotany**

OPEN ACCESS

***Edited by:***
*Nathan Wales,*
*University of York, United Kingdom*

***Reviewed by:***
*Chrsitine Ann Hastorf,*
*University of California, Berkeley,*
*United States*
*Bastien Llamas,*
*University of Adelaide, Australia*

***Correspondence:***
*Natalia A. S. Przelomska*
*n.przelomska@kew.org*
*Logan Kistler*
*kistlerl@si.edu*

***Specialty section:***
*This article was submitted to*
*Paleoecology,*
*a section of the journal*
*Frontiers in Ecology and Evolution*

***Received:*** *13 December 2019*
***Accepted:*** *04 March 2020*
***Published:*** *25 March 2020*

***Citation:***
*Przelomska NAS, Armstrong CG and Kistler L (2020) Ancient Plant DNA as a Window Into the Cultural Heritage and Biodiversity of Our Food System. Front. Ecol. Evol. 8:74. doi: 10.3389/fevo.2020.00074*

## INTRODUCTION

Vascular plants are foundational to all terrestrial ecosystems, and have been around for hundreds of millions of years. The flowering plants—Angiospermae, the largest and most diverse group in the plant kingdom—arose ∼140–250 million years ago (Magallón et al., 2015; Foster et al., 2017), and over this time they have evolved an astounding range of diversity in functional and developmental traits, genome characteristics, and physiology across almost 300,000 known species (Christenhusz and Byng, 2016). The genus *Homo* began to co-habit with plants a mere ∼2.8 million years ago (Hublin, 2015) and a multitude of co-evolutionary relationships between plants and people ensued (Nabhan, 2002; Allaby et al., 2015).

The co-evolution of plants and humans is most prominently exemplified by the emergence of domesticated plants during the Holocene—the epoch beginning around 12,000 years ago after the last ice age—leading to the evolution of agricultural ecosystems and economies, and resulting in massive societal and ecological shifts. Wild plants and their domesticated counterparts thrived in sympatry and continued to interbreed over the course of a protracted domestication process (Purugganan and Fuller, 2011; Allaby et al., 2017). A vast number of other plant species are valuable to humans, but have not been subjected to evolutionary pressures resulting in domestication. The proportion of the world's botanical biodiversity with known uses by humans is estimated to include around 10%, or ∼31,000 species. Of these, 57% are medicinal, 36.5% used for materials and clothing, and 17.8% for food (Royal Botanic Gardens Kew, 2016).

Our current food systems have been profoundly shaped by changes to agriculture that took place after the industrialization of farming methods—breeding for more efficient cultivation, high, uniform yield, large-scale mechanization, and trends toward monocultures. Modern crop practices can therefore overshadow much of the ecological innovation that came before, largely in the hands of Indigenous cultivators. Significant knowledge pertaining to the past usage of plants persists in historical-ecological records, and in traditional knowledge systems and practices. Garnett et al. (2018) estimate that at least 28% of the world's land surface is owned/managed by Indigenous peoples, including some of the most biologically diverse ecosystems on Earth. In some cases these ecosystems include managed perennials and cultivars that have been maintained for millennia (Chouin, 2009; Turner, 2014; Clement et al., 2015; Ford and Nigh, 2016).

The application of archeological science has been hugely successful for complementing traditional knowledge and exploring deeper-time evolutionary and cultural processes surrounding domestication (e.g., Hoffmann et al., 2016). Archeobotany has been valuable for determining the chronology and composition of past agroecosystems (Bakels and Jacomet, 2003; Smith and Yarnell, 2009; Motuzaite-Matuzeviciute et al., 2013; d'Alpoim Guedes et al., 2014), studying the domestication syndrome (Nesbitt, 1998; Fuller et al., 2019), and elucidating past diets with significant time-depth—for example demonstrating a long-term reliance of early humans on starchy foods (Larbey et al., 2019).

At the same time, we live in a moment when the security of our food systems is a subject of global concern (Tscharntke et al., 2012). The most critical conversations for food security and food sovereignty, sustainability, and ecological impact should revolve around land and energy use, agricultural practices, biodiversity priorities, and the threat of worsening global climate uncertainty for predictable growth seasons and crop yields. However, looking to the past 10,000 years of crop evolution and plant use offers context, and could thus contribute to developing solutions, for these important issues.

## ANCIENT DNA: UNRAVELING PATHWAYS TO DOMESTICATION

In combination with archeobotany and other anthropological and analytical methods, ancient DNA (aDNA) is a particularly exciting means for opening doors to the human past using plant remains (Gutaker and Burbano, 2017; Allaby et al., 2018; di Donato et al., 2018; Pont et al., 2019). Ancient plant genomics allows us to address increasingly nuanced questions regarding plant cultivation and domestication, such as the relationships between crops and wild relatives, routes of plant dispersal mediated by humans, and plant traits whose evolution was instrumental to domestication. For example, we are now able to examine shifting allelic makeup for traits related to domestication (e.g., Mascher et al., 2016; Ramos-Madrigal et al., 2016), probe the adaptive trajectories of plants carried into new environments (e.g., da Fonseca et al., 2015; Swarts et al., 2017), infer culinary and other crop-linked cultural preferences (e.g., Castillo et al., 2016), gain insights regarding the source of genetic variants under selection (novel mutations, standing variation, or introgression; Gutaker et al., 2019a), and estimate the extent and timing of admixture with wild relatives and introduced lineages (Mascher et al., 2016; Scott et al., 2019).

The majority of archeobotanical remains are preserved by charring, but unfortunately this process does not preserve ancient DNA at predictable or useful levels for genomic analysis (Nisterlberger et al., 2016). Similar to most subfields of trace sciences (archeology, paleoecology), archeobotanical contexts boasting the outstanding preservation necessary for aDNA analysis over extended archeological time are typically restricted to dry caves or rockshelters, extremely arid localities, or waterlogged deposits. Maize has been particularly productive due to its widespread importance throughout the pre-colonial Americas, and maize remains suitably preserved for genomic analysis have been found in Mexico, the southwestern US and Colorado Plateau, the Andes, and caves in lowland Brazil (da Fonseca et al., 2015; Ramos-Madrigal et al., 2016; Vallebueno-Estrada et al., 2016; Swarts et al., 2017; Kistler et al., 2018).

Application of a target sequence capture approach on 32 maize samples spanning 5,910 to 670 years old in Mesoamerica and the Southwest illuminated routes of maize dispersal into the United States, and the timing of selection on genes associated with adaptation to the challenging new environment (da Fonseca et al., 2015). Further aDNA research revealed that maize adapting to these new environments could be only marginally suited to the unfamiliar growth conditions (Swarts et al., 2017), highlighted a long-lasting, intertwined relationship between maize and its wild progenitor *Zea mays* ssp. *parviglumis* (Ramos-Madrigal et al., 2016; Vallebueno-Estrada et al., 2016), and indicated that the cultivated gene pool was subdivided and carried away from the domestication center before domesticated forms were robust and stable (Kistler et al., 2018). Combining these insights with previous work highlighting the role of adaptive crop–wild gene flow in maize evolution (van Heerwaarden et al., 2011), and the long-term decay of genetic diversity in maize (Wang et al., 2017), aDNA fills in critical elements of the complex, protracted evolutionary process leading to the world's most abundant grain crop (Food and Agriculture Organization of the United Nations, 2018).

Another fortuitous archeobotanical time series presented itself in sorghum remains (*Sorghum bicolor*), preserved in the fortified hilltop city of Qasr Ibrim in the Lower Nubian Desert—present-day Egypt (Smith et al., 2019). Sorghum seeds spanning 1,805 to 450 years old were used along with herbarium and modern specimens to dissect genetic diversity and human selection over time. Additionally, this study showed a long-term loss of genetic diversity through archeological time up to the present day, highlighting the continuous erosion of biodiversity under human influences.

In other species, archeogenomics has provided fundamental insight regarding the evolutionary ecology and long-term niche of domesticates (squashes; Kistler et al., 2015), shifting baselines of genetic diversity through time (sunflowers; Wales et al., 2019), genomic structural upheaval since domestication (cotton; Palmer et al., 2012), and extreme kinship networks and clonal lineages spanning millennia (grapevine; Ramos-Madrigal et al., 2019). In emmer wheat, a species with a particularly large and complex genome, the recently reported genome from a 3,000-year-old specimen revealed significant tracts of diversity absent from modern wheats, underscoring the loss of crop genetic variation over time (Scott et al., 2019). These examples rely on ancient genomics to help build the conceptual foundation of domestication science, which serves as important background to help inform the future of agroecosystems resilience globally.

The genetic basis of local adaptation to new habitats is a subject particularly suited to ancient and historic plant DNA. Throughout their history, crop plants have been routinely presented with novel growth conditions when moved into new habitats by humans. These comprise changes in daylength and temperature, UV exposure, soil conditions, symbiont availability, and other abiotic and biotic variables. The amenability of a staple crop to grow in new regions was critical to the mobility of early farming communities, and many of the same adaptative strategies could be of interest in the present day given our changing climate and conditions of land-use. The mechanisms underpinning these adaptations is therefore of interest to both domestication scientists and contemporary crop breeders.

One such adaptation that can unlock a range of new agroecological niches is the loss of sensitivity to photoperiod (daylength), which in many plants determines time of flowering. Selection pressure on responsiveness to photoperiod has been important in influencing the rate of post-domestication range expansion in many crops (Purugganan and Fuller, 2009), for example permitting the northward spread of rice, barley, beet, and flax from their native ranges (Fujino and Sekiguchi, 2005; Lister et al., 2009; Pin et al., 2012; Gómez-Ariza et al., 2015; Gutaker et al., 2019b). These studies draw on the gene pool available in either landrace or cultivar germplasm material, but even more detailed information can be retrieved using samples that provide time depth. For example, using data from 1,900-years-old maize cobs from Colorado and a modern germplasm panel for comparison, Swarts et al. (2017) demonstrated that alleles associated with early flowering had already been selected for from the extensive standing variation, as an early step in adaptation to more temperate latitudes.

In the absence of archeological samples—the prevailing situation for root crops—herbarium specimens are excellent repositories of valuable genomic information along similar lines at shallower time-depths. For example, Gutaker et al. (2019a) demonstrated with historical genomes from herbaria that long-day adaptation of European potato lineages was most likely due to *de novo* variants arising in their genomes after their arrival in Europe. Furthermore, that study showed that the early European potato genomes were already pre-adapted to this daylength response due to modifications in the gibberellin pathway—a hormonal signaling system mediating tuber bolting time in response to daylength—which occurred prior to their trans-Atlantic dispersal. There has been a recent surge in the use and function of herbarium collections, most notably in conservation biology (Krupnick et al., 2009). However, there has been little inclusion of cultural knowledge, societal values, or domestication history associated with specimens upon their collection. This is surprising given that many of the first herbaria were deeply motivated by ethnobotanical questions (Martín et al., 2010). The early call for collectors to be more mindful of the genetic preservation of specimens (e.g., storing tissues with silica gel), is followed up by another call to curate ethnobotanical and cultural knowledge with specimens in order to better situate collections for research relating to food security/sovereignty, agroecology, and biocultural diversity (Stepp and Thomas, 2007; Nesbitt, 2014).

Past plant–human interactions are best understood when a multidisciplinary approach is applied (e.g., Dalby, 2013), and ancient genomes can be a powerful means to complement or verify historical records. Paris (2015, 2016) summarized how discovery of 4,000-years-old seeds of watermelon (*Citrullus lanatus*) in conjunction with biblical records and iconography, as well as linguistics, can shed light on the cultivation history of this cucurbit crop. A recent study built upon these insights by investigating a 3,500-years-old *Citrullus lanatus* leaf from a Pharaonic tomb using genomic methods, further illuminating the origins of watermelon by revealing that New Kingdom Egyptians were already consuming a pink-fleshed and sweet variety (Renner et al., 2019).

## PLANT BIODIVERSITY AND FOOD SECURITY

One important consideration for agricultural sustainability is population genetic health of crops over time, manifested as both erosion of genetic variation and the accumulation of harmful variants. Current threats to the Cavendish banana—the top commercial banana variety comprising 99% of the global export market—illustrate an extreme case of the risk in lost diversity. Plantations of Cavendish are succumbing to *Fusarium* wilt tropical 4 (TR4), resulting in massive worldwide losses (Solly, 2019). Cavendish is favored because of its storability and stability in shipping, but as a result, the entire export banana market relies on this single cultivar lacking resistance to the current pathogenic threat. This fungal pathogen reached the Americas just last year, and now threatens all the world's key banana-producing regions (Galvis, 2019; Solly, 2019).

A similar scenario created conditions for the Great Famine in nineteenth century Ireland, killing over a million and displacing the same number again (Turner, 2005). Several strains of potato were grown by vegetative propagation in Europe at the time, but their diversity was a subset of their South American landraces of origin, and they had been shaped by adaptive stressors of the new growing environment (Gutaker et al., 2019a). A pathogenic oomycyte, *Phytophthora infestans*, caused years of devastating crop losses since these potato varieties lacked resistance in the population (Austin Bourke, 1964; Saville et al., 2016). DNA from herbarium specimens revealed that it was likely a resurgence of older, blight-resistant landraces with higher proportions of Andean ancestry that re-invigorated European potato production (Gutaker et al., 2019a). These are extreme examples where vegetative propagation facilitates particularly low field-level diversity, but they highlight the vulnerabilities of low diversity and a lack of genetic resilience to such pathogenic onslaughts.

Another consideration for food system resilience is the accumulation of deleterious genetic mutants due to the demographic strain of agriculture on crops. In many other crop (and domestic animal) species analyzed so far, this “mutation load” has increased significantly in domesticated lineages compared with wild forms (Liu et al., 2017; Wang et al., 2017; Smith et al., 2019), with ultimate fitness effects threatening cultivated populations (Kono et al., 2016; Kremling et al., 2018). Recent studies in both plants and animals highlight that the mutation load, though linked with long-term dispersal (Kistler et al., 2018) and demography (Wang et al., 2017), has in some species spiked in recent centuries (Gaunitz et al., 2018; Kistler et al., 2018). Therefore, much of the elevated modern-day mutation load may be the result of recent breeding activities as opposed to the protracted emergence and long-term traditional curation of domesticated species.

Cassava (also called manioc or yuca) is the world’s sixth leading food plant, feeding 800 million as a staple crop and cultivated across a pan-tropical range with an Amazonian origin (Food and Agriculture Organization of the United Nations, 2013, 2014). However, genomic analysis reveals that it suffers an excess of deleterious genetic variants—a high mutation load (Ramu et al., 2017). Cassava is propagated almost exclusively by vegetative stem cuttings, while its wild counterpart reproduces by out-crossing. Therefore, researchers have hypothesized that this shift toward clonal propagation since domestication has pre-empted the wild plant’s abilities to purge its genome of harmful variants through recombination (Ramu et al., 2017). Thus, by bringing the plant into cultivation and altering its reproductive biology, humans have created an important food source with the unintended consequence of introducing genomic vulnerability.

The case of sorghum presents a contrasting view of the mutation load in crops. Sorghum, the world’s most important cereal crop in arid regions, is widely grown as a staple food particularly in developing nations in Asia and Africa (Food and Agriculture Organization of the United Nations International Crops Research Institute for the Semi-Arid Tropics, 1996). Like other domestic species, the mutation load is elevated compared to the wild progenitor. However, archeogenomic time-series analysis from Qasr Ibrim in modern-day Egypt revealed that ancient farmers took advantage of periodic hybridization between lineages to boost diversity and ameliorate the mutation load in real time (Smith et al., 2019). Thus, in some species, traditional management strategies have pre-empted this burdensome accumulation of harmful variants.

Some researchers call for protecting the biodiversity of our food systems at higher levels of organization, namely by prioritizing species and ecosystem biodiversity in our food production pipeline (Hammer et al., 2003; Wood et al., 2015), and this is thought to be most feasible through deployment of diverse local domesticates (Shelef et al., 2017). A vast proportion of human nutrition globally (>90% calories) is derived from just 20 plant species (Massawe et al., 2016), where the dominance of rice, wheat, and maize was facilitated by the Green Revolution commencing in the 1960s. This agricultural transition brought immediate public health and ecological benefits in some regions, including short-term reduction of poverty and temporary reduced deforestation through higher crop yields. The Revolution, however, exposed asymmetries of power dynamics in cache crop economies (Shiva, 1992), and revealed a global Achilles heel in crop field resilience: The global reliance on such a low diversity of plant species can result in higher risks of large-scale famine.

Climatic predictions for the near future raise another set of concerns for crop sustainability (Khoury et al., 2014). Rice, for example, is particularly vulnerable to climate change, where both flooding and drought stress can cause complete crop failure (Nguyen, 2005). However, wild rice (*Zizania palustris*) is highly adaptive to these specific features of climate change and, coupled with Indigenous Anishinaabeg knowledge (e.g., harvest protocols, life histories, grain quality) has the potential to enlighten breeding programs aimed specifically at climate change resiliency (Zilberstein, 2015). Increasing the species diversity and genetic diversity of agricultural systems improves the resilience of our food systems, and may often be best implemented at local scales (Mijatović et al., 2013). Local distribution centers are a key factor in this effort, for example supporting diversity of apple varieties in the US (Goland and Bauer, 2004).

The traditional knowledge of Indigenous famers is a key component of a diversified approach for ensuring that agricultural landscapes thrive into the climatically unstable future (Makondo and Thomas, 2018). As an example, traditional knowledge used for managing yam (*Dioscorea* spp.) crops in Oceania has led to the emergence of valuable hybrids—confirmed by a phylogenetic analysis using herbarium specimens—whose existence could be relevant to the future management of this staple crop (Chaïr et al., 2016). The diversity in traditional landraces is also greater than in breeding programs and industrialized cropping contexts. For example, molecular analysis revealed that diversity sampled from an *ex situ* cassava core collection was matched by indigenous landraces from a single village in Guyana (Elias et al., 2000). Indigenous peoples who have occupied the same landscape for centuries or indeed millennia have had to, unavoidably, manage their food resources through extensive oscillations in weather, climate, and general environmental change. As a result, the codified knowledge of what, and how, to plant is extensive, and the diversity of germplasm maintained is extraordinary (Anderson, 2016).

Studies involving ancient DNA and archeobotanical remains alongside ethnobotanical knowledge may inform and encourage cultivation of valuable plant species which have been marginalized in recent times. For instance, beaked hazelnut (*Corylus cornauta*) in the Pacific Northwest is perceived as an obscure crop in management terms, but surveys of traditional knowledge expose its deep cultural significance, with uses as food, medicine, and construction material (Armstrong et al., 2018). Recently, this species has also been used in hybridization programs given its resistance to Eastern filbert blight (caused by *Anisogramma anomala*), which has been globally devastating for European hazelnut (*Corylus avellana*) crops (Muehlbauer et al., 2014; Oregon State University., 2019). In northeastern North America a whole complex of plants lost for generations and only known as cultigens from archeobotanical evidence include knotweed (*Polygonum erectum*), sumpweed (*Iva annua*), and chenopods (*Chenopodium berlandieri*) of eastern North America (Mueller et al., 2017). These species are good candidates for potential re-domestication, which could simultaneously work against low food biodiversity and economic insecurity in rural North America (Mueller et al., 2017).

Similar marginalized crops elsewhere include the Bambara groundnut (*Vigna subterranea*) native to Africa, and finger millet (*Eleusine coracana*) grown in arid areas of the Indian subcontinent and Africa. Both have been historically dubbed a "poor man's crop" (Azam-Ali et al., 2001; Joshi and Joshi, 2002), but are robust in environments prone to drought, and are nutritionally valuable. Similarly, the Asian millets foxtail (*Setaria italica*) and broomcorn (*Panicum miliaceum*) were not popular enough to become a staple food crop in the past (Lightfoot et al., 2013), but early domestication, flexible seasonality, and persistence seen in the archeological record meant that these hardy crops played a crucial role in facilitating humans' mobility (Jones et al., 2016). The role of multidisciplinary research with an ancient DNA element concerning these marginalized species is much more than strictly scientifically and economically applicable—it highlights the importance of maintaining food sovereignty and preservation of cultural heritage (Coté, 2016).

## CONCLUSIONS

Theories emerging from plant archeogenomics can contribute in a significant way to the understanding of the evolution of plants and plant genomes; domestication serves as an excellent model system for evolution (Piperno, 2017; Zeder, 2017). Contemporary domestication paradigms stemming from this work highlight key concepts that could be instrumental to decision-making with regard to future food systems. However, complementary research on lesser known crops, other plant species that are valuable to humans, and the community ecology of these species is still lacking (see Deur and Turner, 2005). A much more direct application of ancient genomes for the enhancement of our food systems could be mining for useful past traits lost in modern crop fields, but potentially recoverable through crop–wild relatives.

The unstable condition of our food systems is, in part, a biodiversity problem. Several strategies have been put forward to alleviate the burden of scant biological diversity at the highest production levels in our food systems (Tscharntke et al., 2012). One is the improvement of the diversity of elite crops, which can be introducing new traits through breeding or genetic modification. Gene editing using CRISPR-Cas9 technology has in fact been successfully used to produce pathogen-resistant bananas (Tripathi et al., 2019), but these face regulatory obstacles and offer no guarantee of resistance into the future (Maxmen, 2019).

Other approaches involve drawing on diversity already available, which is introduced either through crossing of wild relatives with cultivated varieties or utilizing the wild relatives themselves. Such a "substitution-diversification" strategy could promote species better suited to future climates and push more plant species down the spectrum of domestication (Borrell et al., 2020). Both dimensions of this approach highlight that extensive diversity, which has emerged due to evolutionary pressures over millennial timescales, can be harnessed to its fullest potential, rather than endangered due to an increased reliance on genetically homogenous cultivars (Bevan et al., 2017).

Twenty percent of described angiosperms are threatened with extinction (Brummitt et al., 2015), and the IUCN has recently recognized the attendant risks to food production by recognizing a wide range of crop–wild relatives threatened by habitat disturbance (Maxted and Dulloo, 2017). Ancient plant genomics is not the solution to our food biodiversity predicament. However, it can provide nuanced, long-term insight into where we are and how we got here with respect to the evolution of the species that feed the world.

## DATA AVAILABILITY STATEMENT

All datasets analyzed for this study are included in the article and the supplementary files.

## AUTHOR CONTRIBUTIONS

NP and LK drafted the manuscript. All authors contributed the text, revised, read, and approved the submitted version.

## REFERENCES

Allaby, R. G., Kistler, L., Gutaker, R. M., Ware, R., Kitchen, J. L., Smith, O., et al. (2015). Archaeogenomic insights into the adaptation of plants to the human environment: pushing plant–hominin co-evolution back to the Pliocene. *J. Hum. Evol.* 79, 150–157. doi: 10.1016/j.jhevol.2014.10.014

Allaby, R. G., Smith, O., and Kistler, L. (2018). "Archaeogenomics and crop adaptation," in *Paleogenomics: Genome-Scale Analysis of Ancient DNA*, eds C. Lindqvist and O. P. Rajora (Cham: Springer), 189–203. doi: 10.1007/13836_2018_51

Allaby, R. G., Stevens, C., Lucas, L., Maeda, O., and Fuller, D. Q. (2017). Geographic mosaics and changing rates of cereal domestication. *Philos. Trans. R. Soc. B Biol. Sci.* 372:20160429. doi: 10.1098/rstb.2016.0429

Anderson, E. N. (2016). *Caring for Place: Ecology, Ideology, and Emotion in Traditional Landscape Management*. Abingdon: Routledge.

Armstrong, C. G., Dixon, W. M., and Turner, N. J. (2018). Management and traditional production of beaked hazelnut (k'áp'xw-az', *Corylus cornuta*; Betulaceae) in British Columbia. *Hum. Ecol.* 46, 547–559. doi: 10.1007/s10745-018-0015-x

Austin Bourke, P. M. (1964). Emergence of potato blight, 1843–46. *Nature* 203, 805–808. doi: 10.1038/203805a0

Azam-Ali, S. N., Sesay, A., Karikari, S. K., Massawe, F. J., Aguilar-Manjarrez, J., Bannayan, M., et al. (2001). Assessing the potential of an underutilized crop–a case study using bambara groundnut. *Exp. Agric.* 37, 433–472. doi: 10.1017/S0014479701000412

Bakels, C., and Jacomet, S. (2003). Access to luxury foods in central Europe during the roman period: the archaeobotanical evidence. *World Archaeol.* 34, 542–557. doi: 10.1080/0043824021000026503

Bevan, M. W., Uauy, C., Wulff, B. B. H., Zhou, J., Krasileva, K., and Clark, M. D. (2017). Genomic innovation for crop improvement. *Nature* 543, 346–354. doi: 10.1038/nature22011

Borrell, J. S., Dodsworth, S., Forest, F., Pérez-Escobar, O. A., Lee, M. A., Mattana, E., et al. (2020). The climatic challenge: which plants will people use in the next century? *Env. Exp. Bot.* 170:103872. doi: 10.1016/j.envexpbot.2019.103872

Brummitt, N. A., Bachman, S. P., Griffiths-Lee, J., Lutz, M., Moat, J. F., Farjon, A., et al. (2015). Green plants in the red: a baseline global assessment for the IUCN sampled red list index for plants. *PLoS ONE* 10:e0135152. doi: 10.1371/journal.pone.0135152

Castillo, C. C., Tanaka, K., Sato, Y. I., Ishikawa, R., Bellina, B., Higham, C., et al. (2016). Archaeogenetic study of prehistoric rice remains from Thailand and India: evidence of early japonica in South and Southeast Asia. *Archaeol. Anthropol. Sci.* 8, 523–43. doi: 10.1007/s12520-015-0236-5

Chaïr, H., Sardos, J., Supply, A., Mournet, P., Malapa, R., and Lebot, V. (2016). Plastid phylogenetics of Oceania yams (*Dioscorea* spp., Dioscoreaceae) reveals natural interspecific hybridization of the greater yam (*D. alata*). *Bot. J. Linn. Soc.* 180, 319–333. doi: 10.1111/boj.12374

Chouin, G. (2009). *Forests of Power and Memory: An Archaeology of Sacred Groves in the Eguafo Polity, Southern Ghana (c. 500–1900 AD)*. Syracuse, NY: Syracuse University.

Christenhusz, M. J. M., and Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa* 261, 201–217. doi: 10.11646/phytotaxa.261.3.1

Clement, C. R., Denevan, W. M., Heckenberger, M. J., Junqueira, A. B., Neves, E. G., Teixeira, W. G., et al. (2015). The domestication of amazonia before european conquest. *Proc. Biol. Sci.* 282:20150813. doi: 10.1098/rspb.2015.0813

Coté, C. (2016). "Indigenizing" food sovereignty. Revitalizing indigenous food practices and ecological knowledges in Canada and the United States. *Humanities* 5:57. doi: 10.3390/h5030057

da Fonseca, R. R., Smith, B. D., Wales, N., Cappellini, E., Skoglund, P., Fumagalli, M., et al. (2015). The origin and evolution of maize in the American Southwest. *Nat. Plants* 1:14003. doi: 10.1101/013540

Dalby, A. (2013). *Food in the Ancient World From A to Z*. London: Routledge.

d'Alpoim Guedes, J., Lu, H., Li, Y., Spengler, R. N., Wu, X., and Aldenderfer, M. S. (2014). Moving agriculture onto the Tibetan plateau: the archaeobotanical evidence. *Archaeol. Anthropol. Sci.* 6, 255–269. doi: 10.1007/s12520-013-0153-4

Deur, D., and Turner, N. (2005). *Keeping It Living: Traditions of Plant Use and Cultivation on the Northwest Coast of North America*. Seattle, WA: University of Washington Press.

di Donato, A., Filippone, E., Ercolano, M. R., and Frusciante, L. (2018). Genome sequencing of ancient plant remains: findings, uses and potential applications for the study and improvement of modern crops. *Front. Plant Sci.* 9:441. doi: 10.3389/fpls.2018.00441

Elias, M., Panaud, O., and Robert, T. (2000). Assessment of genetic variability in a traditional cassava (*Manihot esculenta* Crantz) farming system, using AFLP markers. *Heredity* 85, 219–230. doi: 10.1046/j.1365-2540.2000.00749.x

Food and Agriculture Organization of the United Nations (2013). *Save and Grow: Cassava*. Rome.

Food and Agriculture Organization of the United Nations (2014). *Biannual Report on Global Food Markets.* Rome.

Food and Agriculture Organization of the United Nations (2018). *FAOSTAT Statistics Database.* Available online at: http://www.fao.org/faostat/en/#home (accessed March 16, 2020).

Food and Agriculture Organization of the United Nations and International Crops Research Institute for the Semi-Arid Tropics (1996). *The World Sorghum and Millet Economies: Facts, Trends and Outlook.*

Ford, A., and Nigh, R. (2016). *The Maya Forest Garden: Eight Millennia of Sustainable Cultivation of the Tropical Woodlands*. Abingdon: Routledge.

Foster, C. S. P., Sauquet, H., van Der Merwe, M., McPherson, H., Rossetto, M., and Ho, S. Y. W. (2017). Evaluating the impact of genomic data and priors on Bayesian estimates of the angiosperm evolutionary timescale. *Syst. Biol.* 66, 338–351. doi: 10.1093/sysbio/syw086

Fujino, K., and Sekiguchi, H. (2005). Mapping of QTLs conferring extremely early heading in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* 111:393–398. doi: 10.1007/s00122-005-2035-3

Fuller, D. Q., Murphy, C., Kingwell-Banham, E., Castillo, C. C., and Naik, S. (2019). *Cajanus cajan* (L.) Millsp. origins and domestication: the South and Southeast Asian archaeobotanical evidence. *Genet. Resour. Crop Evol.* 66, 1175–1188. doi: 10.1007/s10722-019-00774-w

Galvis, S. (2019). Colombia confirms that dreaded fungus has hit its banana plantations. *Science*. doi: 10.1126/science.aaz1033. [Epub ahead of print].

Garnett, S. T., Burgess, N. D., Fa, J. E., Fernández-Llamazares, Á., Molnár, Z., Robinson, C. J., et al. (2018). A spatial overview of the global importance of indigenous lands for conservation. *Nat. Sustain.* 1, 369–374. doi: 10.1038/s41893-018-0100-6

Gaunitz, C., Fages, A., Hanghøj, K., Albrechtsen, A., Khan, N., Schubert, M., et al. (2018). Ancient genomes revisit the ancestry of domestic and Przewalski's horses. *Science* 360, 111–114. doi: 10.1126/science.aao3297

Goland, C., and Bauer, S. (2004). When the apple falls close to the tree: local food systems and the preservation of diversity. *Renew. Agric. Food Syst.* 19, 228–236. doi: 10.1079/RAFS200487

Gómez-Ariza, J., Galbiati, F., Goretti, D., Brambilla, V., Shrestha, R., Pappolla, A., et al. (2015). Loss of floral repressor function adapts rice to higher latitudes in Europe. *J. Exp. Bot.* 66, 2027–2039. doi: 10.1093/jxb/erv004

Gutaker, R. M., and Burbano, H. A. (2017). Reinforcing plant evolutionary genomics using ancient DNA. *Curr. Opin. Plant Biol.* 36, 38–45. doi: 10.1016/j.pbi.2017.01.002

Gutaker, R. M., Weiß, C. L., Ellis, D., Anglin, N. L., Knapp, S., Luis Fernández-Alonso, J., et al. (2019a). The origins and adaptation of European potatoes reconstructed from historical genomes. *Nat. Ecol. Evol.* 3, 1093–1101. doi: 10.1038/s41559-019-0921-3

Gutaker, R. M., Zaidem, M., Fu, Y. B., Diederichsen, A., Smith, O., Ware, R., et al. (2019b). Flax latitudinal adaptation at LuTFL1 altered architecture and promoted fiber production. *Sci. Rep.* 9:976. doi: 10.1038/s41598-018-37086-5

Hammer, K., Arrowsmith, N., and Gladis, T. (2003). Agrobiodiversity with emphasis on plant genetic resources. *Naturwissenschaften* 90, 241–250. doi: 10.1007/s00114-003-0433-4

Hoffmann, T., Lyons, N., Miller, D., Diaz, A., Homan, A., Huddlestan, S., et al. (2016). Engineered feature used to enhance gardening at a 3800-year-old site on the pacific Northwest coast. *Sci. Adv.* 2:e1601282. doi: 10.1126/sciadv.1601282

Hublin, J. J. (2015). Paleoanthropology: how old is the oldest human? *Curr. Biol.* 25, R453–R455. doi: 10.1016/j.cub.2015.04.009

Jones, M., Hunt, H., Kneale, C., Lightfoot, E., Lister, D., Liu, X., et al. (2016). Food globalisation in prehistory: the agrarian foundations of an interconnected continent. *J. Br. Acad.* 4, 73–87. doi: 10.5871/jba/004.073

Joshi, B. K., and Joshi, M. (2002). A field assessment of finger millet (Kodo) diversity in Sankhuwasava District, Nepal. *Conserv. Newsl. (RRN)* 2, 3–5.

Khoury, C. K., Bjorkman, A. D., Dempewolf, H., Ramirez-Villegas, J., Guarino, L., Jarvis, A., et al. (2014). Increasing homogeneity in global food supplies and the implications for food security. *Proc. Natl. Acad. Sci.* 111, 4001–4006. doi: 10.1073/pnas.1313490111

Kistler, L., Maezumi, S. Y., Gregorio de Souza, J., Przelomska, N. A. S., Malaquias Costa, F., Smith, O., et al. (2018). Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. *Science* 362, 1309–1313. doi: 10.1126/science.aav0207

Kistler, L., Newsom, L. A., Ryan, T. M., Clarke, A. C., Smith, B. D., and Perry, G. H. (2015). Gourds and squashes (*Cucurbita* spp.) adapted to megafaunal extinction and ecological anachronism through domestication. *Proc. Natl. Acad Sci. U.S.A.* 112, 15107–15112. doi: 10.1073/pnas.1516109112

Kono, T. J. Y., Fu, F., Mohammadi, M., Hoffman, P. J., Liu, C., Stupar, R. M., et al. (2016). The role of deleterious substitutions in crop genomes. *Mol. Biol. Evol.* 33, 2307–2317. doi: 10.1093/molbev/msw102

Kremling, K. A. G., Chen, S.-Y., Su, M.-H., Lepak, N. K., Romay, M. C., Swarts, K. L., et al. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* 555, 520–523. doi: 10.1038/nature25966

Krupnick, G. A., Kress, W. J., and Wagner, W. L. (2009). Achieving target 2 of the global strategy for plant conservation: building a preliminary assessment of vascular plant species using data from herbarium specimens. *Biodivers. Conserv.* 18, 1459–1474. doi: 10.1007/s10531-008-9494-1

Larbey, C., Mentzer, S. M., Ligouis, B., Wurz, S., and Jones, M. K. (2019). Cooked starchy food in hearths ca. 120 kya and 65 kya (MIS 5e and MIS 4) from Klasies River Cave, South Africa. *J. Hum. Evol.* 131, 210–227. doi: 10.1016/j.jhevol.2019.03.015

Lightfoot, E., Liu, X., and Jones, M. K. (2013). Why move starchy cereals? A review of the isotopic evidence for prehistoric millet consumption across Eurasia. *World Archaeol.* 45, 574–623. doi: 10.1080/00438243.2013.852070

Lister, D. L., Thaw, S., Bower, M. A., Jones, H., Charles, M. P., Jones, G., et al. (2009). Latitudinal variation in a photoperiod response gene in European barley: insight into the dynamics of agricultural spread from "historic" specimens. *J. Archaeol. Sci.* 36, 1092–1098. doi: 10.1016/j.jas.2008.12.012

Liu, Q., Zhou, Y., Morrell, P. L., and Gaut, B. S. (2017). Deleterious variants in Asian rice and the potential cost of domestication. *Mol. Biol. Evol.* 34, 908–924. doi: 10.1093/molbev/msw296

Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L., and Hernández-Hernández, T. (2015). A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol.* 207, 437–453. doi: 10.1111/nph.13264

Makondo, C. C., and Thomas, D. S. G. (2018). Climate change adaptation: linking indigenous knowledge with western science for effective adaptation. *Environ. Sci. Policy* 88, 83–91. doi: 10.1016/j.envsci.2018.06.014

Martín, J. L., Cardoso, P., Arechavaleta, M., Borges, P. A. V., Faria, B. F., Abreu, C., et al. (2010). Using taxonomically unbiased criteria to prioritize resource allocation for oceanic island species conservation. *Biodivers. Conserv.* 16, 1659–1682. doi: 10.1007/s10531-010-9795-z

Mascher, M., Schuenemann, V. J., Davidovich, U., Marom, N., Himmelbach, A., Hübner, S., et al. (2016). Genomic analysis of 6000-year-old cultivated grain illuminates the domestication history of barley. *Nat. Genet.* 48, 1089–1093. doi: 10.1038/ng.3611

Massawe, F., Mayes, S., and Cheng, A. (2016). Crop diversity: an unexploited treasure trove for food security. *Trends Plant Sci.* 21, 365–368. doi: 10.1016/j.tplants.2016.02.006

Maxmen, A. (2019). CRISPR might be the banana's only hope against a deadly fungus. *Nature* 574:15. doi: 10.1038/d41586-019-02770-7

Maxted, N., and Dulloo, E. (2017). *IUCN SSC Crop Wild Relative Specialist Group*. Available online at: https://www.iucn.org/sites/dev/files/2016-2017_crop_wild_relative_sg_report.pdf (accessed March 16, 2020).

Mijatović, D., van Oudenhoven, F., Eyzaguirre, P., and Hodgkin, T. (2013). The role of agricultural biodiversity in strengthening resilience to climate change: towards an analytical framework. *Int. J. Agric. Sustain.* 11, 95–107. doi: 10.1080/14735903.2012.691221

Motuzaite-Matuzeviciute, G., Staff, R. A., Hunt, H. V., Liu, X., and Jones, M. K. (2013). The early chronology of broomcorn millet (*Panicum miliaceum*) in Europe. *Antiquity*. 338:1073. doi: 10.1017/S0003598X00049875

Muehlbauer, M. F., Honig, J. A., Capik, J. M., Vaiciunas, J. N., and Molnar, T. J. (2014). Characterization of eastern filbert blight-resistant hazelnut germplasm using microsatellite markers. *J. Am. Soc. Hortic. Sci.* 139, 399–432. doi: 10.21273/JASHS.139.4.399

Mueller, N. G., Fritz, G. J., Patton, P., Carmody, S., and Horton, E. T. (2017). Growing the lost crops of eastern North America's original agricultural system. *Nat. Plants* 3:17092. doi: 10.1038/nplants.2017.92

Nabhan, G. P. (2002). *Enduring Seeds: Native American Agriculture and Wild Plant Conservation*. Tuscon, AZ: University of Arizona Press.

Nesbitt, M. (1998). Wheat domestication: archaeobotanical evidence. *Science* 279:1431. doi: 10.1126/science.279.5356.1431e

Nesbitt, M. (2014). "Chapter 22. Use of herbarium specimens in ethnobotany," in *Curating Biocultural Collections: A Handbook* (Richmond: Royal Botanic Gardens, Kew), 313–328.

Nguyen, N. V. (2005). *Global Climate Changes and Rice Food Security*. Rome: International Rice Commission, FAO.

Nisterlberger, H. M., Smith, O., Wales, N., and Boessenkool, S. (2016). The efficacy of high-throughput sequencing and target enrichment on charred archaeobotanical remains. *Sci. Rep.* 6:37347. doi: 10.1038/srep37347

Oregon State University. (2019). *Hazelnut Breeding Program*. Available online at: https://plantbreeding.oregonstate.edu/plantbreeding/research/hazelnut-breeding-program (accessed December 13, 2019).

Palmer, S. A., Clapham, A. J., Rose, P., Freitas, F. O., Owen, B. D., Beresford-Jones, D., et al. (2012). Archaeogenomic evidence of punctuated genome evolution in gossypium. *Mol. Biol. Evol.* 29, 2031–2038. doi: 10.1093/molbev/mss070

Paris, H. S. (2015). Origin and emergence of the sweet dessert watermelon, *Citrullus lanatus. Ann. Bot.* 116, 133–148. doi: 10.1093/aob/mcv077

Paris, H. S. (2016). Overview of the origins and history of the five major cucurbit crops: issues for ancient DNA analysis of archaeological specimens. *Veg. Hist. Archaeobot.* 25, 405–414. doi: 10.1007/s00334-016-0555-1

Pin, P. A., Zhang, W., Vogt, S. H., Dally, N., Büttner, B., Schulze-Buxloh, G., et al. (2012). The role of a pseudo-response regulator gene in life cycle adaptation and domestication of beet. *Curr. Biol.* 22, 1095–1101. doi: 10.1016/j.cub.2012.04.007

Piperno, D. R. (2017). Assessing elements of an extended evolutionary synthesis for plant domestication and agricultural origin research. *Proc. Natl. Acad. Sci. U.S.A.* 114, 6429–6437. doi: 10.1073/pnas.1703658114

Pont, C., Wagner, S., Kremer, A., Orlando, L., Plomion, C., and Salse, J. (2019). Paleogenomics: Reconstruction of plant evolutionary trajectories from modern and ancient DNA. *Genome Biol.* 20:29. doi: 10.1186/s13059-019-1627-1

Purugganan, M. D., and Fuller, D. Q. (2009). The nature of selection during plant domestication. *Nature* 457, 843–848. doi: 10.1038/nature07895

Purugganan, M. D., and Fuller, D. Q. (2011). Archaeological data reveal slow rates of evolution during plant domestication. *Evolution* 65, 171–183. doi: 10.1111/j.1558-5646.2010.01093.x

Ramos-Madrigal, J., Runge, A. K. W., Bouby, L., Lacombe, T., Samaniego Castruita, J. A., Adam-Blondon, A. F., et al. (2019). Palaeogenomic insights into the origins of French grapevine diversity. *Nat. Plants*. 5, 595–603. doi: 10.1038/s41477-019-0437-5

Ramos-Madrigal, J., Smith, B. D., Moreno-Mayar, J. V., Gopalakrishnan, S., Ross-Ibarra, J., Gilbert, M. T. P., et al. (2016). Genome sequence of a 5310-year-old maize cob provides insights into the early stages of maize domestication. *Curr. Biol.* 26, 3195–3201. doi: 10.1016/j.cub.2016.09.036

Ramu, P., Esuma, W., Kawuki, R., Rabbi, I. Y., Egesi, C., Bredeson, J. V., et al. (2017). Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* 49, 959–963. doi: 10.1038/ng.3845

Renner, S. S., Pérez-Escobar, O. A., Silber, M. V., Nesbitt, M., Preick, M., Hofreiter, M., et al. (2019). A 3500-year-old leaf from a Pharaonic tomb reveals that New Kingdom Egyptians were cultivating domesticated watermelon. *bioRxiv*. doi: 10.1101/642785

Royal Botanic Gardens Kew (2016). *State of the World's Plants 2016*. Available online at: https://stateoftheworldsplants.org/2016/ (accessed December 13, 2019).

Saville, A. C., Martin, M. D., and Ristaino, J. B. (2016). Historic late blight outbreaks caused by a widespread dominant lineage of *Phytophthora infestans* (Mont.) de Bary. *PLoS ONE* 11:e0168381. doi: 10.1371/journal.pone.0168381

Scott, M. F., Botigué, L. R., Brace, S., Stevens, C. J., Mullin, V. E., Stevenson, A., et al. (2019). A 3000-year-old Egyptian emmer wheat genome reveals dispersal and domestication history. *Nat. Plants*. 5, 1120–1128. doi: 10.1038/s41477-019-0534-5

Shelef, O., Weisberg, P. J., and Provenza, F. D. (2017). The value of native plants and local production in an era of global agriculture. *Front. Plant Sci.* 8:2069. doi: 10.3389/fpls.2017.02069

Shiva, V. (1992). *The Violence of the Green Revolution: Third World Agriculture, Ecology and Politics*. London, UK: Zed Books Ltd.

Smith, B. D., and Yarnell, R. A. (2009). Initial formation of an indigenous crop complex in eastern North America at 3800 BP. *Proc. Natl. Acad. Sci. U.S.A.* 106, 6561–6566. doi: 10.1073/pnas.0901846106

Smith, O., Nicholson, W. V., Kistler, L., Mace, E., Clapham, A., Rose, P., et al. (2019). A domestication history of dynamic adaptation and genomic deterioration in *Sorghum. Nat. Plants*. 5, 369–379. doi: 10.1038/s41477-019-0397-9

Solly, M. (2019). *A Banana-Destroying Fungus Has Arrived in the Americas*. Smithson Mag. Available online at: https://www.smithsonianmag.com/smart-news/banana-destroying-fungus-has-arrived-americas-180972892/.

Stepp, J. R., and Thomas, M. B. (2007). "Managing ethnopharmacological data: herbaria, relational databases, literature," in *Ethnopharmacology in Encyclopedia of Life Support Systems (EOLSS)*, eds E. Elisabetsky and N. L. Etkin (UNESCO).

Swarts, K., Gutaker, R. M., Benz, B., Blake, M., Bukowski, R., Holland, J., et al. (2017). Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. *Science* 357, 512–515. doi: 10.1126/science.aam9425

Tripathi, L., Ntui, V. O., and Tripathi, J. N. (2019). Application of genetic modification and genome editing for developing climate-smart banana. *Food Energy Secur.* 8:e00168. doi: 10.1002/fes3.168

Tscharntke, T., Clough, Y., Wanger, T. C., Jackson, L., Motzke, I., Perfecto, I., et al. (2012). Global food security, biodiversity conservation and the future of agricultural intensification. *Biol. Conserv.* 151, 53–59. doi: 10.1016/j.biocon.2012.01.068

Turner, N. J. (2014). *Ancient Pathways, Ancestral* Knowledge: *Ethnobotany and Ecological Wisdom of Indigenous Peoples of Northwestern North America.* Montreal, QC: McGill-Queen's University Press.

Turner, R. S. (2005). After the famine: plant pathology, Phytophthora infestans, and the late blight of potatoes, 1845–1960. *Hist. Stud. Phys. Biol. Sci.* 35, 341–370. doi: 10.1525/hsps.2005.35.2.341

Vallebueno-Estrada, M., Rodríguez-Arévalo, I., Rougon-Cardoso, A., Martínez González, J., García Cook, A., Montiel, R., et al. (2016). The earliest maize from San Marcos Tehuacán is a partial domesticate with genomic evidence of inbreeding. *Proc. Natl. Acad. Sci. U.S.A.* 113, 14151–14156. doi: 10.1073/pnas.1609701113

van Heerwaarden, J., Doebley, J., Briggs, W. H., Glaubitz, J. C., Goodman, M. M., de Jesus Sanchez Gonzalez, J., et al. (2011). Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proc. Natl. Acad. Sci. U.S.A.* 108, 1088–1092. doi: 10.1073/pnas.1013011108

Wales, N., Akman, M., Watson, R. H. B., Sánchez Barreiro, F., Smith, B. D., Gremillion, K. J., et al. (2019). Ancient DNA reveals the timing and persistence of organellar genetic bottlenecks over 3000 years of sunflower domestication and improvement. *Evol. Appl.* 12, 38–53. doi: 10.1111/eva.12594

Wang, L., Beissinger, T. M., Lorant, A., Ross-Ibarra, C., Ross-Ibarra, J., and Hufford, M. B. (2017). The interplay of demography and selection during maize domestication and expansion. *Genome Biol.* 18:215. doi: 10.1186/s13059-017-1346-4

Wood, S. A., Karp, D. S., DeClerck, F., Kremen, C., Naeem, S., and Palm, C. A. (2015). Functional traits in agriculture: agrobiodiversity and ecosystem services. *Trends Ecol. Evol.* 30, 531–539. doi: 10.1016/j.tree.2015.06.013

Zeder, M. A. (2017). Domestication as a model system for the extended evolutionary synthesis. *Interface Focus* 7:20160133. doi: 10.1098/rsfs.2016.0133

Zilberstein, A. (2015). Inured to empire: wild rice and climate change. *William Mary Q.* 72, 127–158. doi: 10.5309/willmaryquar.72.1.0127

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a past co-authorship with one of the authors, LK.

frontiers in Ecology and Evolution

REVIEW
published: 31 March 2020
doi: 10.3389/fevo.2020.00083

# Kinship Determination in Archeological Contexts Through DNA Analysis

**Stefania Vai[1]*, Carlos Eduardo G. Amorim[2], Martina Lari[1] and David Caramelli[1]**

[1] Laboratorio di Antropologia Molecolare e Paleogenetica, Dipartimento di Biologia, Università di Firenze, Florence, Italy,
[2] Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, CA, United States

Knowing kinship relations between individuals in archeological contexts is of great importance to understand social habits and structure in past human communities. Archeological and anthropological analyses of burial sites and skeletal remains often allow us to infer connections between individuals, but only genetic analysis can provide a sound determination of kinship. Several case studies are now available in the literature that show the potentiality and limitations of different methodological approaches based on ancient DNA (aDNA). Both experimental and computational strategies for kinship estimation on ancient samples are described in this review and we argue that, within a multidisciplinary approach, kinship inference contributes to the understanding of the biological and cultural patterns that characterized past societies.

**Keywords: ancient DNA, kinship, relatedness, paleogenomics, anthropology, archeology, archeogenetics**

OPEN ACCESS

***Edited by:***
*Nathan Wales,*
*University of York, United Kingdom*

***Reviewed by:***
*Kristian Hanghoej,*
*University of Copenhagen, Denmark*
*Christine Keyser,*
*Université de Strasbourg, France*

****Correspondence:***
*Stefania Vai*
*stefania.vai@unifi.it*

***Specialty section:***
*This article was submitted to Paleoecology, a section of the journal Frontiers in Ecology and Evolution*

***Received:*** *18 November 2019*
***Accepted:*** *12 March 2020*
***Published:*** *31 March 2020*

***Citation:***
*Vai S, Amorim CEG, Lari M and Caramelli D (2020) Kinship Determination in Archeological Contexts Through DNA Analysis. Front. Ecol. Evol. 8:83. doi: 10.3389/fevo.2020.00083*

## INTRODUCTION

With the development of paleogenetic methods, both laboratorial and computational, ancient DNA (aDNA) analysis has been taking on an increasing importance in the study of archeological contexts. For instance, in the last few years, several aDNA studies have revealed past migration routes and connections between different human groups, shedding light on large-scale population dynamics of our ancestors (i.e., Allentoft et al., 2015; Haak et al., 2015; Mathieson et al., 2018; Olalde et al., 2018 and references therein). While this information is of utmost importance for reconstructing our history and understanding human evolution, evidence at a finer scale – e.g., the degree of relatedness between individuals interred in a cemetery or the existence of fine-population structure within an archeological site – have always been of crucial interest to anthropologists and archeologists. These details can reveal interesting patterns about the social structure and behavior of a community and are thus very relevant for a complete description of archeological sites (Amorim et al., 2018; Veeramah, 2018). In addition, in a practical sense, this type of information may be an important source for "storytelling" in museum presentations, outreach, and science communication.

To infer the relationship between individuals in an archeological context or between populations from the past, classic methods in physical anthropology and archeology leverage information from morphological traits (e.g., a shared hereditary disease marker) and elements of material culture. However, these data present some limitations. For instance, morphological traits for kinship determination are mostly represented by non-metric traits that are not commonly retrieved in archeological studies, and often do not have enough resolution to detect close relationships between pairs of individuals. This happens because a polygenic genetic architecture usually underlies these traits and environmental factors can affect their expression (Alt and Vach, 1998; Hassett, 2006; Ricaut et al., 2010; Stojanowski and Hubbard, 2017). In addition, it is possible that two unrelated individuals share a morphological trait by chance. For these reasons, kinship in some cases can at most be hypothesized considering the spatial organization of graves, relative position of burials, connections of grave goods, and age classes. In this scenario genetic data is the only way to reach a sound determination of kinship relations.

In this review, we describe molecular strategies for kinship estimation, from the classic PCR-based methods to Next-Generation Sequencing (NGS), with an overview of the computational approaches for kinship inference using aDNA data. For the purposes of this review, "kinship" strictly denotes the biological relationship between individuals, though we acknowledge that kinship in archeology and anthropology encompasses a much broader range of social relationships.

## FIRST APPROACHES TO KINSHIP ANALYSIS

The first attempts to infer kinship in ancient individuals by genetic analysis focused on a limited number of loci that were genotyped with methods based on PCR (Polymerase Chain Reaction). In these early studies, mostly mitochondrial DNA (mtDNA) fragments were used as target (Mooder et al., 2005; Rudbeck et al., 2005). Hypervariable Region I and II (HVR-I, HVR-II) of the mtDNA and Single Nucleotide Polymorphisms (SNPs) in its Coding Region are usually typed allowing one to define individual profiles and haplogroups, which are then used to establish a possible maternal relationship between samples. Usually mtDNA markers are amplified using custom primers; Coding Region SNPs are also detected by restriction enzymes.

An advantage of using mtDNA over other types of genetic markers is that it is available in a high amount in the cell and thus its amplification success in degraded samples is generally higher than that for nuclear loci (Bouwman et al., 2008; Baca et al., 2012; Deguilloux et al., 2014; Cui et al., 2015; Esparza et al., 2017). In spite of that, attempts to analyze nuclear DNA are common, as they provide higher resolution in kinship estimates than mtDNA lineages alone. Two types of markers are commonly used in such studies: fast evolving, multi-allelic genetic markers known as short-tandem repeats (STRs) or microsatellites and SNPs. To establish relationships between father and son, STRs and SNPs on the Y-chromosome can be used, while autosomal STRs are used more broadly to infer this and other types of family relationships (Haak et al., 2008; Gamba et al., 2011; Baca et al., 2012; Alt et al., 2016).

Commercial forensic kits are usually used for the amplification of STRs in archeological studies. One example of such kits is the *AmpFl*STR1Y-Filer PCR Amplification kit (Thermo Fisher) (Mulero et al., 2006), which has been used, for instance, to type Y-chromosome STRs in individuals from a 7th century burial-place in Germany, together with custom primers designed for shorter amplicons (Vanek et al., 2009). In addition to Y-chromosome markers, this study also analyzed autosomal STRs, which were amplified with the commercial kits *AmpFl*STR1Identifiler (Wang et al., 2011) and *AmpFl*STR1MiniFiler PCR Amplification Kit (Thermo Fisher) (Mulero et al., 2008). The same kits also allowed us to reconstruct the genealogy in Tutankhamun's family (Hawass et al., 2010). These are just a couple of examples of studies that used forensic kits to infer kinship in an archeological context. The literature bears some other examples of these such as: *AmpFl*STR NGM SElect$^{TM}$ PCR Amplification Kit (Green et al., 2013) (Thermo Fisher) and PowerPlex® ESX SYSTEM (ESX) (Sprecher et al., 2009) (Promega) for the analysis of a Bronze Age pit burial in Spain (Esparza et al., 2017; Palomo-Dìez et al., 2018); *AmpFl*STR Profiler Plus kit (Thermo Fisher) for samples from Corded Ware Culture burials in Germany (Haak et al., 2008); and finally the AGCU mini STR Kit (AGCU ScienTech, China) to determine kinship relations in Mongolian noble burials from the beginning of the fourteenth century (Cui et al., 2015). When different kits for analysis of autosomal STRs are used on the same samples, the *AmpFl*STR1MiniFiler PCR Amplification Kit shows a higher rate of amplification success in terms of typed loci and number of samples (Vanek et al., 2009). This kit is indeed specifically designed for degraded DNA, producing short amplicons (71–250 bp), also called miniSTRs (Butler et al., 2003; Nastainczyk et al., 2009), and remains the best strategy to obtain complete or almost complete autosomal profiles in ancient samples (Gamba et al., 2011).

The first study to reconstruct kinship relations in an archeological context using maternal, paternal, and biparental markers focused on 62 individuals from the Egyin Gol necropolis in Mongolia (Keyser-Tracqui et al., 2003). The site was dated from the 3rd century BCE to the 2nd century CE and is associated with the Xiongnu period. The skeletal material was well preserved and climatic conditions of the area are favorable to DNA preservation. Genetic analysis was based on autosomal STRs first, because of their high discriminatory power in kinship inference, then on Y-chromosome and mtDNA HVR-I. Multiple amplifications for each marker on independent DNA extracts of the same specimen were performed to authenticate the results. Nine autosomal STRs and sex determination markers were amplified using the AmpFlSTR profiler Plus kit (Thermo Fisher). Samples belonging to a putative family were analyzed for 10 additional loci using the AmpFlSTR SGM Plus kit (Applied Biosystems). Forty-nine partially complete profiles were obtained. Eight Y-chromosome STR markers were amplified using the Y-Plex6 kit (ReliaGene Technologies) and custom primers, and 27 out of 35 male samples were typed for at least three loci. An inverse proportionality between amplification success and size of the amplified fragment was shown. mtDNA HVR-I was amplified with a different combination of custom primers for 46 samples out of 56. After pairwise comparison of the profiles, it was possible to detect close relationships between several samples: for example, one parentage trio, mother/father/child, was found; nine possible parent-child relationships and three siblings were also identified. The finding of male relatives buried in-group allowed us to better understand the funeral practices of Xiongnu people from the Egyin Gol necropolis in Mongolia. These data, along with archeological and chronological characterization, provided important information about the social history of the necropolis. When kinship estimates based on genetic data were compared to those based on osteological non-metric traits, there appeared to be a correlation between them, even though the number of relationships detected by non-metric trait analysis was 50% lower than those highlighted by genetic markers, confirming the importance of molecular analysis to detect close relationships between individuals (Ricaut et al., 2010).

The high success rate of amplification for the Egyin Gol samples, along with the high number of individuals analyzed, still represents a *unicum* in the PCR-based kinship studies. PCR-based approaches are indeed often characterized by technical problems that could lead to partial or wrong results. The degree of DNA preservation strongly influences the success of the analysis: low copy number of amplifiable DNA, molecule damage and fragmentation, and the presence of PCR inhibitors can determine no results or produce incomplete profiles for some samples. Except for a few cases of well-preserved samples (as seen in the study of the Egyin Gol necropolis), in most cases nuclear data are limited to few individuals and to partial profiles. Furthermore, when nuclear DNA is available in low amounts, allelic dropout in autosomal loci can occur and lead to false homozygous profiles (Haak et al., 2008; Palomo-Dìez et al., 2018). For instance, the analysis of four multiple burials in Eulau, Germany, attributed to the Corded Ware Culture, shows a typical pattern for ancient nuclear DNA: amplification success inversely correlated to the length of loci and alleles and loci dropout. Only 3 out of 12 individuals (25%) yield reliable results for four/five autosomal STRs loci. On the other hand, the success rate with mtDNA is high (75%), with 9 samples genotyped for HVR-I (Haak et al., 2008). Consequently, most of the kinship studies are exclusively [for example, the analysis of a Late Neolithic megalithic tomb in Alto de Reinoso, Spain, (Alt et al., 2016) and of a Merovingian necropolis in France (Deguilloux et al., 2014)] or mainly based on mtDNA, because, as explained above, it is present in higher amounts than nuclear DNA, although it yields limited information, restricted to maternal relations, and does not allow one to obtain a complete reconstruction of possible relationships. Furthermore, kinship estimates using mtDNA should be supported by proper evaluation of the significance of match, since identical haplotypes can be carried by unrelated individuals (Just et al., 2009). Even with mtDNA, the success rate can vary greatly because of micro-environmental conditions that lead to a different level of DNA degradation and presence of inhibitor substances: from more than 90% of the individuals successfully genotyped in an early Danish Christian Cemetery that were analyzed (Rudbeck et al., 2005) and from Alto de Reinoso (Alt et al., 2016), to only 4 samples out of 22 from a grave circle in Mycenae (18%) (Bouwman et al., 2008).

To improve the experimental performance, specific silica-based extraction protocols can optimize DNA recovery in highly degraded samples and overcome inhibition problems. Success in amplification is usually improved by using custom primers or commercial kits set up for short amplicons (Gamba et al., 2011). A further problem that characterizes the PCR approach is the authentication of the result, which can only be attested with difficulty especially in highly manipulated specimens. Presence of exogenous human contamination can produce false negative results for kinship attribution as well as leading to wrong matches in case of contamination from the same source spreading to more samples. To exclude possible contaminations, some precautions are generally considered: data for an individual are retained if coming from multiple independent DNA extracts, having phylogenetic sense, and differing from researchers' profiles. Cloning of PCR products and sequencing multiple clones are an efficient strategy to detect contamination and to observe possible nucleotide misincorporations due to post-mortem damage (Briggs et al., 2007; Brotherton et al., 2007) that represent a further indication of authenticity of the result (Rudbeck et al., 2005; Haak et al., 2008). As previously mentioned, an inverse relationship between amplification efficiency and size of the amplicons is also characteristic of authentic ancient data.

To overcome most of the limitations of the PCR-based approach, in recent years Next Generation Sequencing (NGS) methods have been applied to aDNA. Thanks to the primers-independent strategy, very short molecules can be recovered and target sequences can be reconstructed even if the DNA is highly fragmented and damaged and from samples that could not be analyzed by PCR. High-throughput sequencing and enrichment strategies allow one to obtain data also in case of a very low amount of DNA. With the NGS approach it is possible to dramatically increase the number of loci and individuals successfully typed and consequently to obtain higher-resolution kinship estimates and more complete reconstruction of past societies.

## NGS METHODS

Since 2005, NGS methods have started to be used in aDNA research, providing several benefits in the study of degraded samples. High-throughput sequencing platforms generate data from billions of DNA fragments per sequencing run, with a fast time and cost-efficient data production.

The sample preparation strategy allows one to preserve and analyze the original characteristics of degradation of the DNA molecules (Ginolhac et al., 2011; Jónnson et al., 2013), improving the possibility of detecting possible contamination and of authenticating the results. Even very short molecules (<50 bp), that are not analyzable by PCR, can be recovered and sequenced by NGS. For these reasons, samples that are not suitable for PCR analysis because of degradation and contamination can often be analyzed through NGS experiments and yield important results. The advantages of NGS methods also lie in the number of analyzable samples and loci that can be sequenced. Furthermore, accompanying NGS with the choice of the skeletal element that can provide the highest amount of endogenous DNA, it is possible to dramatically increase the informative power of ancient samples, allowing one to sequence even entire genomes. Recently, the petrous part of the temporal bone was identified as the best source for aDNA (Gamba et al., 2014; Pinhasi et al., 2015). Thanks to this knowledge, it is now possible to obtain a high number of comparable loci for several samples and to assess kinship relations with increased resolution, going deeper in the evaluation of the degree of relationship between individuals. With this approach, a new era for kinship analysis in archeological contexts has started.

Sample preparation for NGS consists of adding universal oligonucleotide adapters and specific indexing sequences (barcodes) to the extracted DNA molecules, producing the so called NGS library. Specific protocols have been developed for ancient samples instead of commercial library preparation kits to improve sequence retrieval even in case of low DNA amounts and to take into account characteristics due to degradation. The protocol proposed by Meyer and Kircher (2010) for double-stranded DNA libraries is commonly used in aDNA studies. For highly degraded samples a method was also developed suitable to recover single-stranded DNA (Gansauge and Meyer, 2013). Uracil-DNA-glycosylase (UDG) treatment can be used to reduce the occurrence of nucleotide misincorporations that can lead to false mismatches to the reference genome in the final sequence (Briggs et al., 2010). The protocol most commonly used in recent years provides for a partial UDG treatment that preserves a damage signal at the terminal nucleotides useful for validating the authenticity of the result, while nearly eliminating misincorporations in the interior of the molecule in order to increase confidence in SNPs calling (Rohland et al., 2015). A target-enrichment strategy can be associated with NGS to improve sequencing depth on particular loci of interest. This approach is usually followed when the sample is characterized by a low percentage of endogenous DNA (generally lower than 30%), and it is usually conducted by in-solution capture with DNA probes. The whole mitochondrial genome is generally captured using custom made PCR products as probes (Maricic et al., 2010). More than 1 million informative SNPs are used as target on the nuclear genome. Probes are usually designed as described in Haak et al. (2015) and Fu et al. (2015), and their sequences derive mostly from commercial arrays such as Affymetrix Human Origins SNP (Patterson et al., 2012). Y-chromosome probes for target enrichment have been designed for aDNA (Cruz-Davalos et al., 2018), but not yet used specifically for kinship analysis.

Through target enrichment strategies, it was possible to reveal the absence of maternal kinship in the Neolithic site of Çatalhöyük with the study of mtDNA whole genomes (Chylénski et al., 2019). Possible maternal and paternal relationships were found in ten necropolises of the Avar period (7th–8th century CE) in the Carpathian Basin thanks to the enrichment of whole mtDNA genomes and Y chromosome STRs amplified with AmpFLSTR Yfiler PCR Amplification Kit (Thermo Fisher Scientific) (Csàky et al., 2019). Thanks to mtDNA and nuclear SNPs enrichment, a matrilineal dynasty was found at Pueblo Bonito in Chaco Canyon, Mexico, between 800 and 1130 CE, associated with one of North America's earliest complex societies (Kennett et al., 2016).

Samples with a high amount of endogenous DNA can be successfully sequenced with a shotgun approach. With Whole-Genome Sequencing (WGS), the kinship relations recognized in an Early Medieval Alemannic graveyard dated to the 7th century CE, together with archeological data, highlighted that closely related individuals showed different cultural characteristics (O'Sullivan et al., 2018).

A recent study focused on 24 individuals from 5 megalithic tombs of the fourth millennium BCE, located in northern and western Europe. First-degree (parent-offspring, full siblings) and second-degree (half-siblings, grandparent-grandchild, aunt/uncle-niece/nephew) kinship relations between individuals buried in the same as well as in different megaliths were identified and an association of these monuments with patrilineal kindred group were found. These data provided important information on the social dynamics of the megalithic culture (Sanchez-Quinto et al., 2019). The analysis of a Late Neolithic mass grave from Poland, associated with the Globular Amphora Culture, demonstrated that the 15 buried individuals belonged to the same extended family. The relative position of the bodies was in accordance with kin relationship, revealing that, after a violent death, someone who knew these people, took care of their burial (Schroeder et al., 2019).

To illustrate how kinship inference in high-density palaeogenomic data can inform anthropological and archeological studies, here we compare the studies of Amorim et al. (2018) and Mittnik et al. (2019). Amorim et al. (2018) obtained ancient genomic DNA, thanks to a combination of target enrichment and whole genome sequencing, from 63 samples from two 6th century CE cemeteries associated with the Longobard culture, Collegno in North Italy and Szólád in western Hungary. This article presented the largest sample size for a single archeological site (Szólád; $N = 39$) among all studies at genomic level in human aDNA at the time it was published. Mittnik et al., 2019, published a year later, also presented high-resolution genetic data for prehistorical human societies and, similarly to Amorim et al. (2018), these authors combined genetic data for 1.2M SNPs with archeological and isotopic data. Mittnik et al. (2019) focused on populations in the Lech Valley in Germany, associated with the Corded Ware Culture (∼2750–2460 BCE), the Bell Beaker Complex (∼2480–2150 BCE) and the Bronze Age (∼2150–1300 BCE), genotyping 104 individuals.

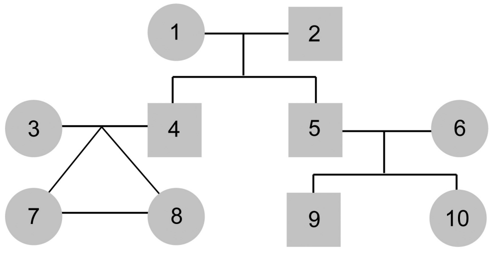
Both studies (Amorim et al., 2018; Mittnik et al., 2019) assessed aspects of societies that could only be visible through the lenses of high-resolution datasets. Although chronologically distantly related (at least 19 centuries apart), the organization of the burial sites presented striking similarities. First, with one exception (one pair in the Lech Valley), all first and second-degree relationships were found between individuals buried in the same burial site. In total, both studies describe 11 large pedigrees, sometimes including four generations, with a clear lack of female individuals in specific family groups. This feature possibly reflects female exogamy, higher migration rates for females, differential mortuary practices for females, or a combination of the three. In both studies, the authors describe a higher prevalence of grave goods in burials belonging to multi-generational families. Grave goods were also often found in graves with a certain wooden structure (as opposed to simple pits). Finally, members of these multi-generational families were often buried next to each other. All these features were observed in cemeteries that are chronologically distant (at least 1,900 years apart). The deep characterization of kinship relationships and fine population structure in association with archeological data analysis that both articles implement represents a novel analytical paradigm in aDNA studies.

Published studies based on NGS data specifically focused on kinship inference using ancient samples are still limited in number, but recent years are characterized by an increasing interest in developing specific experimental and data analysis strategies, to take advantage of the high informative power of this kind of data for understanding past societies. In particular, specific analytic approaches are followed to overcome the problem of low coverage data and the difficulty of reconstructing diploid genomes for degraded samples.

## COMPUTATIONAL METHODS FOR KINSHIP INFERENCE

Two individuals that are biologically related share alleles that are identical-by-descent. Because it is not possible to directly assess identity-by-descent (IBD) using population-based genetic datasets, methods for inferring biological relatedness in population studies rely on estimates of the probability of IBD between genetic variants. The probabilities of IBD, in turn, are calculated based on the observed fraction of the genome of two individuals that are identical-by-state. In this section, we will first discuss general aspects related to kinship inference using population genetics data (i.e., without prior knowledge of the pedigree) and then we will discuss some issues related to kinship inference using aDNA data.

To estimate IBD probabilities for a locus between two individuals, one needs to know (i) the genotype sharing pattern (i.e., whether one, two or no genetic variants are shared in a given locus), and (ii) the frequency of these variants in the population. There are a few methods available in the literature that, based on these two features, can determine the kinship coefficient between two individuals (Thompson, 1975; Gusev et al., 2009; Albrechtsen et al., 2010; Manichaikul et al., 2010). There are usually two types of information obtained from these methods: (i) the inbreeding coefficient $\phi$ that describes the probability of two random alleles sampled from two individuals being identical-by-descent, and (ii) the probabilities $k_0$, $k_1$, $k_2$ of a given pair of individuals sharing, respectively zero, one or two alleles that are identical-by-descent. For instance (**Figure 1**), the expectation for non-twin siblings will be $\phi$, $k_0$, and $k_2$ equals to 0.25 each, and $k_1$ equals to 0.50 (Weir et al., 2006). The different values of $k$ are equivalent to the expected fraction of their genomes that will have zero, one or two alleles identical-by-descent, respectively, following the principles of Mendel's independent segregation law. Clearly, individuals who are distantly related will have a relatively larger value for $k_0$ and, conversely, twins will have $k_2$ equals to 1.00. Notably, if the different values of $k$ can be computed, then it is possible to infer the degree of relatedness between two individuals (**Figure 1**). In computing these probabilities across genomic loci, rare alleles (i.e., those that are seen in low frequency in the population) are especially informative, since the sharing of such rare variants is a strong indication of IBD.



| Relationship | Example | $\phi$ | $k_0$ | $k_1$ | $k_2$ |
|---|---|---|---|---|---|
| Identical Twins | 7-8 | 0.5000 | 0 | 0 | 1.00 |
| Full-siblings | 9-10 | 0.2500 | 0.25 | 0.50 | 0.25 |
| Parent-child | 1-5 | 0.2500 | 0 | 1.00 | 0 |
| Grandparent-grandchild | 1-10 | 0.1250 | 0.50 | 0.50 | 0 |
| First cousins | 8-9 | 0.0625 | 0.75 | 0.25 | 0 |

**FIGURE 1 |** Pedigree showing the degrees of relationship between selected pairs of individuals and corresponding inbreeding coefficients ($\phi$) and $k_0$, $k_1$, $k_2$ values describing the probabilities of a given pair of individuals of sharing, respectively zero, one or two alleles that are identical-by-descent.

Software packages that include tools for evaluation of kinship probability are available and were used also on ancient samples. Examples are Patcan (Riancho and Zarrabeitia, 2003), Familias (Egeland et al., 2000; Kling et al., 2014), GenoProof (Qualitype AG, Dresden), Relatedness (Goodnight and Queller, 2001), and GenAlEx (Peakall and Smouse, 2012). In some cases, it is possible to combine non-DNA evidence (for example the age of individuals) and DNA profiles calculating posterior probabilities through a Bayesian approach (Egeland et al., 2000). These tools were successfully applied in aDNA in PCR-based studies. Other bioinformatic tools, generally used with NGS data are, for example, kinship inference tools included in ANGSD (Korneliussen et al., 2014), as well as ERSA (Huff et al., 2011), REAP (Thornton et al., 2012), READ (Monroy Kuhn et al., 2018), and KING (Manichaikul et al., 2010).

KING is freely available with the genome analysis package PLINK (Purcell et al., 2007) and, like other similar methods available online, it uses as an input polymorphism data for two to several individuals and a file with population allele frequencies. Direct application of this tool to aDNA data obtained from archeological samples is, however, not possible for two reasons: (i) because diploid genotypes (i.e., the information for both alleles in a given locus) are often not possible to call in aDNA data due to low coverage sequencing; and (ii) there is usually not a good reference sample to estimate population allele frequencies.

The difficulty in calling diploid genotypes comes from the fact that aDNA is often found in very low concentrations, yielding sequencing data with really low depth of coverage, even <1x (for instance, Allentoft et al., 2015; Haak et al., 2015; Mathieson et al., 2015). If a site is covered by a single sequencing read, it is technically impossible to call diploid genotypes at that site. A strategy commonly used in palaeogenomic studies is to randomly sample one allele per SNP site, reconstructing a pseudo-haploid genome for each individual. READ allows one to infer kinship relations up to second degree starting from pseudo-haploid genotypes (Monroy Kuhn et al., 2018). A different approach to circumvent this problem (the difficulty of calling diploid genotypes) is to estimate IBD probabilities from genotype likelihoods, instead of observed genotypes. Examples of implementation using this approach are lcMLkin (Lipatov et al., 2015) and NgsRelate2 (Korneliussen and Moltke, 2015; Hanghøj et al., 2019). These methods incorporate the uncertainty in genotype calls in order to infer kinship. lcMLkin, for instance, sums the probabilities of IBD over all possible genotypes, weighted by their likelihoods, instead of using the single best genotype for the statistical inferences.

The second problem, namely, the lack of a good reference sample to estimate population allele frequencies, comes from the fact that large databases of frequencies are not available for past populations. Datasets for mtDNA and nuclear genomes from ancient populations have been growing over the years, but sample size is often limited. Furthermore, available ancient samples may not properly represent allele frequencies in the original population because of sampling biases, caused for instance by burial pattern. To estimate population allele frequencies in the absence of a reference sample, one may use a modern dataset as the reference population. In choosing the best modern reference population, one should consider populations that are historically related to the study population. However, this may not always be obvious to assess. Usually, the corresponding modern population of the same geographical area is considered as a proxy to represent the population allele frequencies; in such cases, possible differences between ancient and modern allelic frequencies should be taken into account (Vanek et al., 2009; Esparza et al., 2017; Palomo-Dìez et al., 2018). Notably, *lcMLkin* (Lipatov et al., 2015), the method mentioned above, is robust even when the considered population allele frequencies diverge from the true allele frequency. If there are multiple potential reference sets, it may be worth performing the kinship inference using different reference sets, one at a time [see, for instance, (Amorim et al., 2018)]. Moreover, depending on whether there is a large enough sample, it may be possible to estimate allele frequencies from the target set of ancient samples [see, for instance, (Amorim et al., 2018)]. On their supplementary figure S85, Amorim et al. (2018) compare kinship coefficients using modern datasets versus the target set of ancient samples. They find that using modern datasets as the reference sample yields larger kinship coefficients. With the lack of a good reference set, simple pairwise comparison may be performed considering the profiles from the samples analyzed (Keyser-Tracqui et al., 2003; Rudbeck et al., 2005; Alt et al., 2016). The latter approach could be considered appropriate especially when genetic data are required to support links that have been established based on archeological or anthropological data. If thousands of matching loci are present among samples coming from a restricted community and if other elements of connection are present, the alleged kinship relation can be considered as supported even without comparison to a large dataset. Finally, when there is no good proxy to estimate ancestral population allele frequencies, it is possible to infer kinship without a reference set for allele frequencies, such as for instance in the methods implemented by Waples et al. (2019) and Sikora et al. (2017).

As previously mentioned, in forensic routines, kinship inference is often performed with STRs. Their high mutation rates and high heterozygosis make them ideal for accurately discerning relatives. Laboratory protocols to genotype this type of marker in highly degraded specimens have been developed (Butler et al., 2003; Nastainczyk et al., 2009; Vanek et al., 2009) and are suitable for samples recovered from archeological contexts. In palaeogenomic studies, another type of marker is employed, namely Single-Nucleotide Polymorphisms (SNPs). Comparatively, ∼50 SNPs are needed to have the same informative power of ∼10 STRs (Gill, 2001; Amorim and Pereira, 2005). In the genomic era, where a few hundreds of thousands of markers are used, this is not a problem. Even for aDNA data it is common to have a few orders of magnitude more than 50 SNPs. As described in the previous sections, the availability of so many loci allows the determination of relationships of up to 4th-degree (see for instance Amorim et al., 2018; Mittnik et al., 2019) illuminating important facets of past human societies that were previously unknown.

## CONCLUSION

In the study of an archeological context, one of the questions archeologists and anthropologists are mainly interested in is the possible kinship relations between individuals. Genetic analysis of skeletal remains can support kinship estimates coming from morphological study and archeological inference and provide a sound determination even in absence of other data. Some contexts show very elaborate funerary rituals, with handling, moving, and fragmentations of the remains in secondary and multiple burials. In these cases, only an accurate molecular analysis can help the attribution of the remains to single individuals and the identification of possible relations between them, for a better understanding of past funerary practices. Molecular determination is indicative of biological kinship, but also non-biological connections between individuals can be highlighted in an archeological context. In this case, merging together genetic and archeological data is fundamental for a proper reconstruction and interpretation of social and cultural habits. Different archeological patterns can be found, that show how biological kinship does not always agree with other, non-biological forms of kinship. For example, some members of a biological family could be buried far away from their relatives and sometimes be associated with biologically unrelated individuals. Even simple contexts such as three skeletons of a male, female, and child buried together should be accurately studied, since not always is the easier interpretation of a nuclear family correct. Different cultural ideology of what a family is can underlie the burial pattern, and the common interpretation of a heteronormative nuclear family should be proven instead of simply assumed. Molecular sex and kinship estimates are sometimes essential to provide a correct interpretation of the context, as demonstrated for the Bronze Age site of Los Tolmos in Spain by Esparza et al. (2017).

Since past ideologies, habits, social structures, and rituals can be very different and sometimes unexpected, a multidisciplinary approach is strongly recommended to analyze and properly interpret archeological evidence. Data from different disciplines should always be considered and put together: spatial distribution and structure of the burials, presence/absence and distribution of grave goods studied by archeologists are necessary to formulate hypotheses of connections between individuals and for a final interpretation of the results. In this context, a detailed examination of the written record and the oral history, when available, could reveal for instance whether changes in specific cultural patterns are associated with historical events like the contact between two peoples. Anthropological characterization of age and sex, morphological traits, traces of diseases, stress markers, and cause of death provide fundamental details to investigate individual history and role. Furthermore, radiocarbon dating is necessary to understand temporal relationships between individuals and isotopic analysis reveals geographical origins and dietary conditions. All these data together with the genetic analysis for a fine reconstruction of kinship relations – integrated into a multidisciplinary approach – allow a complete description of past communities with their social structure. In this regard, works such as, for example, Amorim et al., 2018 and Mittnik et al. (2019) provide a model for future studies in archeogenetics where information from biological material, historical sources, and archeological evidence come together for a "bottom-up" characterization of past societies. As other archeological contexts in Europe and elsewhere start to be studied in this way, we will gain insight about whether this type of kin-based structures and mortuary practices are a common feature of past human societies and how these structures have evolved through time in different regions of the world.

## AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

## REFERENCES

Albrechtsen, A., Moltke, I., and Nielsen, R. (2010). Natural selection and the distribution of identity-by-descent in the human genome. *Genetics* 186, 295–308. doi: 10.1534/genetics.110.113977

Allentoft, M. E., Sikora, M., Sjögren, K. G. R., Rasmussen, S., Rasmussen, M., Stenderup, J., et al. (2015). Population genomics of bronze age Eurasia. *Nature* 522:167. doi: 10.1038/nature14507

Alt, K. W., and Vach, W. (1998). "Kinship studies in skeletal remains-concepts and examples," in *Dental Anthropology: Fundamentals, Limits, and Prospects*, eds K. W. Alt, W. Rosing, and M. Techler-Nicola (Wien: Springer Verlag), 537–554. doi: 10.1007/978-3-7091-7496-8_27

Alt, K. W., Zesch, S., Garrido-Pena, R., Knipper, C., Szécsényi-Nagy, A., Roth, C., et al. (2016). A community in life and death: the late neolithic megalithic tomb at alto de reinoso (Burgos. Spain). *PLoS One* 11:e0146176. doi: 10.1371/journal.pone.0146176

Amorim, A., and Pereira, L. (2005). Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. *Forensic Sci. Int.* 150, 17–21. doi: 10.1016/j.forsciint.2004.06.018

Amorim, C. E. G., Vai, S., Posth, C., Modi, A., Koncz, I., Hakenbeck, S., et al. (2018). Understanding 6th-century barbarian social organization and migration through paleogenomics. *Nat. Commun.* 9:3547. doi: 10.1038/s41467-018-06024-4

Baca, M., Doan, K., Sobczyk, M., Stankovic, A., and Węgleński, P. (2012). Ancient DNA reveals kinship burial patterns of a pre-Columbian Andean community. *BMC Genet.* 13:30. doi: 10.1186/1471-2156-13-30

Bouwman, A. S., Brown, K. A., Prag, A. J. N. W., and Brown, T. A. (2008). Kinship between burials from Grave Circle B at Mycenae revealed by ancient DNA typing. *J. Archaeol. Sci.* 35, 2580–2584. doi: 10.1016/j.jas.2008.04.010

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., et al. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U.S.A.* 104, 14616–14621. doi: 10.1073/pnas.0704665104

Briggs, A. W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., and Paabo, S. (2010). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acid Res.* 38:87. doi: 10.1093/nar/gkp1163

Brotherton, P., Endicott, P., Sanchez, J. J., Beaumont, M., Barnett, R., Austin, J., et al. (2007). Novel high-resolution characterization of ancient DNA reveals C (U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Res.* 35, 5717–5728. doi: 10.1093/nar/gkm588

Butler, J., Shen, Y., and McCord, B. (2003). The Development of Reduced Size STR amplicons as tools for analysis of degraded DNA. *J. Forensic Sci.* 48, 1054–1064.

Chylénski, M., Ehler, E., Somel, M., Yaka, R., Krzewinska, M., Dabert, M., et al. (2019). Ancient mitochondrial genomes reveal the absence of maternal kinship in the burials of Çatalhöyük people and their genetic affinities. *Genes* 10:207. doi: 10.3390/genes10030207

Cruz-Davalos, D. I., Nieves-Colon, M. A., Sockell, A., Poznik, G. D., Schroeder, H., Stone, A. C., et al. (2018). In-solution Y-chromosome capture-enrichment on ancient DNA libraries. *BMC Genomics* 19:608. doi: 10.1186/s12864-018-4945-x

Csàky, V., Gerber, D. N., Koncz, I., Csiky, G., Mende, B. G., Szeifert, B., et al. (2019). Genetic insights into the social organisation of the Avar period elite in the 7th century AD Carpathian Basin. *Sci. Rep.* 10, 948. doi: 10.1038/s41598-019-57378-8

Cui, Y., Song, L., Wei, D., Pang, Y., Wang, N., Ning, C., et al. (2015). Identification of kinship and occupant status in Mongolian noble burials of the Yuan Dynasty through a multidisciplinary approach. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20130378. doi: 10.1098/rstb.2013.0378

Deguilloux, M. F., Pemonge, M. H., Mendisco, F., Thibon, D., Cartron, I., and Castex, D. (2014). Ancient DNA and kinship analysis of human remains deposited in Merovingian necropolis sarcophagi (Jau Dignac et Loirac. France, 7th-8th century AD). *J. Archaeol. Sci.* 41, 399–405. doi: 10.1016/j.jas.2013.09.006

Egeland, T., Mostad, P. F., Mevåg, B., and Stenersen, M. (2000). Beyond traditional paternity and identification cases: selecting the most probable pedigree. *Forensic Sci. Int.* 110, 47–59. doi: 10.1016/s0379-0738(00)00147-x

Esparza, A., Palomo-Dìez, S., Velasco-Vàzquez, J., Delibes, G., Arroyo-Pardo, E., and Salazar-Garcìa, D. C. (2017). Familiar Kinship? palaeogenetic and isotopic evidence from a triple burial of the cogotas i Archaeological culture (Bronze Age, Iberian Peninsula). *Oxford J. Archaeol.* 36, 223–242. doi: 10.1111/ojoa.12113

Fu, Q., Hajdinjak, M., Moldovan, O. T., Constantin, S., Mallick, S., Skoglund, P., et al. (2015). An early modern human from Romania with a recent Neanderthal ancestor. *Nature* 524, 216–219. doi: 10.1038/nature14558

Gamba, C., Fernandez, E., Tirado, M., Pastor, F., and Arroyo-Pardo, E. (2011). Brief communication: ancient nuclear DNA and kinship analysis: the case of a medieval burial in San Esteban Church in Cuellar (Segovia, Central Spain). *Am. J. Phys. Anthropol.* 144, 485–491. doi: 10.1002/ajpa.21451

Gamba, C., Jones, E. R., Teasdale, M. D., McLaughlin, R. L., Gonzalez-Fortes, G., Mattiangeli, V., et al. (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* 5:5257. doi: 10.1038/ncomms6257

Gansauge, M. T., and Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damage DNA. *Nat. Protoc.* 8, 737–748. doi: 10.1038/nprot.2013.038

Gill, P. (2001). An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *Int. J. Legal Med.* 114, 204–210. doi: 10.1007/s004149900117

Ginolhac, A., Rasmussen, M., Gilbert, M. T., Willerslev, E., and Orlando, L. (2011). mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics* 27, 2153–2155. doi: 10.1093/bioinformatics/btr347

Goodnight, K., and Queller, D. (2001). *Relatedness v.5.0.8*. Huston: G. Software.

Green, R. L., Lagacé, R. E., Oldroyd, N. J., Hennessy, L. K., and Mulero, J. J. (2013). Developmental validation of the AmpFâ„"STR® NGM SElect PCR Amplification Kit: a next-generation STR multiplex with the SE33 locus. *Forensic Sci. Int.* 7, 41–51. doi: 10.1016/j.fsigen.2012.05.012

Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., et al. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* 19, 318–326. doi: 10.1101/gr.081398.108

Haak, W., Brandt, G., Jong, H. N. D., Meyer, C., Ganslmeier, R., Heyd, V., et al. (2008). Ancient DNA, Strontium isotopes, and osteological analyses shed light on social and kinship organization of the Later Stone Age. *PNAS* 105, 18226–18231. doi: 10.1073/pnas.0807592105

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522, 207–211. doi: 10.1038/nature14317

Hanghøj, K., Moltke, I., Andersen, P. A., Manica, A., and Korneliussen, T. S. (2019). Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. *GigaScience* 8:giz034. doi: 10.1093/gigascience/giz034

Hassett, B. (2006). Mandibular Torus: etiology and bioarchaeological utility. *Dental Anthropol.* 9, 1–9.

Hawass, Z., Gad, Y. Z., Ismail, S., Khairat, R., Fathalla, D., Hasan, N., et al. (2010). Ancestry and pathology in king Tutankhamun's family. *JAMA* 303, 638–647. doi: 10.1001/jama.2010.121

Huff, C. D., Witherspoon, D. J., Simonson, T. S., Xing, J., Watkins, W. S., Zhang, Y., et al. (2011). Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Res.* 21, 768–774. doi: 10.1101/gr.115972.110

Just, R. S., Leney, M. D., Barritt, S. M., Los, C. W., Smith, B. C., Holland, T. D., et al. (2009). The Use of mitochondrial DNA single nucleotide polymorphisms to assist in the resolution of three challenging forensic cases. *J. Forensic Sci.* 54, 887–891. doi: 10.1111/j.1556-4029.2009.01069.x

Jónnson, H., Ginolhac, A., Schubert, M., Johnson, P. L., and Oralndo, L. (2013). mapDamage2.0: fast approzimate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684. doi: 10.1093/bioinformatics/btt193

Kennett, D. J., Plog, S., George, R. J., Culleton, B. J., Watson, A. S., Skoglund, P., et al. (2016). Archaeogenomic evidence reveals prehistoric matrilineal dynasty. *Nat. Commun.* 8:14115.

Keyser-Tracqui, C., Crubézy, E., and Ludes, B. (2003). Nuclear and mitochondrial DNA analysis of a 2,000-year-old necropolis in the Egyin Gol Valley of Mongolia. *Am. J. Hum. Genet.* 73, 247–260. doi: 10.1086/377005

Kling, D., Tillmar, A. O., and Egeland, T. (2014). Familias 3 - Extensions and new functionality. *Forensic Sci. Int.* 13, 121–127. doi: 10.1016/j.fsigen.2014.07.004

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* 15:356. doi: 10.1186/s12859-014-0356-4

Korneliussen, T. S., and Moltke, I. (2015). NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics* 31, 4009–4011. doi: 10.1093/bioinformatics/btv509

Lipatov, M., Sanjeev, K., Patro, R., and Veeramah, K. R. (2015). Maximum likelihood estimation of biological relatedness from low coverage sequencing data. *biorxiv [Preprint]* doi: 10.1101/023374

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873. doi: 10.1093/bioinformatics/btq559

Maricic, T., Whitten, M., and Paabo, S. (2010). Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One* 5:e14004. doi: 10.1371/journal.pone.0014004

Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., Szecsenyi-Nagy, A., Rohland, N., Mallick, S., et al. (2018). The genomic history of southeastern Europe. *Nature* 555, 197–203. doi: 10.1038/nature25778

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. I. A., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528:499. doi: 10.1038/nature16152

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 6:pdb.prot5448. doi: 10.1101/pdb.prot5448

Mittnik, A., Massy, K., Knipper, C., Wittenborn, F., Friedrich, R., Pfrengle, S., et al. (2019). Kinship-based social inequality in Bronze Age Europe. *Science* 366, 731–734. doi: 10.1126/science.aax6219

Monroy Kuhn, J. M., Jakobsson, M., and Günther, T. (2018). Estimating genetic kin relationships in prehistoric populations. *PLoS One* 13:e0195491. doi: 10.1371/journal.pone.0195491

Mooder, K. P., Weber, A. W., Bamforth, F. J., Lieverse, A. R., Schurr, T. G., Bazaliiski, V. I., et al. (2005). Matrilineal affinities and prehistoric Siberian mortuary practices: a case study from Neolithic Lake Baikal. *J. Archaeol. Sci.* 32, 619–634. doi: 10.1016/j.jas.2004.12.002

Mulero, J. J., Chang, C. W., Calandro, L. M., Green, R. L., Li, Y., Johnson, C. L., et al. (2006). Development and validation of the AmpFlSTR(Yfiler PCR Amplification Kit: a male specific, single amplification 17 Y-STR multiplex system. *J. Forensic Sci.* 51, 64–75. doi: 10.1111/j.1556-4029.2005.00016.x

Mulero, J. J., Chang, C. W., Lagacé, R. E., Wang, D. Y., Bas, J. L., McMahon, T. P., et al. (2008). Development and validation of the AmpFlSTR(MiniFilerTM PCR Amplification Kit: a MiniSTR multiplex for the analysis of degraded and/or PCR inhibited DNA. *J. Forensic Sci.* 53, 838–852. doi: 10.1111/j.1556-4029.2008.00760.x

Nastainczyk, M., Schulz, S., Kleiber, M., and Immel, U. D. (2009). STR analysis of degraded DNA using a miniplex. *Forensic Sci. Int.* 2, 53–54. doi: 10.1016/j.fsigss.2009.08.107

Olalde, I., Brace, S., Allentoft, M. E., Armit, I., Kristiansen, K., Booth, T., et al. (2018). The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* 555, 190–196. doi: 10.1038/nature25738

O'Sullivan, N., Posth, C., Coia, V., Schuenemann, V. J., Price, T. D., Wahl, J., et al. (2018). Ancient genome-wide analyses infer kinship structure in an Early Medieval Alemannic graveyard. *Sci. Adv.* 4:eaao1262. doi: 10.1126/sciadv.aao1262

Palomo-Dìez, S., Esparza Arroyo, A., Tirado-Vizcaìno, M., Velasco Vàzquez, J., Lòpez-Parra, A. M., Gomes, C., et al. (2018). Kinship analysis and allelic dropout: a forensic approach on an archaeological case. *Ann. Hum. Biol.* 45, 365–368. doi: 10.1080/03014460.2018.1484159

Patterson, N., Moorjani, P., Luo, Y., Mallik, S., Rohland, N., Zhan, Y., et al. (2012). Ancient admixture in human history. *Genetics* 192, 1065–1093. doi: 10.1534/genetics.112.145037

Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* 28, 2537–2539. doi: 10.1093/bioinformatics/bts460

Pinhasi, R., Fernandes, D., Sirak, K., Novak, M., Connell, S., Alpaslan-Roodenberg, S., et al. (2015). Optimal ancient DNA yields from the inner ear part of the human petrous bone. *PLoS One* 10:e0129102. doi: 10.1371/journal.pone.0129102

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Riancho, J. A., and Zarrabeitia, M. ÌA. T. (2003). A Windows-based software for common paternity and sibling analyses. *Forensic Sci. Int.* 135, 232–234. doi: 10.1016/s0379-0738(03)00217-2

Ricaut, F. O. X., Auriol, V., von Cramon-Taubadel, N., Keyser, C., Murail, P., Ludes, B., et al. (2010). Comparison between morphological and genetic data to estimate biological relationship: the case of the Egyin Gol necropolis (Mongolia). *Am. J. Phys. Anthropol.* 143, 355–364. doi: 10.1002/ajpa.21322

Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., and Reich, D. (2015). Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. B* 370:20130624. doi: 10.1098/rstb.2013.0624

Rudbeck, L., Gilbert, M. T. P., Willerslev, E., Hansen, A. J., Lynnerup, N., Christensen, T., et al. (2005). mtDNA analysis of human remains from an early Danish Christian cemetery. *Am. J. Phys. Anthropol.* 128, 424–429. doi: 10.1002/ajpa.20294

Sanchez-Quinto, F., Malmstrom, H., Fraser, M., Girdland-Flink, L., Svensson, E. M., Simoes, L. G., et al. (2019). Megalithic tombs in western and northern Neolithic Europe were linked to a kindred society. *Proc. Natl. Acad. Sci. U.S.A.* 116, 9469–9474. doi: 10.1073/pnas.1818037116

Schroeder, H., Margaryan, A., Szmyt, M., Theulot, B., Wlodarczak, P., Rasmussen, S., et al. (2019). Unraveling ancestry, kinship, and violence in a Late Neolithic mass grave. *Proc. Natl. Acad. Sci. U.S.A.* 116, 10705–10710. doi: 10.1073/pnas.1820210116

Sikora, M., Seguin-Orlando, A., Sousa, V. C., Albrechtsen, A., Korneliussen, T., Ko, A., et al. (2017). Ancient genomes show social and reproductive behaviour of early Upper Paleolithic foragers. *Science* 3, 659–662. doi: 10.1126/science.aao1807

Sprecher, C. J., McLaren, R. S., Rabbach, D., Krenke, B., Ensenberger, M. G., Fulmer, P. M., et al. (2009). PowerPlex(ESX and ESI Systems: a suite of new STR systems designed to meet the changing needs of the DNA-typing community. *Forensic Sci. Int.* 2, 2–4. doi: 10.1016/j.fsigss.2009.08.058

Stojanowski, C. M., and Hubbard, A. R. (2017). Sensitivity of dental phenotypic data for the identification of biological relatives. *Int. J. Osteoarchaeol.* 27, 813–827. doi: 10.1002/oa.2596

Thompson, E. A. (1975). The estimation of pairwise relationships. *Ann. Hum. Genet.* 39, 173–188. doi: 10.1111/j.1469-1809.1975.tb00120.x

Thornton, T., Tang, H., Hoffmann, Thomas, J., Ochs-Balcom, HeatherÂ, M., et al. (2012). Estimating Kinship in Admixed Populations. *Am. J. Hum. Genet.* 91, 122–138. doi: 10.1016/j.ajhg.2012.05.024

Vanek, D., Saskova, L., and Koch, H. (2009). Kinship and Y-chromosome analysis of 7th century human remains: novel DNA extraction and typing procedure for ancient material. *Croat. Med. J.* 50, 286–295. doi: 10.3325/cmj.2009.50.286

Veeramah, K. R. (2018). The importance of fine-scale studies for integrating paleogenomics and archaeology. *Curr. Opin. Genet. Dev.* 53, 83–89. doi: 10.1016/j.gde.2018.07.007

Wang, D. Y., Chang, C.-W., Lagacé, R. E., Oldroyd, N. J., and Hennessy, L. K. (2011). Development and validation of the AmpFlSTR(Identifiler(Direct PCR Amplification Kit: a multiplex assay for the direct amplification of single-source samples. *J. Forensic Sci.* 56, 835–845. doi: 10.1111/j.1556-4029.2011.01757.x

Waples, R. K., Albrechtsen, A., and Moltke, I. (2019). Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. *Mol. Ecol.* 28, 35–48. doi: 10.1111/mec.14954

Weir, B. S., Anderson, A. D., and Hepler, A. B. (2006). Genetic relatedness analysis: modern data and new challenges. *Nat. Rev. Genet.* 7, 771–780. doi: 10.1038/nrg1960

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

ORIGINAL RESEARCH
published: 03 April 2020
doi: 10.3389/fevo.2020.00084

# PIA: More Accurate Taxonomic Assignment of Metagenomic Data Demonstrated on sedaDNA From the North Sea

*Becky Cribdon[1], Roselyn Ware[1], Oliver Smith[1†], Vincent Gaffney[2] and Robin G. Allaby[1]**

*[1] School of Life Sciences, University of Warwick, Coventry, United Kingdom, [2] School of Archaeological and Forensic Sciences, University of Bradford, Bradford, United Kingdom*

Assigning metagenomic reads to taxa presents significant challenges. Existing approaches address some issues, but are mostly limited to metabarcoding or optimized for microbial data. We present PIA (Phylogenetic Intersection Analysis): a taxonomic binner that works from standard BLAST output while mitigating key effects of incomplete databases. Benchmarking against MEGAN using sedaDNA suggests that, while PIA is less sensitive, it can be more accurate. We use known sequences to estimate the accuracy of PIA at up to 96% when the real organism is not represented in the database. For ancient DNA, where taxa of interest are frequently over-represented domesticates or absent, poorly-known organisms, more accurate assignment is critical, even at the expense of sensitivity. PIA offers an approach to objectively filter out false positive hits without the need to manually remove taxa and so make presuppositions about past environments and their palaeoecologies.

**Keywords: ancient DNA, BLAST, MEGAN, metagenomics, sedaDNA, taxonomic assignment**

OPEN ACCESS

***Edited by:***
*Michael Knapp,*
*University of Otago, New Zealand*

***Reviewed by:***
*John Kenneth Pearman,*
*Cawthron Institute, New Zealand*
*Kieren James Mitchell,*
*The University of Adelaide, Australia*

****Correspondence:***
*Robin G. Allaby*
*r.g.allaby@warwick.ac.uk*

***†Present address:***
*Oliver Smith,*
*Micropathology Ltd., University of Warwick Science Park, Coventry, United Kingdom*

***Specialty section:***
*This article was submitted to Paleoecology, a section of the journal Frontiers in Ecology and Evolution*

***Received:*** *22 October 2019*
***Accepted:*** *12 March 2020*
***Published:*** *03 April 2020*

***Citation:***
*Cribdon B, Ware R, Smith O, Gaffney V and Allaby RG (2020) PIA: More Accurate Taxonomic Assignment of Metagenomic Data Demonstrated on sedaDNA From the North Sea. Front. Ecol. Evol. 8:84. doi: 10.3389/fevo.2020.00084*

## INTRODUCTION

Next-generation sequencing allows detailed metagenomic analysis of a wide range of ancient samples. Studies have attempted to recreate biological communities from material including coprolites (Bon et al., 2012; Appelt et al., 2014), dental calculus (Warinner et al., 2015; Weyrich et al., 2017), ice cores (Willerslev et al., 2007), sediment (Birks and Birks Hilary, 2015; Smith et al., 2015), stalagmites (Stahlschmidt et al., 2019), rodent middens (Kuch et al., 2002) and mollusc shells (Der Sarkissian et al., 2016). Our understanding of contamination and best laboratory practice has made good progress (Gilbert et al., 2005; Shapiro et al., 2019) and methods for authenticating ancient DNA sequences are developing (Key et al., 2017; Renaud et al., 2019). However, identifying ancient metagenomic sequences is still a challenge, particularly for shotgun data.

Shotgun sequencing has three key advantages over metabarcoding for ancient metagenomics. First, it can capture information from anywhere in the genome, greatly increasing sensitivity. Every DNA molecule extracted from a sample has the potential to be identified, provided that reference databases are adequate. Second, read count and genome size could be used to calculate biogenomic mass: a proxy of biomass (Gaffney et al., 2020). Third, metabarcoding is far less likely to record DNA damage signals. Damage accumulates in DNA over time (Kistler et al., 2017), so is important for authentication of ancient reads, and occurs most rapidly on the single-stranded overhangs at the ends of molecules. A characteristic damage signal is C-T deamination; changes to the base sequence make it less likely that metabarcoding primers will anneal, so damaged molecules are less likely to be sequenced. Furthermore, primer regions are typically removed during analysis, so even if the very ends of molecules are amplified, they will not be considered. Shotgun sequencing can potentially sequence whole molecules, especially when fragments are short, as is the case for ancient DNA. This preserves any damage signal intact. Overall, shotgun data has the potential to supply highly sensitive and informative metagenomic data.

However, because sequences can come from anywhere in the genome, accurately assigning shotgun reads to taxa requires a much larger reference database than for metabarcoding. The GenBank database is the most comprehensive (Benson et al., 2016), but even this is highly incomplete. Only a tiny fraction of organisms have had their full genomes sequenced and most are not represented at all. Reads from unrepresented organisms may go unassigned. Worse, the uneven representation of taxa that are in a database can create two additional problems that may lead to incorrect assignments.

The first problem is the over-representation of some taxa. This was recently identified as an issue for BLAST (Zhang et al., 2000), the "gold standard" of taxonomic binning (Herbig et al., 2016), by Shah et al. (2018). When BLAST searches against a database, it starts at the top and returns the first *n* hits that pass a quality filter, not the best *n* hits. If an over-represented taxon is a reasonable match, BLAST could return *n* hits and finish before it has a chance to identify closer but less represented taxa further down the database. Better matches may be missing from the list of hits. Even if BLAST does check the whole database, the list of hits may be disproportionately full of over-represented taxa. Taxonomic assignment methods that consider this list may then assign with too much weight to these taxa.

The second problem with an uneven database is "oasis" taxa in "sparse" areas. Consider a sparsely-populated area of the database with just one or a few taxa represented, not including the real taxon (**Figure 1B**). A specific sequence is unlikely to hit anything and will probably be left unassigned. But a conservative sequence may hit that one or few taxa, not necessarily because they are a good match, but because there is nothing else closer. The list of BLAST hits for that read will not be empty, but will have very low diversity. This can give the illusion of a confident match. Taxonomic binners that use a phylogenetic intersection or "lowest common ancestor" approach, robust to conservative sequences, can produce false positives because of oasis taxa.

BLAST and BLAST-like algorithms have a minimum quality filter that affects how similar a reference sequence must be to count as a hit and how much empty space there must be around a read for it to go unassigned (**Figure 1**, "hit radius"). But as with many aspects of taxonomic assignment, this filter has a trade-off between accuracy and sensitivity. A very strict filter would increase the resistance of reads to not-very-similar oases, but make them less attracted to more similar sequences that could be informative. This is especially an issue for aDNA, where even a read from an organism that is in the database may not share an identical sequence because of DNA damage or mutations over time. The minimum quality filter cannot protect from oasis taxa alone.

One of the main arguments in favor of metabarcoding is its use of confined, curated databases that aim to be functionally complete for the study taxon in the study area, such as the Arctic flora database in Sønstebø et al. (2010). Uneven representation is limited if all taxa are represented to some degree. It is currently realistic to sequence a barcode region of a several hundred species for a study, as in Sønstebø et al. But because shotgun sequencing can access the whole genome, a complete shotgun database must have the full genome of all organisms, which will not happen in the foreseeable future. Metabarcoding databases are typically far more "complete" in that more of the study taxa are represented. However, this still assumes that an environment can even be well-studied enough for a complete list of taxa. This is debatable, especially for ancient ecosystems. Despite metabarcoding databases being easier to fill, arguably neither can ever be truly complete. Metabarcoding does not fully address uneven representation in databases. Both metabarcoding and shotgun approaches would benefit from an alternative solution.

A method that accepts shotgun data while also improving the database is SPARSE (Zhou et al., 2018). It rebuilds a given database as hierarchical clusters of similar sequences. If a taxon is represented by several very similar genomes, these genomes will be combined into a single cluster. The final SPARSE database has every present taxon represented by one genome, addressing the problem of over-represented taxa. However, SPARSE is designed for microbial data in relatively well-studied systems, where the database is both relatively well-populated and small enough to be rebuilt on a typical lab server. It does not address the problem of oasis taxa in sparse areas, nor would it be easily applicable to studies of organisms with larger genomes.

A popular standard tool for metagenomic studies not limited to microbes is MEGAN (Huson et al., 2007, 2016). This analyses output from various reference-matching programs, including BLAST. Its sister program, MALT (Herbig et al., 2016), aims to generate comparable output to BLAST at greatly increased speed before assigning taxonomy in the same way as MEGAN. This shared method is the LCA (Lowest Common Ancestor) algorithm (Huson et al., 2016). The default naive LCA is best suited to taxonomic binning. For each read, hits are first quality-filtered against multiple criteria. Good hits are assumed to belong not to the single organism they were sequenced from, but the "lowest common ancestor" (ancestral node) of all associated taxa. Being associated with multiple taxa suggests that the hit sequence is conservative, so should be assigned to a higher taxon. The more conserved the sequence, the more diverse the associated taxa, so the higher the taxon to which the hit is assigned. Following the same logic, the *read* is then assigned to the lowest common ancestor of its list of processed hits.

The LCA is robust to overrepresented taxa in the list of hits. The lowest common ancestor is calculated on presence/absence, not number of occurrences. However, accurate assignment still depends on the list containing accurate hits to begin with, which overrepresented taxa can prevent (Shah et al., 2018).

The LCA also addresses unrepresented taxa: even if the real taxon is not in the database, the list of hits should include relatives, so the read should be assigned to an "ancestor" that encompasses the real taxon. The more sparse the database,

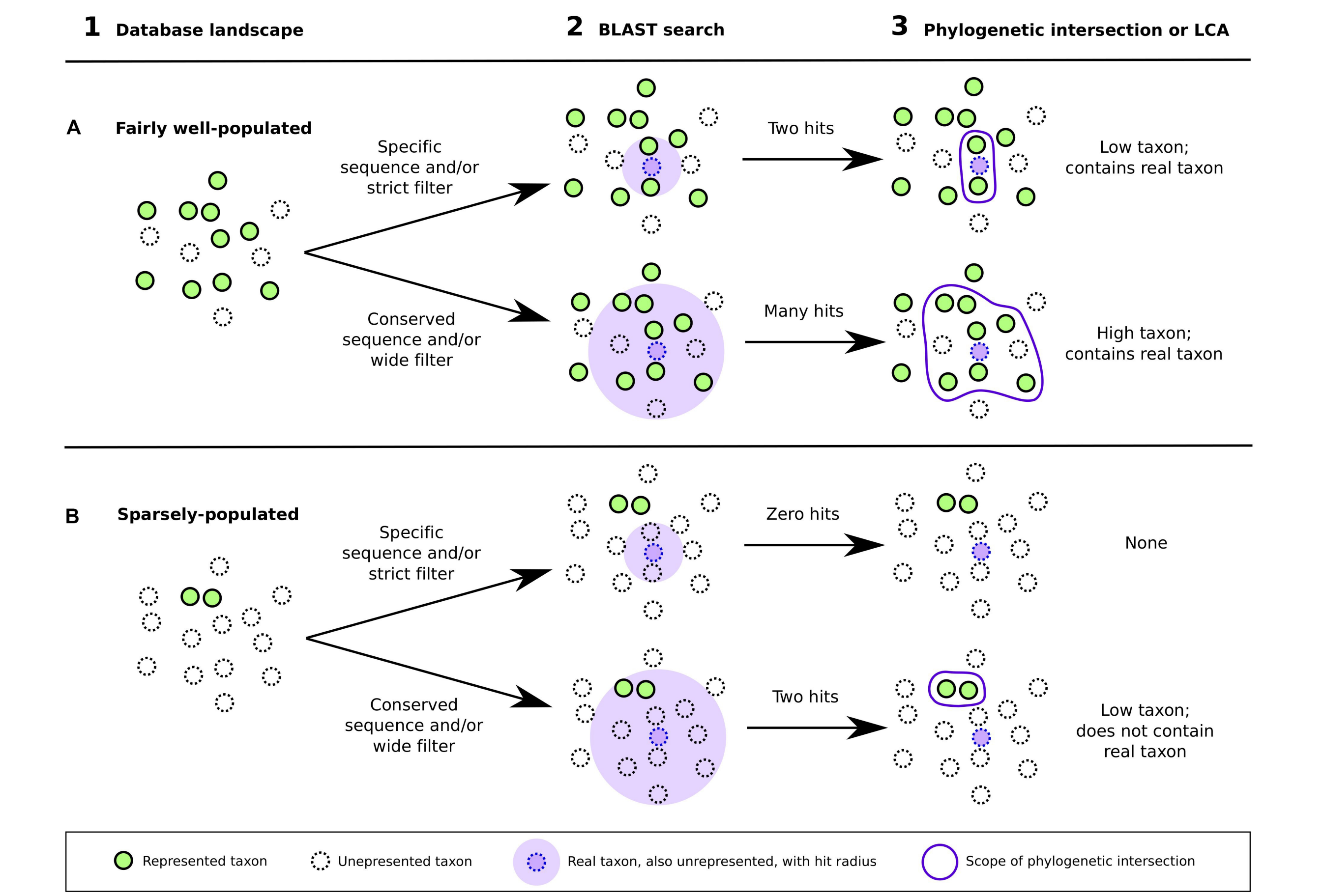

1 Database landscape
2 BLAST search
3 Phylogenetic intersection or LCA
A Fairly well-populated
Specific sequence and/or strict filter
Conserved sequence and/or wide filter
Two hits
Many hits
Low taxon; contains real taxon
High taxon; contains real taxon
B Sparsely-populated
Specific sequence and/or strict filter
Conserved sequence and/or wide filter
Zero hits
Two hits
None
Low taxon; does not contain real taxon
Represented taxon
Unepresented taxon
Real taxon, also unrepresented, with hit radius
Scope of phylogenetic intersection

**FIGURE 1 |** Database landscape diagrams of an **(A)** well-populated and **(B)** sparsely-populated region. Column (1) is the state of the database region. (2) Shows a BLAST search for a sequence from an unrepresented taxon; for a reference sequence to be a hit, it must be inside the hit radius. Column (3) is what phylogenetic intersection or LCA ("lowest common ancestor") those BLAST hits would produce: what the read would be assigned to. In a well-populated region **(A)**, the more conservative the sequence (or the wider the quality filter), the higher the taxon the read would be assigned to. The intersection should contain diverse relatives of the real taxon, so is likely to contain the real taxon. In a sparsely-populated region **(B)**, a specific sequence (or strict filter) is unlikely to be assigned. But a more conservative sequence (or wider filter) may hit the one or few represented taxa in the region: oases. These are probably not close relatives of the real taxon. However, the resulting low-diversity list of hits gives the intersection an illusion of confidence.

the more diverse the list of hits, so the higher the taxon the read is assigned to. In very sparse regions, this means that reads are likely to be under- or unassigned but not incorrectly over-assigned (Huson et al., 2007). However, we argue that the LCA approach may incorrectly assign these reads if they are influenced by oasis taxa. If, for instance, a sparse region were occupied by clumps of taxa rather than an even spread of relatives around the unrepresented taxon (**Figure 1B**), the list of hits may be dominated by one of those taxon clumps, resulting in a relatively specific "ancestor" close to the oasis but not necessarily the real taxon.

MEGAN does have a further check against false positives: the *min-support* filter (Huson et al., 2007; Huson, 2019). Once all reads have been assigned, resulting taxa are only reported if they contain a minimum number of reads. If a read was assigned to a taxon that does not meet this threshold, it is pushed up the taxonomy until it reaches a taxon that does. This excludes very rare taxa, which Huson et al. argue are more likely to be false positives. However, we argue that oasis taxa could escape this check. Being the only represented taxon in that database region, an oasis could potentially pull in reads that would otherwise be assigned to multiple local taxa. The fewer other taxa around, the stronger the oasis effect, and the greater the number of reads incorrectly assigned to that taxon. Oasis taxa can systematically generate false positives that are not necessarily rare.

In this paper, we present Phylogenetic Intersection Analysis (PIA) as a taxonomic binner which, like MEGAN, works from gold-standard BLAST output and is not designed specifically for microbial data, yet goes further to address the shortcomings of BLAST and databases. It also filters BLAST hits by a strict quality threshold. It also accounts for over-represented taxa by only counting each hit taxon once. It also avoids over-assigning conservative hits and sequences by finding a lowest common ancestor, here called a phylogenetic intersection to avoid ambiguity when dealing with ancient sequences that may genuinely be ancestral. However, there are two key differences between MEGAN and PIA. First is a difference with finding the intersection. MEGAN accepts an LCA calculated from just one taxon (i.e., that taxon itself), but if PIA does not have at least two taxa, it discards the read. It assumes that the real taxon is not in the database, so will not assign directly to a taxon in the database. It only assigns to a higher taxon, assuming that the real taxon lies within that phylogenetic range. This avoids over-assigning unrepresented reads to close relatives. Second is a diversity check that measures the extent of population in the region of the database. Reads assigned in sparse regions, vulnerable to the influence of oasis taxa, are discarded. PIA discards the majority of reads, but those that remain are robustly assigned. The resulting assignations are reliable despite low read counts.

This study evaluates PIA by benchmarking its performance against MEGAN with empirical and simulated data. The empirical data was generated as part of the Europe's Lost Frontiers project. This aims to reconstruct submerged palaeolandscapes around the United Kingdom, particularly Doggerland, which now lies under the North Sea. One arm of the project is multi-proxy analysis of sediment cores. This study uses our sedaDNA data from core ELF039, chosen because most samples had a relatively high data yield and the geological context suggested a potentially interesting story. For more information, see Gaffney et al. (2020).

## ALGORITHM

A very early version of PIA was originally presented in Smith et al. (2015). Although the central approach has not changed, it has been substantially rewritten and refined. Scripts are available from https://github.com/Allaby-lab/PIA.

### The Input BLAST File

The two inputs for PIA are a FASTA of query sequences and a corresponding BLAST file. The BLAST file must be in format six (tabular) with all standard columns followed by an additional column containing taxonomic IDs associated with the reference sequence hit. This column is how PIA assigns hits to taxa. We also use the "-max_target_seqs" parameter to limit the number of hits returned per query sequence, recognizing that the hits returned will be the first *n* to meet a quality threshold (Shah et al., 2018). Although PIA aims to reduce the impact of overrepresented taxa in databases once the BLAST is complete, it is important that this BLAST takes enough hits to reach underrepresented taxa. "-max_target_seqs" should be as high as practical. We suggest 500 as a default. Finally, note that BLAST can be run with *x* number of threads. Many of our larger samples took days to BLAST despite using several threads. This is by far the most computationally expensive part of the pipeline.

A typical pre-PIA BLAST command:

```
blastn -db [nucleotide database] -num_threads [x] -query
[input FASTA] -out [output] -max_target_seqs 500 -outfmt
"6 std staxids"
```

The resulting BLAST file (**Figure 2**) lists hits first by query sequence, so all hits to a query are together, and then by descending Expect value (*E*), so better matches are generally further up the list. However, within *E* value, the order is simply the order in which the hits occur in the database.

## PIA

The PIA algorithm itself is computationally light enough to be run on a laptop with small sample files (FASTA ∼ <3 MB). The index-building step required before first use should take no more than a few minutes. Time to analyze the seven samples used in this study on one thread ranged from approximately 10 s to 10 min. PIA can also be multi-threaded for larger samples, for which we recommend a server.

**Figure 3** illustrates the PIA algorithm. PIA considers one read at a time. Reads with no BLAST hits are discarded. For reads with hits, PIA first calculates the coverage of the top hit:

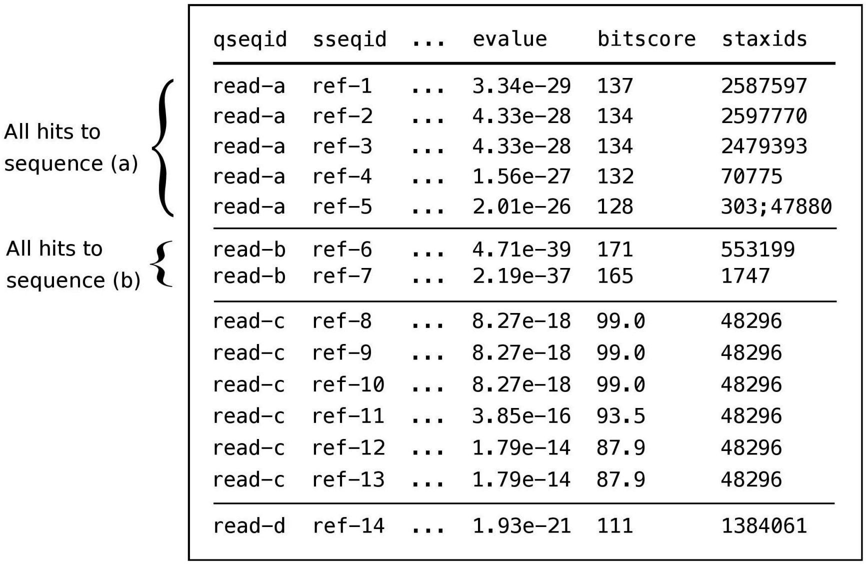
$$\%\ coverage = \frac{match\ length}{read\ length} \times 100$$

If the coverage does not meet a threshold (default 95%), the read is discarded. The taxonomic assignment of the read is strongly influenced by the top hit, so it only accepts a very close match.

PIA then considers each hit in order of the BLAST file. First, the hit is assigned to a taxon. If a hit is associated with multiple taxa, PIA assumes that this indicates a conservative sequence and assigns the hit to the phylogenetic intersection of those taxa. The assigned taxon is then evaluated. If there has already been a hit to the taxon, the hit is discarded. Because hits are listed in order of *E* value, this means that only the best hit for each taxon is retained. This taxon check aims to mitigate the problem of overrepresented taxa. Provided that the BLAST found enough hits to reach underrepresented taxa in the database at all, this check gives them equal weight to overrepresented taxa. Every taxon is reduced to a single hit.

The second check performed on each hit is the *E* value. If there has already been a hit that passed the taxon check with this *E* value, those hits are grouped together. Once all hits for this read have been taxon-checked and grouped by *E* value, the *E* value groups are collapsed to a single "hit" per *E* value. This "hit" is the phylogenetic intersection of the group members. If a read is found to be equally similar to sequences from several different organisms, PIA again assumes that this indicates a conservative sequence. Finally, if these new "hits" are to previously seen taxa, then as before, only the hit with the best *E* value is retained.

Once the list of BLAST hits for the read has been reduced to one (best) hit per taxon, PIA assigns the read to the phylogenetic intersection of the top and second-top hits. If only one hit remains, there cannot be an intersection, so the read is discarded. Finding the intersection firstly avoids over-assigning conservative sequences. Secondly, it avoids over-assigning reads from unrepresented taxa to represented relatives. PIA assumes that the real taxon is not in the database, so it will not assign directly to any organism in the database. The intersection is only taken between the top two hits because, after the taxon check and grouping by *E* value, those two hits may already be to distantly-related and/or high taxa.



```
qseqid  sseqid  ...  evalue    bitscore  staxids
read-a  ref-1   ...  3.34e-29  137       2587597
read-a  ref-2   ...  4.33e-28  134       2597770
read-a  ref-3   ...  4.33e-28  134       2479393
read-a  ref-4   ...  1.56e-27  132       70775
read-a  ref-5   ...  2.01e-26  128       303;47880
read-b  ref-6   ...  4.71e-39  171       553199
read-b  ref-7   ...  2.19e-37  165       1747
read-c  ref-8   ...  8.27e-18  99.0      48296
read-c  ref-9   ...  8.27e-18  99.0      48296
read-c  ref-10  ...  8.27e-18  99.0      48296
read-c  ref-11  ...  3.85e-16  93.5      48296
read-c  ref-12  ...  1.79e-14  87.9      48296
read-c  ref-13  ...  1.79e-14  87.9      48296
read-d  ref-14  ...  1.93e-21  111       1384061
```

**FIGURE 2 |** Example partial BLAST output structure in format "6 std staxids". The standard (std) fields are the first columns, starting with query sequence (qseqid) and ending with Expect (*E*) value (evalue) and score (bitscore). Additional fields, here the taxonomic IDs (staxids) associated with the reference, are at the end. Each row is a hit between the query sequence and a reference sequence from the database. Hits are ordered first by query sequence, then by *E* value from lowest to highest.

The final step is the diversity check, which filters reads by taxonomic diversity score:

$$\textit{Taxonomic diversity score} = \frac{t-1}{c}$$

Where $t$ is the number of different taxa in the original list of BLAST hits and $c$ is a predefined cap on the number of hit taxa to consider. The score measures how populated this area of the database is. A well-populated region will have more hits. If the region is sparsely-populated, there may be a disproportionately high number of hits to oasis taxa. Reads which seem to match an organism in a too sparsely-populated area are discarded.

# METHODS

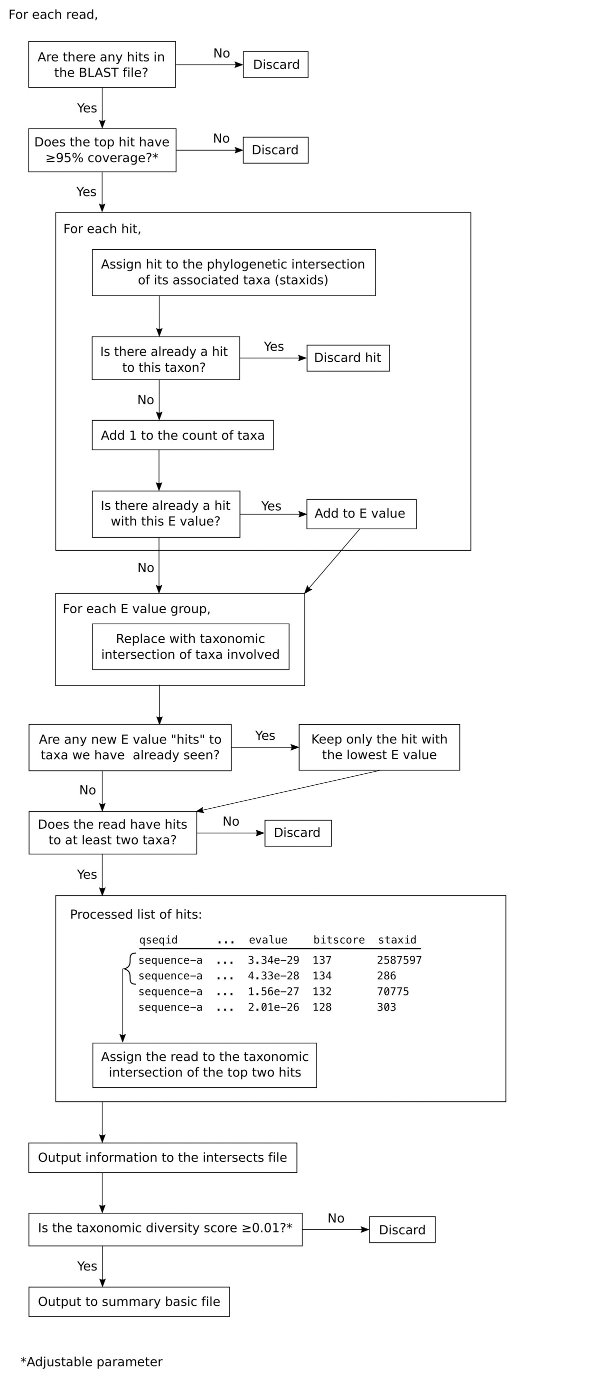
## Analysis of Empirical sedaDNA Data

PIA and MEGAN were compared in a parallel analysis of seven samples from the Europe's Lost Frontiers project (Gaffney et al., 2020). These samples are from sediment core ELF039 which was taken from a palaeochannel approximately 50 km north of the present Norfolk coast. No dates are available for that core at the time of writing, but the channel is interpreted as a river valley that underwent marine inundation during the early Holocene. The samples were shotgun sequenced on a NextSeq 550 as part of our work using sedaDNA for palaeoenvironmental reconstruction. We typically focus on plants because of their high biomass in most environments, increasing the chance of DNA deposition, and the abundance of ecological and distribution information available. Accordingly, this study made use of reads from Viridiplantae.

Raw FASTQ files were adapter-trimmed and collapsed in AdapterRemoval 2.2.2 (Lindgreen, 2012), converted to FASTA, and had duplicates removed using fastx_collapser from the FAST-X Toolkit 0.0.13 (Gordon and Hannon, 2010). Then an initial BLAST was performed against the full nucleotide GenBank database (downloaded on 05-09-2019) using blastn 2.6.0 (Zhang et al., 2000) with -outfmt "6 std staxids" and -num_alignments 10. Output format six is tabular, reducing file size, and reference sequence taxonomic IDs were included to allow full parsing by MEGAN. In format 6, -num_alignments states the maximum number of hits per query. Ten was sufficient for this stage. An RMA file was generated from that BLAST output using the MEGAN5 command line interface with default settings (Huson et al., 2016). Reads assigned to Viridiplantae or below were extracted to a new FASTA. This FASTA was then BLASTed more thoroughly, with -max_target_seqs set to 500 to give up to approximately 500 hits per read.

For the MEGAN analysis, an RMA file was again generated from this final BLAST output using the default settings. All nodes were exported to a text file in the format "taxonID_to_count". The BLAST output and corresponding FASTA were also run through PIA. A custom script[1] (see **Supplementary Material**) was then used to filter both sets of output by a negative control: taxa with a control:sample hit ratio of at least 0.02 were discarded from the sample data. The control is the sum of all negative controls in the wider sequencing run of 142 samples from the same project. The seven filtered sample files were concatenated together and visualized with Krona (Ondov et al., 2011; see **Supplementary Material**).

[1] https://github.com/Allaby-lab/PIA-accessories


For each read,
Are there any hits in the BLAST file?
No
Discard
Yes
Does the top hit have ≥95% coverage?*
No
Discard
Yes
For each hit,
Assign hit to the phylogenetic intersection of its associated taxa (staxids)
Is there already a hit to this taxon?
Yes
Discard hit
No
Add 1 to the count of taxa
Is there already a hit with this E value?
Yes
Add to E value
No
For each E value group,
Replace with taxonomic intersection of taxa involved
Are any new E value "hits" to taxa we have already seen?
Yes
Keep only the hit with the lowest E value
No
Does the read have hits to at least two taxa?
No
Discard
Yes
Processed list of hits:
Assign the read to the taxonomic intersection of the top two hits
Output information to the intersects file
Is the taxonomic diversity score ≥0.01?*
No
Discard
Yes
Output to summary basic file
*Adjustable parameter

```
qseqid       ...  evalue     bitscore  staxid
sequence-a   ...  3.34e-29   137       2587597
sequence-a   ...  4.33e-28   134       286
sequence-a   ...  1.56e-27   132       70775
sequence-a   ...  2.01e-26   128       303
```

**FIGURE 3 |** Flowchart illustrating the PIA algorithm. There are three key checks that may result in a read being discarded: sufficient coverage of the top BLAST hit, at least two hits remaining after processing, and a high enough taxonomic diversity score. Reads that pass are assigned to the intersection of the top two remaining BLAST hits.

## Accuracy Testing With Simulated Data

Benchmarking against MEGAN suggested that PIA may successfully increase the accuracy of taxonomic assignations at the cost of sensitivity. To test the accuracy more objectively, we ran both MEGAN and PIA on two test datasets of known GenBank sequences. For each dataset, the control condition used the original BLAST database from the benchmarking analysis (downloaded on 05-09-2019). An "exclusion" condition excluded all taxa in the test dataset from the BLAST database. This aimed to simulate the unrepresented taxa, common in metagenomic data, that PIA is designed to analyze. In each condition, we tracked the assignations of individual sequences and compared them to the actual source organisms. Most stages involved custom scripts available from https://github.com/Allaby-lab/PIA-accessories and detailed in the **Supplementary Material**.

Each test dataset comprised 250 GenBank sequences downloaded through the NCBI website. For the first dataset, sequences were first filtered to Embryophyta and to a length of 30–150 bp to reflect typical aDNA. We then iterated through "All other taxa" from the "Results by taxon" option until taxa were represented by no more than 44 relevant sequences. Metagenomic data is likely to contain poorly-represented organisms. Single sequences from 245 taxa were downloaded as a FASTA with GIs included. An additional five 30–150 bp sequences were added from well-represented domesticates: *Hordeum vulgare, Musa acuminata*, *Triticum dicoccon*, *Triticum aestivum*, and *Zea mays*. These were run through BLASTn to check that they did match their taxa labels, as model organism sequences are frequently assigned to incorrect taxa. The second dataset was constructed in a similar way, but first filtered to Mammalia instead of Embryophyta. The low-frequency taxa were represented by up to 47 relevant sequences and the five high-frequency taxa were *Camelus bactrianus*, *Camelus dromedarius*, *Balaenoptera bonaerensis*, *Chlorocebus aethiops*, and *Papio anubis*. Finally, each FASTA file was re-formatted to single-line using fasta_formatter from the FAST-X toolkit 0.0.13 (Gordon and Hannon, 2010). The final FASTAs are included as **Supplementary Data Sheets S2, S3** in the **Supplementary Material**.

The FASTAs were run through BLAST with the same settings as in benchmarking. The exclusion condition only differed in the reduced database. For every taxon, a list of GIs for all sequences from that taxon was downloaded from GenBank. These lists were concatenated into a master GI list. The BLAST option "-negative_gilist" was used to exclude this list from the database. For each BLAST file, the MEGAN and PIA analyses were performed with the same settings as in benchmarking. See the **Supplementary Material** for details.

It became apparent after analysis that two Mammalia sequences may be affected by human contamination: GI 2198752 (accession no. U84666.1, *Cavia porcellus* Y5 scRNA gene, partial sequence) and GI 13508496 (accession no. AY028924.1, *Mammut americanum* 16S ribosomal RNA gene, partial sequence; mitochondrial gene for mitochondrial product). We ran BLAST on both sequences to check, changing "-max_target_seqs 500" to "-num_alignments 1" to produce easily readable output with the default limit of 500 hits. Other settings were the same as in benchmarking.

Finally, a small separate test of GenBank data was used to evaluate the performance of PIA on highly divergent taxa.

Because of the diversity check, we expect PIA to unnecessarily discard reads assigned to taxa with few living relatives because their region of the database will always appear incomplete. We ran BLAST and PIA on the available GenBank sequences from two monotypic orders: Ginkgoales (containing the gymnosperm *Ginkgo biloba*; 22,600 sequences) and Microbiotheria (containing the marsupial *Dromiciops gliroides*; 417 sequences). This used the same settings as in benchmarking.

# RESULTS

## Analysis of Empirical sedaDNA Data

Taxonomic assignations of early Holocene sedaDNA from a submerged palaeochannel in the North Sea by MEGAN and PIA are compared in **Figures 4A,B**. The most frequent taxa are labeled in full. Of these, taxa not native to Europe are highlighted in bold (see below). The original interactive HTML chart is included as **Supplementary Data Sheet S4** in the **Supplementary Material**.

The taxonomic profiles of the MEGAN and PIA outputs are broadly similar (**Figures 4A,B**). **Figure 4** begins at Mesangiospermae, to which the vast majority of reads are assigned by both methods. Most reads are assigned to *Zostera marina* (eelgrass), related taxa in Potamogetonaceae or to its parent order Alismatales, suggesting a wetland or fully aquatic environment with at least some saltwater influence. There is also a sizeable signal from grasses (Poaceae). In the largest remaining segment, Pentapetalae (**Figure 4B**), both profiles show a diverse range of taxa found in northwest Europe today. This includes Rosaceae (strawberry, bramble, apple, drupe trees), *Salix* (willow), *Populus* (poplar), and Fagales (birch, oak).

However, the numbers of reads making up these taxa differ significantly. Though proportionally similar, the MEGAN profile was built from 88,497 reads compared to just 27,547 accepted by PIA. The MEGAN profile also has higher taxonomic richness, containing 374 taxa versus 210 (**Table 1**). Those MEGAN taxa are also generally more specific. MEGAN assigned far more reads to genus or lower. Overall, the results are consistent with MEGAN placing more emphasis on sensitivity than PIA.

Because the samples originate from northwest Europe in the early Holocene, we would expect DNA sequences to be comparable to European taxa today. The samples have been filtered by negative controls which should have removed most assignations to common modern contaminant taxa present in reagents. We therefore assume any assignations to non-European taxa to be false positives.

Many of the most frequent non-European taxa assigned to by MEGAN are domesticated grasses such as *Oryza*, *Setaria italica* and *Sorghum bicolor* (**Figure 4A**). In Pentapetalae (**Figure 4B**), most of the terminal taxa in the MEGAN output – those genera and species that suggest a higher sensitivity than PIA – are non-European and therefore likely false positives. **Table 1** quantifies all assignations: 40.11% of taxa in the MEGAN profile are suspect compared to 20.95% for PIA. In total, MEGAN assigned 12.78% of reads to non-European taxa and PIA assigned just 0.52%. The false positive taxa have lower counts on average, suggesting that the minimum support filter in MEGAN is a valid approach, but in this case PIA was more effective at removing this sort of false positive.

It appears that the lower sensitivity of PIA is associated with higher accuracy. To investigate this more objectively, we ran PIA on test sequences of known origin.

## Accuracy Testing With Simulated Data

### Embryophyta

Individual reads, their source organism and all four assignations are listed in the first worksheet of **Supplementary Table S2**. **Table 2** provides a summary. We considered an assignation correct if it was to the actual taxon or one of its parent taxa. For example, if PIA assigned a read from *Betula* to the family Betulaceae, it would be a correct assignment at family level. Family level is typically precise enough to be useful for environmental reconstruction in plants. An assignment to Viridiplantae would be correct at kingdom level. An assignment to Poaceae would be incorrect.

In the control condition, MEGAN assigned 91% of sequences and PIA 52%, mirroring the higher sensitivity of MEGAN observed in the analysis of real data. Both were highly accurate at 97 and 100%, respectively. MEGAN was somewhat more precise, with 62% of assignments correct to family level or below, compared to 53.49% for PIA. Overall, MEGAN showed a much greater ability to assign sequences at the cost of a very small drop in accuracy compared to PIA.

The exclusion condition, where the source taxa had been removed from the database, shows a similar pattern of results with generally worse performance by both tools. However, MEGAN appears to suffer more. The “Change” columns in **Table 2** show that MEGAN assigns proportionally fewer sequences at all, correctly, and with precision than PIA. Notably, accuracy of MEGAN falls to 80% but that of PIA remains at a healthy 96%.

Despite the exclusion database generally presenting more of a challenge, there were a small number of sequences that were assigned better than with the complete database. PIA did not assign the *Lapageria rosea* and *Lupinus luteus* sequences in the control condition but matched MEGAN’s broad Mesangiospermae assignment for the exclusion. Both MEGAN and PIA assigned the *Metasequoia glyptostroboides* and *Magnolia x soulangeana* sequences more precisely in the exclusion condition, although not particularly so. This unexpected behavior may be due to peculiarities of the database around those sequences.

### Mammalia

Full results are listed in the second worksheet of **Supplementary Table S2**. **Table 3** provides a summary. In the control condition, the Mammalia dataset showed a similar pattern to Embryophyta. MEGAN assigned more reads and with more precision; both programs were very accurate. The exclusion condition resulted in worse performance for both programs, again with a greater impact on MEGAN. However, the decrease in accuracy was even more pronounced than for Embryophyta. MEGAN only assigned

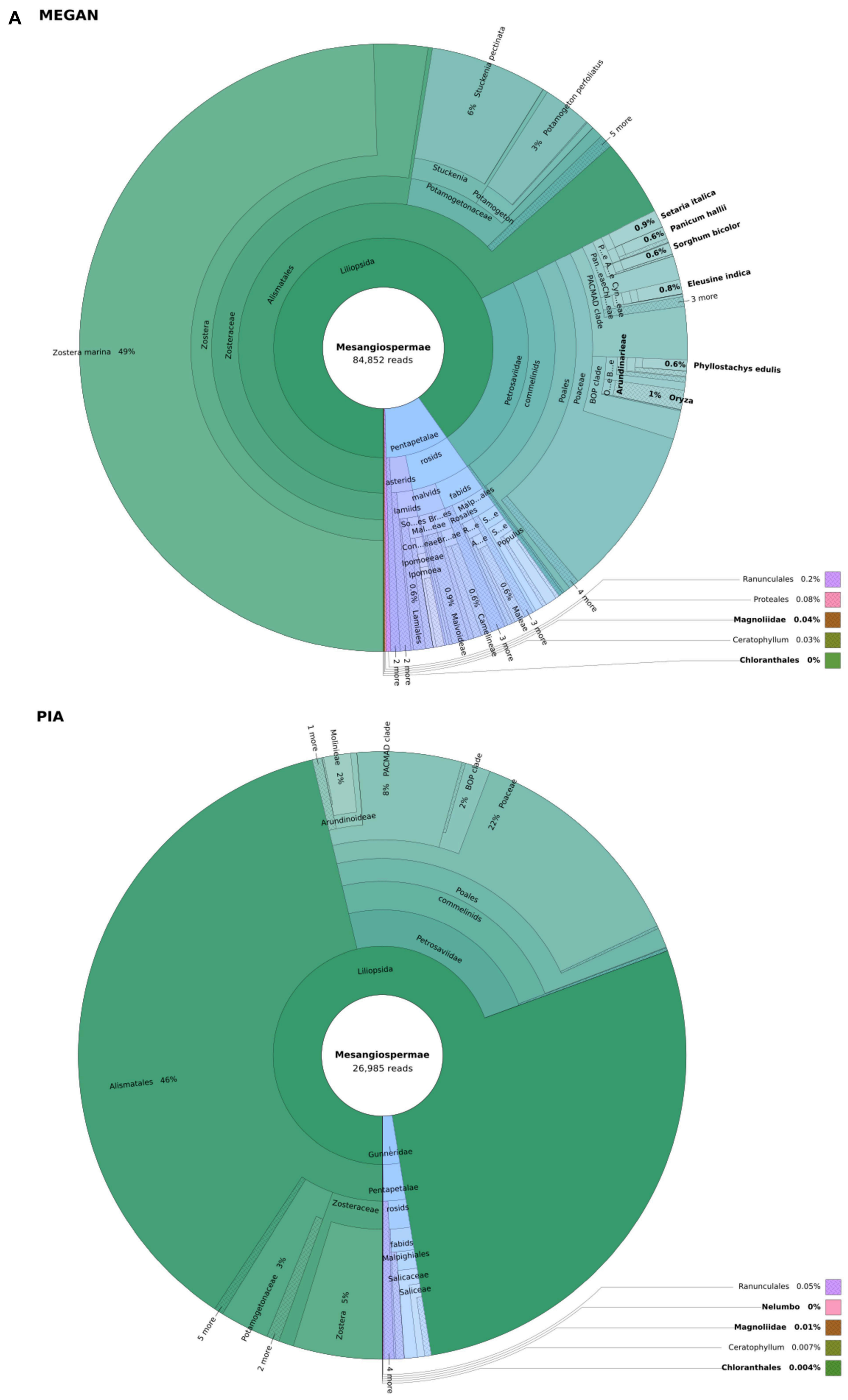


**FIGURE 4 |** Continued

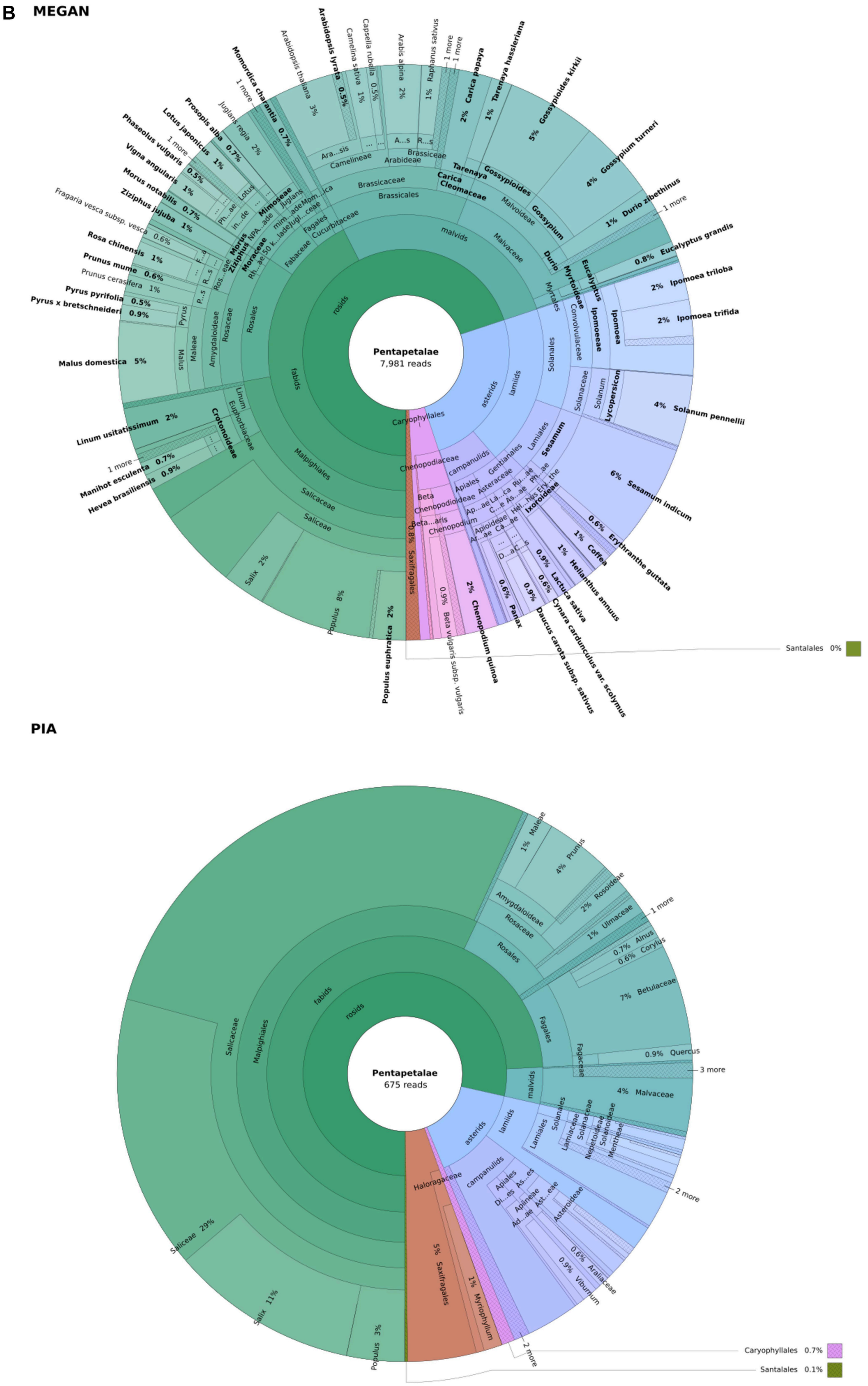


**FIGURE 4 |** Taxonomic profiles of the combined MEGAN and PIA outputs for the seven sediment samples, after filtering each by negative controls. **(A)** shows Mesangiospermae, which includes the vast majority of reads. **(B)** zooms in on Pentapetalae, the largest segment in panel **(A)** that cannot easily be seen. Taxa not native to Europe, which are suspected to be false positive assignments for this data, are highlighted in bold. Colors indicate taxon frequency. See **Supplementary Data Sheet S4** in the **Supplementary Material** for the original interactive HTML chart, which was produced using Krona (Ondov et al., 2011).

**TABLE 1 |** Numbers of European and non-European taxa hit and the numbers of reads assigned to each category in the MEGAN and PIA benchmarking output.

| | MEGAN | PIA |
|---|---|---|
| Total taxa | 374 | 210 |
| European | 224 (59.89%) | 166 (79.05%) |
| Non-European | 150 (40.11%) | 44 (20.95%) |
| Total reads | 88,497 | 27,547 |
| To European taxa | 77,189 (87.22%) | 27,405 (99.48%) |
| To non-European taxa | 11,308 (12.78%) | 142 (0.52%) |

*Reads assigned to non-European taxa are suspected to be false positives for this data.*

60% of sequences accurately. PIA assigned 83% accurately, which while better, is far from the 96% accuracy seen for Embryophyta.

Note that these accuracy results are likely a slight underestimate, as the two questionable sequences (to *Cavia porcellus* and *Mammut americanum*), do indeed appear to be mislabeled. Both BLAST outputs are dominated by *Homo sapiens* and other primates. MEGAN and PIA generally assigned them either to high mammal taxa or close parent taxa of humans, both of which are reasonable if the sequences are actually human.

As with Embryophyta, a small number of sequences were assigned better with their taxa excluded from the database. MEGAN assigned the *Stenella attenuata* sequence incorrectly in the control but broadly correct after exclusion. PIA assigned the *Kogia sima* sequence more precisely after exclusion, though only by one level. Finally, the only time the *Halichoerus grypus* sequence was assigned was by PIA after exclusion, and it did so correctly to family.

## Monotypic Taxa

Phylogenetic Intersection Analysis assigned 5% of reads from Ginkgoales and with only 77% accuracy. For Microbiotheria, PIA assigned 37% of reads; 100% were accurate but the most precise was only to Metatheria. The proportion of reads assigned to each was considerably lower than the ∼50–60% from the mixed test datasets above.

# DISCUSSION

Ancient metagenomics has much potential, but taxonomic assignation of reads can be improved. Databases are highly uneven, resulting in the joint problems of over-represented taxa filling up hit lists at the expense of poorly-represented but closer matches, and oasis taxa in sparsely-populated areas drawing in reads and giving an illusion of confident assignation. There are methods that partly address these problems in some circumstances, but we demonstrate here that PIA performs strongly, providing an objective approach to remove false positives from data sets.

Benchmarking on plant sedaDNA data against a standard tool, MEGAN, showed that PIA produces a comparatively low-resolution taxonomic profile. Far fewer reads are assigned and those that are rarely make it to genus. However, we argue that

**TABLE 2 |** Percentages of the 250 sequences assigned by MEGAN and PIA in the Embryophyta accuracy test.

| Embryophyta | Control BLAST | | Exclusion BLAST | | Change | |
|---|---|---|---|---|---|---|
| | MEGAN | PIA | MEGAN | PIA | MEGAN | PIA |
| Assigned | 91.20% | 51.60% | 76.00% | 45.60% | −15.20% | −06.00% |
| Incorrect | 03.07% | 00.00% | 20.00% | 04.39% | 16.93% | 04.39% |
| Correct | 96.93% | 100.00% | 80.00% | 95.61% | −16.93% | −04.39% |
| Correct to above family | 35.09% | 46.51% | 46.84% | 60.53% | 11.75% | 14.02% |
| Correct to family or below | 61.84% | 53.49% | 33.16% | 35.09% | −28.68% | −18.40% |

*The control condition BLASTed against the full GenBank nucleotide database (downloaded on 05-09-2019). The exclusion condition omitted the source taxa from the database. Of those reads assigned, percentages assigned incorrectly or correctly are given. The final two rows detail whether correctly-assigned reads were assigned to higher taxa or to at least family. These rows sum to the total percent correct.*

**TABLE 3 |** Percentages of the 250 sequences assigned by MEGAN and PIA in the Mammalia accuracy test.

| Mammalia | Control BLAST | | Exclusion BLAST | | Change | |
|---|---|---|---|---|---|---|
| | MEGAN | PIA | MEGAN | PIA | MEGAN | PIA |
| Assigned | 93.60% | 57.60% | 76.40% | 52.40% | −17.20% | −05.20% |
| Incorrect | 02.99% | 00.00% | 40.31% | 16.79% | 37.32% | 16.79% |
| Correct | 97.01% | 100.00% | 59.69% | 83.21% | −37.32% | −16.79% |
| Correct to above family | 28.21% | 45.14% | 41.36% | 49.62% | 13.36% | 4.48% |
| Correct to family or below | 68.80% | 54.86% | 18.32% | 33.59% | −50.48% | −21.27% |

*The control condition BLASTed against the full GenBank nucleotide database (downloaded on 05-09-2019). The exclusion condition omitted the source taxa from the database. Of those reads assigned, percentages assigned incorrectly or correctly are given. The final two rows detail whether correctly-assigned reads were assigned to higher taxa or to at least family. These rows sum to the total percent correct.*

much of the sensitivity of MEGAN in this context is over-sensitivity. Both methods describe core ELF039 as coming from a primarily wetland environment, with a clear signal from fresh and saltwater plants in Alismatales and the riverine *Salix*, along with some signal from grasses in Poaceae and woodland trees in Fagales. Yet the MEGAN profile assigned nearly 13% of reads to clearly questionable taxa, such as the tropical *Sorghum bicolor*, Australasian *Eucalyptus* and American *Carica papaya*, that if taken at face value would present a radical departure from the established palaeoecology of Europe. Once such taxa are removed as "known" false positives, the MEGAN analysis only retrieves a few more taxa than PIA (**Figure 4B**), which add little to the palaeoecological reconstruction and likely still contain false positives. One example is *Arabidopsis thaliana*, a known model organism not expected to feature greatly in the Mesolithic. In our context, the additional accuracy of PIA appears to outweigh the increased sensitivity of MEGAN.

The accuracy test on simulated data returned similar results. With a full BLAST database, MEGAN assigned nearly twice as many sequences with greater precision and only marginally lower accuracy than PIA. However, when the source taxa were excluded from the database, exacerbating the problems caused by incomplete databases and better representing real metagenomic data, the improvements of MEGAN over PIA diminished and the difference in accuracy became substantial. For Embryophyta sequences, PIA maintained a very high accuracy of 96%, whereas that of MEGAN fell to 80%.

Both programs performed less well with the Mammalia dataset, but PIA still returned 83% accuracy after exclusion of source taxa compared to 60% from MEGAN. We suspect that this difference may simply be due to the fact that there are far fewer species of mammal than embryophyte, so removing 250 mammal taxa will have removed proportionally more of the relevant database than removing the same number from Embryophyta. Both PIA and MEGAN performed very well in the control condition, so it is unlikely to be directly due to the mammal sequences themselves. Instead, we suggest that the exclusion condition simulated a more incomplete database for Mammalia than Embryophyta. PIA still outperformed MEGAN. However, it is clear that while PIA copes better with incomplete databases, it is not a perfect solution.

Additionally, two specific limitations of PIA are apparent from its algorithm. First, PIA cannot assign to leaf taxa. It can only assign to a species if there are subspecies in the database, for example. PIA does not fully take advantage of sequences with very good taxonomic resolution. If better resolution is desired, it may be helpful to first identify reads to higher taxa more accurately using PIA, then further analyze any sequences assigned to taxa of interest using a different approach.

The second limitation is a result of the taxonomic diversity check. PIA discards assignments to taxa in sparse areas of the database because these areas are vulnerable to the influence of oasis taxa. However, this assumes that sparsity is due to incompleteness. There are divergent taxa with very few living relatives that will occupy a naturally sparse database region. PIA is less likely to accept assignments to these taxa. To demonstrate this, we ran PIA on the available GenBank sequences from Ginkgoales and Microbiotheria, which are orders containing a single species. PIA assigned fewer reads from these taxa than from the mixed Embryophyta or Mammalia datasets. Such divergent taxa are unusual, but are less likely to be recovered by PIA. Again, PIA shows a lack of sensitivity that may limit its application in some studies.

However, even with these caveats, we have demonstrated that the improved ability of PIA to address the challenges of an incomplete reference database can result in highly accurate taxonomic assignation of metagenomic shotgun data. PIA produced fewer false positives than the standard approach. The more likely false positives are to occur, the more necessary it becomes to manually sort taxa into plausible and implausible, which requires subjective presuppositions about the source of the data. This is particularly problematic for ancient metagenomics where little is known about the study environment. PIA offers an objective alternative with an estimated 96% accuracy for plants.

## DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the European Nucleotide Archive under the project code PRJEB33717. See **Supplementary Table S1** for sample accession codes.

## AUTHOR CONTRIBUTIONS

RA, OS, RW, and BC wrote and designed the PIA. BC performed benchmarking and accuracy testing with some input from RW. BC was the primary author of the manuscript with review and editing by RA and RW. VG was the Principal Investigator of the project through which the sedaDNA dataset was obtained.

## FUNDING

The sedaDNA dataset used for benchmarking was generated as part of the Europe's Lost Frontiers project. This received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (ERC funded project no. 670518 LOST FRONTIER, https://europa.eu/european-union/index_en, https://lostfrontiers.teamapp.com/).

## ACKNOWLEDGMENTS

Thanks to the Europe's Lost Frontiers Project for supplying the sedaDNA data.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.00084/full#supplementary-material

## REFERENCES

Appelt, S., Fancello, L., Le Bailly, M., Raoult, D., Drancourt, M., Desnues, C., et al. (2014). Viruses in a 14th-Century Coprolite. *Appl. Environ. Microbiol.* 80, 2648–2655. doi: 10.1128/AEM.03242-13

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2016). GenBank. *Nucleic Acids Res.* 45, D37–D42. doi: 10.1093/nar/gkw1070

Birks, H. J. B., and Birks Hilary, H. (2015). How have studies of ancient DNA from sediments contributed to the reconstruction of Quaternary floras? *New Phytol.* 209, 499–506. doi: 10.1111/nph.13657

Bon, C., Berthonaud, V., Maksud, F., Labadie, K., Poulain, J., Artiguenave, F., et al. (2012). Coprolites as a source of information on the genome and diet of the cave hyena. *Proc. R. Soc. B Biol. Sci.* 279, 2825–2830. doi: 10.1098/rspb.2012.0358

Der Sarkissian, C., Pichereau, V., Dupont, C., Ilsøe, P. C., Perrigault, M., Butler, P., et al. (2016). Ancient DNA analysis identifies marine mollusc shells as new metagenomic archives of the past. *Mol. Ecol. Resour.* 17, 835–853. doi: 10.1111/1755-0998.12679

Gaffney, V., Fitch, S., Bates, M., Ware, R. L., Kinnaird, T., Gearey, B., et al. (2020). Multi-proxy evidence for the impact of the Storegga Slide Tsunami on the early Holocene landscapes of the southern North Sea. *BioRxiv* [Preprint]. doi: 10.1101/2020.02.24.962605

Gilbert, M. T. P., Bandelt, H.-J., Hofreiter, M., and Barnes, I. (2005). Assessing ancient DNA studies. *Trends Ecol. Evol.* 20, 541–544. doi: 10.1016/j.tree.2005.07.005

Gordon, A., and Hannon, G. J. (2010). *Fastx-toolkit*. Cold Spring Harbor, NY: Hannon Laboratory.

Herbig, A., Maixner, F., Bos, K. I., Zink, A., Krause, J., Huson, D. H., et al. (2016). MALT: fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. *bioRxiv* [Preprint]. doi: 10.1101/050559

Huson, D. H. (2019). *User Manual for MEGAN V6.17.0. 0–74*. Available online at: http://ab.inf.uni-tuebingen.de/data/software/megan6/download/manual.pdf (accessed September 26, 2019)

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386. doi: 10.1101/gr.5969107

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., et al. (2016). MEGAN community edition – interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput. Biol.* 12:e1004957. doi: 10.1371/journal.pcbi.1004957

Key, F. M., Posth, C., Krause, J., Herbig, A., and Bos, K. I. (2017). Mining metagenomic data sets for ancient DNA: recommended protocols for authentication. *Trends Genet.* 33, 508–520. doi: 10.1016/j.tig.2017.05.005

Kistler, L., Ware, R., Smith, O., Collins, M., and Allaby, R. G. (2017). A new model for ancient DNA decay based on paleogenomic meta-analysis. *Nucleic Acids Res.* 45, 6310–6320. doi: 10.1093/nar/gkx361

Kuch, M., Rohland, N., Betancourt, J. L., Latorre, C., Steppan, S., and Poinar, H. N. (2002). Molecular analysis of a 11 700-year-old rodent midden from the Atacama Desert. *Chile. Mol. Ecol.* 11, 913–924. doi: 10.1046/j.1365-294x.2002.01492.x

Lindgreen, S. (2012). AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res. Notes* 5:337. doi: 10.1186/1756-0500-5-337

Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011). Interactive metagenomic visualization in a web browser. *BMC Bioinformatics* 12:385. doi: 10.1186/1471-2105-12-385

Renaud, G., Schubert, M., Sawyer, S., and Orlando, L. (2019). "Authentication and assessment of contamination in ancient DNA," in *Ancient DNA: Methods and Protocols*, ed. B. Shapiro, A. Barlow, P. D. Heintzman, M. Hofreiter, J. L. A. Paijmans, and A. E. R. Soares (New York, NY: Springer), 163–194. doi: 10.1007/978-1-4939-9176-1_17

Shah, N., Nute, M. G., Warnow, T., and Pop, M. (2018). Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows. *Bioinformatics* 35, 1613–1614. doi: 10.1093/bioinformatics/bty833

Shapiro, B., Barlow, A., Heintzman, P. D., Hofreiter, M., Paijmans, J. L. A., and Soares, A. E. R. (eds). (2019). *Ancient DNA Methods and Protocols*, 2nd Edn. New York, NY: Humana Press. doi: 10.1007/978-1-4939-9176-1

Smith, O., Momber, G., Bates, R., Garwood, P., Fitch, S., Pallen, M., et al. (2015). Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago. *Science* 347, 998–1001. doi: 10.1126/science.1261278

Sønstebø, J. H., Gielly, L., Brysting, A. K., Elven, R., Edwards, M., Haile, J., et al. (2010). Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Mol. Ecol. Resour.* 10, 1009–1018. doi: 10.1111/j.1755-0998.2010.02855.x

Stahlschmidt, M. C., Collin, T. C., Fernandes, D. M., Bar-Oz, G., Belfer-Cohen, A., Gao, Z., et al. (2019). Ancient mammalian and plant DNA from late quaternary stalagmite layers at Solkota Cave. *Georgia. Sci. Rep.* 9:6628. doi: 10.1038/s41598-019-43147-0

Warinner, C., Speller, C., and Collins, M. J. (2015). A new era in palaeomicrobiology: prospects for ancient dental calculus as a long-term record of the human oral microbiome. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20130376. doi: 10.1098/rstb.2013.0376

Weyrich, L. S., Duchene, S., Soubrier, J., Arriola, L., Llamas, B., Breen, J., et al. (2017). Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. *Nature* 544, 357–361. doi: 10.1038/nature21674

Willerslev, E., Cappellini, E., Boomsma, W., Nielsen, R., Hebsgaard, M. B., Brand, T. B., et al. (2007). Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science* 317, 111–114. doi: 10.1126/science.1141758

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. A. (2000). Greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* 7, 203–214. doi: 10.1089/10665270050081478

Zhou, Z., Luhmann, N., Alikhan, N.-F., Quince, C., and Achtman, M. (2018). "Accurate reconstruction of microbial strains from metagenomic sequencing using representative reference genomes," in *Research in Computational Molecular Biology*, ed. B. J. Raphael (Cham: Springer International Publishing), 225–240. doi: 10.1007/978-3-319-89929-9_15

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 15 April 2020
doi: 10.3389/fevo.2020.00094

# Ancient DNA From Museum Specimens and Next Generation Sequencing Help Resolve the Controversial Evolutionary History of the Critically Endangered Puebla Deer Mouse

***Susette Castañeda-Rico[1,2,3]*, Livia León-Paniagua[4], Cody W. Edwards[2,3] and Jesús E. Maldonado[1,2]***

*[1] Center for Conservation Genomics, Smithsonian's National Zoo & Conservation Biology Institute, Washington, DC, United States, [2] Department of Biology, George Mason University, Fairfax, VA, United States, [3] Smithsonian-Mason School of Conservation, Front Royal, VA, United States, [4] Departamento de Biología Evolutiva, Facultad de Ciencias, Universidad Nacional Autónoma de México, Mexico City, Mexico*

Key insights into the evolutionary history of recently extinct or critically endangered species can be obtained through analysis of genomic data collected using high-throughput sequencing and ancient DNA from museum specimens, particularly where specimens are rare. For instance, the evolutionary history of the critically endangered Puebla deer mouse, *Peromyscus mekisturus*, remains unclear due to discordance between morphological and molecular phylogenetic analyses. However, previous molecular analyses were based on PCR and Sanger sequencing of only a few mitochondrial genes. Here, we used ancient DNA from historical museum specimens followed by target enrichment and high-throughput sequencing of several thousand nuclear ultraconserved elements and whole mitochondrial genomes to test the validity of the previous phylogenetic placement of *P. mekisturus*. Based on UCEs and mitogenomes, our results revealed that *P. mekisturus* forms a well-supported distinct lineage outside the clade containing all other members of the *Peromyscus melanophrys* group. Additionally, the mitogenome phylogeny further supports the placement of *P. mekisturus* as the sister species of the genus *Reithrodontomys*. This conflicts with the previous mtDNA phylogenetic reconstruction, in which *P. mekisturus* was nested within the species *P. melanophrys*. Our study demonstrates that high-throughput sequencing of ancient DNA, appropriately controlling for contamination and degradation, can provide a robust resolution of the evolutionary history and taxonomic status of species for which few or no modern genetic samples exist. In light of our results and pending further analysis with denser taxon sampling and the addition of morphological data, a re-evaluation of the taxonomy and conservation management plans of *P. mekisturus* is needed to ensure that the evolutionary distinctiveness of this species is recognized in future conservation efforts.

**Keywords: biodiversity, diversification, mitogenomes, *Peromyscus mekisturus*, phylogenomics, Sanger sequencing, scientific collections, ultraconserved elements**

OPEN ACCESS

***Edited by:***
*Michael Knapp,*
*University of Otago, New Zealand*

***Reviewed by:***
*Alicia Grealy,*
*Australian National University, Australia*
*Kieren James Mitchell,*
*The University of Adelaide, Australia*

****Correspondence:***
*Susette Castañeda-Rico*
*susetteazul@gmail.com;*
*castanedaricos@si.edu*

***Specialty section:***
*This article was submitted to Paleoecology, a section of the journal Frontiers in Ecology and Evolution*

***Received:*** *16 November 2019*
***Accepted:*** *24 March 2020*
***Published:*** *15 April 2020*

***Citation:***
*Castañeda-Rico S, León-Paniagua L, Edwards CW and Maldonado JE (2020) Ancient DNA From Museum Specimens and Next Generation Sequencing Help Resolve the Controversial Evolutionary History of the Critically Endangered Puebla Deer Mouse. Front. Ecol. Evol. 8:94. doi: 10.3389/fevo.2020.00094*

## INTRODUCTION

Scientific collections are unique repositories of biodiversity. They preserve and make accessible specimens that capture variation across taxonomic boundaries, space, and time (Webster, 2018). The specimens deposited in museums allow us to study demographic changes of populations through space and time, including those of extinct species. These specimens also provide a historical context to examine patterns of genetic variation and offer direct measures of evolutionary processes (Burrell et al., 2015; Buerki and Baker, 2016). Genetic studies incorporating natural history collections also allow us to study rare or critically endangered species. Some of these species remain poorly understood due to the scarcity of samples, either because small population sizes make them difficult to find or because they have become extinct in the wild.

Despite their enormous potential, museum specimens are difficult to work with because of postmortem DNA fragmentation and damage. Additionally, the relationship between specimen age and DNA fragmentation is not linear, and even some recently collected samples (<20 years) can have highly fragmented DNA (Zimmermann et al., 2008; Allentoft et al., 2012; Sawyer et al., 2012; Burrell et al., 2015). DNA quality and quantity may be more strongly influenced by preservation methods, storage conditions, the type of tissue targeted, or how quickly the sample was desiccated than to the age of the specimen itself (Pääbo, 1989; Casas-Marce et al., 2010; Mason et al., 2011). Despite these challenges, museum specimens represent a unique repository of valuable information, making it worthwhile to unlock their potential with novel genomic protocols.

These novel genomic methods have spurred a renaissance in studies of natural history collections. Scientists have been studying ancient DNA for more than three decades, but the advent of high-throughput (HT) or next-generation sequencing (NGS) has made the process of sequencing ancient DNA from both model and non-model organisms much easier (Church, 2006; Lemmon and Lemmon, 2013; Hawkins et al., 2016a,b; McCormack et al., 2017; McDonough et al., 2018; Webster, 2018). The improvement is reflected by decreasing the cost and increasing the efficiency of genomic data collection by several orders of magnitude (Lemmon and Lemmon, 2013; Buerki and Baker, 2016; McCormack et al., 2017). NGS methods like sequence capture or target enrichment have been changing the field of phylogenetics and are especially well-suited for sequencing ancient DNA or other degraded samples. This method involves hybridizing genomic DNA to biotinylated DNA or RNA 'baits' present in solution and then washing away unbound, non-target DNA. The result is a DNA solution enriched for specific targets that can then be sequenced using NGS platforms (Burrell et al., 2015). Using sequence capture, researchers can focus sequencing efforts on loci useful for their particular genomic scope, enabling them to increase the number of taxa or samples that can be processed, analyze samples that were difficult to use in the past, and improve phylogenetic resolution (Lemmon and Lemmon, 2013). Filling gaps in the tree of life should significantly improve topological and branch-length estimation and will also allow more accurate biogeographical reconstructions (Buerki and Baker, 2016).

Members of the rodent genus *Peromyscus* are commonly referred to as deer mice. *Peromyscus* is the most common and speciose genus within the subfamily Neotominae. This genus comprises more than 70 new world species that diverged within the last 6–10 million years (Platt et al., 2015). Despite intensive and extensive studies of this genus, understanding its phylogenetic relationships has been difficult. Phylogenetic studies have suggested that the genus *Peromyscus* is paraphyletic, including *Habromys*, *Isthmomys*, *Megadontomys*, *Neotomodon*, *Osgoodomys*, and *Podomys* at the generic (*sensu stricto*) or subgeneric (*sensu lato*) level (Bradley et al., 2007; Miller and Engstrom, 2008; Platt et al., 2015; Sullivan et al., 2017). The large number of species, both described and undescribed, as well as the cryptic variation present in the group, have yielded numerous distinct phylogenetic hypotheses (Sullivan et al., 2017). Osgood (1909) placed related species into monophyletic species groups based mainly on morphological similarities. At present, several lines of evidence support the recognition of 13 *Peromyscus* species groups (Carleton, 1989; Hogan et al., 1993; Musser and Carleton, 1993, 2005, Dawson, 2005; Bradley et al., 2007). However, the composition of the groups has been modified several times based on new evidence and many of the species have been re-categorized (Carleton, 1989; Riddle et al., 2000; Álvarez-Castañeda and González-Ruiz, 2008).

The *Peromyscus melanophrys* group, endemic to México, comprises three species: *Peromyscus melanophrys*, *P. perfulvus*, and *P. mekisturus* (Osgood, 1909; Carleton, 1989; Musser and Carleton, 1993, 2005; Bradley et al., 2007; Castañeda-Rico et al., 2014). Recent field surveys searching for additional specimens of *P. mekisturus* have failed, suggesting that this species is likely extinct or close to extinction. To date, only one Sanger sequencing-based study has included all three species in phylogenetic analyses. Castañeda-Rico et al. (2014) analyzed all three species using the *ND3, tRNA-Arg, ND4L,* and partial *ND4* mitochondrial genes. However, owing to the degraded condition of the *P. mekisturus* specimen (collected in 1947), only *P. melanophrys* and *P. perfulvus* samples produced sequences of the nuclear gene (*GHR*). Castañeda-Rico et al. (2014) found that the three species form a monophyletic group, which is concordant with long-standing morphology-based taxonomic hypotheses (Osgood, 1909; Carleton, 1989; Musser and Carleton, 1993, 2005). Critically, the Sanger sequencing data placed the single *P. mekisturus* specimen within a clade where all of the other samples are considered to be *P. melanophrys*, which contrasts with morphological evidence that supports all three groups as distinct species (Osgood, 1909; Carleton, 1989; Musser and Carleton, 1993, 2005).

Given the discrepancy between the morphological data and the molecular data based on Sanger sequencing, we aimed to test whether we could corroborate the previous phylogenetic hypothesis proposed by Castañeda-Rico et al. (2014) by using NGS methods to obtain dense nuclear and mitochondrial datasets for *P. mekisturus* and the other members of the *Peromyscus melanophrys* species group. Crucially, since *P. mekisturus* is only known from two specimens—Merriam (1898) holotype from

**TABLE 1 |** Specimens used in this study with corresponding species, Museo de Zoología "Alfonso L. Herrera" Facultad de Ciencias UNAM (MZFC) and University of Michigan Museum of Zoology (UMMZ) accession number collection, ID for this study, tissue type that was destructively sampled, date collected, number of UCE reads after filtering, number of UCE loci, reads mapped to the reference mitogenome and mean coverage mitogenome.

| Species | Accession number collection | ID | Tissue | Date collected | Reads UCEs | UCEs loci | Reads mapped mitogenome | Mean coverage mitogenome |
|---|---|---|---|---|---|---|---|---|
| *P. mekisturus* | UMMZ_88967 | UMMZ88967 | Dry skin | 1947 | 5,578,674 | 2,996 | 99,865 | 581.7 |
| *P. melanophrys* | MZFC_3907 | MQ1229 | Dry skin | 1984 | 1,811,856 | 2,700 | 1,310 | 7.3 |
| *P. perfulvus* | – | MCP119 | Internal organ | 2010 | 4,779,011 | 1,353 | 54,235 | 10 |
| *P. mexicanus* | MZFC_11150 | MRM030 | Internal organ | 2010 | 2,876,361 | 1,691 | – | – |
| *P. eremicus* | MZFC_10465 | FCR176 | Internal organ | 2009 | 2,637,980 | 3,219 | 8,806 | 62.3 |
| *H. simulatus* | MZFC_10104 | HBR031 | Internal organ | 2006 | 3,106,129 | 3,316 | – | – |

Chalchicomula, and Hooper's (1947) record from Tehuacán, both in Puebla, México—its phylogenetic position can only be resolved by denser sequencing of the genomes of the specimens in hand. Recent developments in NGS technology and ancient DNA protocols offer a huge advantage for rare and under-collected species such as this one, where sometimes only their holotypes are known to science.

# MATERIALS AND METHODS

## Sampling and Lab Work

We sampled six museum specimens deposited in scientific collections in México and the United States (**Table 1**). We followed strict protocols to avoid contamination during sampling, including the use of a new disposable scalpel blade and gloves for each specimen, and bleaching 50% household bleach (5.25% sodium hypochlorite solution) followed by rinsing with HPLC-grade water of all work surfaces and utensils prior to each use (McDonough et al., 2018). We performed all laboratory work at the ancient DNA facilities at the Center for Conservation Genomics (CCG), Smithsonian Conservation Biology Institute, Washington, DC. We extracted genomic DNA from one tissue sample from each of the six specimens, including *Peromyscus melanophrys*, *P. perfulvus*, *P. mekisturus*, *P. mexicanus*, *P. eremicus*, and *Habromys simulatus*. The last three species were included as outgroups, following Castañeda-Rico et al. (2014).

We extracted ethanol-preserved internal organ samples (kidney, liver, or heart) using a DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA, United States) following the manufacturer's protocol, and visualized on 1% agarose gel to assess quality. Gels were run at 120 volts for 45 min using 1× TBE (Tris-Borate-EDTA) buffer. We quantified each extraction using a Qubit® (Life Technologies) fluorometer with a 1× dsDNA HS assay kit. We extracted dried skin clips using a standard phenol-chloroform protocol in an ancient DNA facility at the CCG following established ancient DNA standards (Pääbo et al., 2004; Willerslev and Cooper, 2005; McDonough et al., 2018). We sheared 100 μl of each DNA extraction, after normalizing concentrations to approximately 400–500 ng/μl, to an average length of ca. 500 bp using a Bioruptor® Pico sonicator (Diagenode) with a pulse of 30 s on/30 s off for 90 cycles. Afterward, we concentrated the samples via centrifugation to 25 μl and cleaned them using 2× solid-phase reversible immobilization (SPRI) magnetic beads (Rohland and Reich, 2012) following the manufacturer's instructions to remove small fragments. We did not shear, quantify, or clean extractions from dried skin due to the inherent degradation and fragmentation of the DNA.

We prepared each DNA sample (22 μl) as a dual-indexed library using Kapa LTP Library Preparation kit (Kapa Biosystems, Boston, MA, United States) for Illumina® Platforms sequencing following the manufacturer's protocol (Kapa Biosystems V6.17), with 1/4 reactions. We performed all pre-PCR steps for the skin samples in a laboratory specifically dedicated to processing of ancient DNA. The ancient DNA lab is physically separated from the main laboratory, and no modern tissue/DNA samples or PCR amplifications are allowed. We performed dual indexing PCR with Nextera-style indices (Faircloth and Glenn, 2012) using Kapa HiFi Hotstart Ready Mix (Kapa Biosystems) according to the manufacturer's protocols, with 14 cycles for organ samples and 18 cycles for skin samples. We purified the resulting indexed libraries using 1.6× SPRI magnetic beads and visualized on a 1% agarose gel (conditions as mentioned above). We quantified library concentrations using Qubit® 1× dsDNA HS assay and we inspected library size-ranges and qualities using a Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, United States) with High Sensitivity DNA kits. We pooled libraries equimolarly in pairs for the organ samples; we did not pool the more degraded skin samples. We performed enrichment for UCEs using an in-solution DNA hybridization with synthetic RNA baits. We used the myBaits® UCE Tetrapods 5Kv1 kit (Arbor Biosciences) following the myBaits protocol v3. The *P. mekisturus* sample was library prepped and enriched twice to confirm the results (data not shown). While both sample enrichments were sequenced, only one of the samples was included in the final analyses upon confirmation that both samples yielded the same results.

To sequence the mitochondrial genome of *P. mekisturus*, we followed the single tube library preparation method described by Carøe et al. (2017). We performed dual indexing PCR with TruSeq-style indices (Meyer and Kircher, 2010) using Kapa HiFi Uracil + kit (Kapa Biosystems) according to the Carøe et al. (2017) protocol. We performed all steps in duplicate. Prior to the index PCR reaction, we performed a quantitative

PCR (qPCR) using SYBR green fluorescence (Kapa Biosystems Illumina Library Quantification kit) in order to determine the number of cycles for library amplification. We used 16 and 20 cycles, and pooled post-PCR products. We then enriched for the mitochondrial genome using the myBaits® Mito kit (Arbor Biosciences) designed for the house mouse, *Mus musculus*, following the myBaits protocol v4.

We amplified post-enrichment UCE libraries with 14–18 cycles of PCR using universal Illumina primers (see myBaits protocol v3 and v4) and Kapa HiFi Hotstart Ready Mix (Kapa Biosystems) according to the manufacturer's protocol, and sequenced them on a MiSeq (Illumina, Inc., San Diego, CA, United States) using a 600-cycle Reagent Kit v3 (2 × 300 bp) at the CCG. We split samples into two groups using a different kit for each group. In order to ensure that we would obtain enough coverage for the *P. mekisturus* sample, we sequenced it at a much deeper coverage compared to the other samples. We amplified the post-enrichment mitogenome library with 18 cycles following the same protocol as above, and we sequenced it using a 2 × 150 bp – PE SP kit on a NovaSeq 6000 (Illumina, Inc., San Diego, CA, United States) at the Vincent J. Coates Genomics Sequencing Laboratory, UC Berkeley (combined with samples from other projects). We evaluated the quantity and quality of each sequencing pool using a Qubit® 1× dsDNA HS assay and a Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, United States) before sequencing.

In order to test the reliability of our results based on NGS generated data, we also PCR amplified and Sanger sequenced using aliquots from the same *P. mekisturus* DNA sample. Because we only had limited amount of DNA extract after conducting the NGS protocols, we decided to only test the first five out of the eight fragments from the Castañeda-Rico et al. (2014) study. In order to ensure that primer stocks used in this assay were free of contamination, we ordered a new set based on the published sequences of these five primer pairs (Integrated DNA Technologies, Inc.). All pre-PCR reactions were prepared in our ancient DNA facility taking same precautions as described above. We amplified DNA in a 25 μl reaction volume containing the following: 9 ng of template DNA, 1 unit of AmpliTaq Gold (Thermo Fisher Scientific), 1.5 mM of $MgCl_2$, 100 μM of deoxynucleoside triphosphate, and 0.25 μM of each primer. Polymerase chain reaction (PCR) conditions were as follow: Taq activation at 95°C for 10 min, followed by 17 cycles at 95°C for 30 s, 50°C annealing temperature for 1 min, and 72°C for 1 min, followed by 27 cycles at 95°C for 30 s, 55°C for 1 min, and 72°C for 1 min with a final extension of 72°C for 5 min. We cleaned PCR products with ExoSAP-IT (Thermo Fisher Scientific), followed by the sequencing reaction using BigDye Terminator v3.1cycle sequencing kit (Thermo Fisher Scientific). We cleaned the sequencing reaction using Sephadex G-50 Fine (GE Healthcare Life Sciences), following the manufacturer's protocols. Bi-directional Sanger sequencing was performed in an ABI Prism 3130XL Genetic Analyzer (Applied Biosystems, Foster City, CA, United States) at the CCG. To control for contamination, we included negative controls in all amplification reactions and sequenced them. Each fragment was also amplified and sequenced in duplicate to corroborate our results.

In addition to the sequences we generated in our lab, we also reanalyzed previously published data including the following: mitochondrial gene cytochrome b (*Cytb*) gene sequences from Bradley et al. (2007) and Platt et al. (2015), mitochondrial genes *ND3*, *tRNA-Arg*, *ND4L*, and partial *ND4* genes from Castañeda-Rico et al. (2014) and full mitogenomes from Sullivan et al. (2017). A list of these analyzed samples is found in **Supplementary Table S1**.

## Bioinformatic Processing of Ultraconserved Elements and Mitogenomes

### Ultraconserved Elements

We demultiplexed samples in FASTQ format using BaseSpace (Illumina, Inc.). We processed raw FASTQ files following the PHYLUCE v1.6.7 bioinformatic pipeline (Faircloth, 2016) with default parameters, available at https://phyluce.readthedocs.io/en/latest/tutorial-one.html (accessed on August 2019). We used Illumiprocessor 2.0.7 (Faircloth, 2013), which allows processing of Illumina sequencing reads using the trimming tool Trim Galore 0.6.5[1] to clean the data (to remove adapter contamination, barcode regions and low-quality bases). We assembled reads into contigs using Trinity 2.8.5 (Grabherr et al., 2011). Following contig assembly, we identified contigs matching UCE loci in the 5K UCE locus set[2]. We created a "taxon set" containing all of our samples to query the database generated during UCE contig identification and created a list of UCE loci by sample. We generated a monolithic FASTA file to extract sequences from each sample. We aligned FASTA sequences using MAFFT 7.4 (Katoh and Standley, 2013) and performed internal trimming using Gblocks 0.91b (Castresana, 2000). We quantified informative sites with the *phyluce_align_get_informative_sites.py* script. We filtered the resulting alignment to create a 75 and 95% complete matrix. All of these analyses were performed on the Smithsonian Institution High Performance Cluster (SI/HPC).

### Mitogenomes

In order to obtain mitogenomes as "off-target sequences" from UCE capture sequences, we used cleaned reads (paired P1 and P2, plus singletons) generated by Illumiprocessor (Faircloth, 2013) through the PHYLUCE pipeline. We removed exact duplicates (—derep1,4) using Prinseq-lite v0.20.4 (Schmieder and Edwards, 2011). We mapped the reads to a reference (*Peromyscus megalops*, GenBank KY707305) using the Geneious algorithm in Geneious Prime® 2019.2.3[3] with default parameters (Medium-Low sensitivity, Maximum mismatches = 20%, Maximum gaps = 10%). We generated consensus sequences with Geneious, using 5× as the lowest coverage to call a base, a Highest Quality control, and the remaining default parameters, and aligned them using MAFFT 7.4 (Katoh and Standley, 2013). We transferred annotations from *P. megalops* reference (GenBank KY707305) and translated genes to check for stop codons.

[1] https://github.com/FelixKrueger/TrimGalore

[2] https://github.com/faircloth-lab/uce-probe-sets

[3] https://www.geneious.com

We demultiplexed the sample sequenced on a NovaSeq in FASTQ format using BaseSpace (Illumina, Inc.). We checked sequence read quality using FastQC v0.11.5[4] (Andrews, 2010). We quality filtered the reads using Trim Galore 0.6.5[5] to remove adapter sequences and low-quality reads using the default parameters as Phred:20, mean min-len:20. We further filtered the trimmed DNA sequencing reads using Prinseq-lite v0.20.4 (Schmieder and Edwards, 2011) to remove exact duplicates (−derep1,4). We used the resulting high-quality, de-duplicated reads in all subsequent steps, following the steps outlined above. We compared the partial mitogenome obtained from off-target UCE enrichment to the one obtained using mitogenome probes in order to corroborate the *P. mekisturus* mitogenome sequence using data generated independently.

### Sanger Sequencing

We edited and cleaned sequences using BioEdit 7.2.5 (Hall, 1999). We extracted *tRNA-Gly*, *ND3*, *tRNA-Arg*, *ND4L*, and partial *ND4* mitochondrial genes from the *P. mekisturus* mitogenome. We used MAFFT 7.4 (Katoh and Standley, 2013) in Geneious to align and compare the extracted mitochondrial genes with the new Sanger sequences obtained in this study and the *P. mekisturus*_KF885810 sequences obtained by Castañeda-Rico et al. (2014).

### *Peromyscus mekisturus* – Mitogenome

We performed three additional analyses for *P. mekisturus* in order to validate that the mitogenome sequence was not the result of contamination. First, in addition to the mitogenome that we obtained by mapping sequence reads to a reference genome (*P. megalops*, GenBank KY707305), we performed a *de novo* assembly using MIRA 4.0 (Chevreux et al., 1999) in Geneious with default paraments, which include an accurate quality level of assembly. Second, we exported from Geneious the reads mapped to the reference genome as FASTQ files and we converted them to FASTA files with seqtk version1.2[6] (Li, 2013). We used mega-BLAST (in BLAST + version 2.6.0 – Camacho et al., 2009) to align the reads against the nucleotide database (accessed on January 29, 2020). We then used MEGAN version 6.18.4 (Huson et al., 2016) to visualize and analyze the BLAST output, following the default parameters and parsed with the LCA (Lowest Common Ancestor) method. Finally, we used mapDamage2.0 (Jónsson et al., 2013) to examine the patterns of DNA damage sequencing artifacts. We analyzed the reads obtained from the mitogenome enrichment and mapped to the closest reference genome. We used the default parameters in mapDamage2.0.

## Phylogenetic Analyses

We performed four independent phylogenetic analyses using: (1) UCE dataset, (2) full mitogenomes, (3) *Cytb* mitochondrial gene, and (4) *ND3*, *tRNA-Arg*, *ND4L*, and partial *ND4* mitochondrial genes.

First, we analyzed the 75% complete UCE matrix using RAxML 8.12 (Stamatakis, 2014) with a GTRGAMMA site rate substitution model and 20 maximum-likelihood (ML) searches for the phylogenetic tree that best fit each set of data. The GTRGAMMA is the most recommended model for ML analyses using RaxML because it represents an acceptable trade-off between speed and accuracy (RAxML 8.12 manual). We generated non-parametric bootstrap replicates using the autoMRE option which runs until convergence is reached. We reconciled the best fitting ML tree with the bootstrap replicate to obtain the final phylogenetic tree. We also analyzed a 95% complete UCE matrix following the protocol above. We ran these analyses on the Smithsonian High-Performance Computing cluster Hydra (SI/HPC). We performed a Bayesian Inference (BI) analysis in MrBayes 3.2.6 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) on the CIPRES infrastructure (Miller et al., 2010). In order to do it, we first estimated the best evolutionary model of nucleotide substitution in jModelTest 2.1.1 (Guindon and Gascuel, 2003; Darriba et al., 2012) using the Akaike Information Criterion (AIC). The TVM + I model was selected as the best fitting model for the UCE dataset (95% matrix) with the following parameters: base frequencies $A = 0.2975$, $C = 0.2037$, $G = 0.2038$, $T = 0.2950$; nst = 6; and portion of invariant sites = 0.7730. We used 50 million generations sampling every 1,000 generations with four Markov chains (three heated and one cold). Heating temperature was set at 0.02 to facilitate greater movement between the four Markov chains. We visualized output parameters using Tracer v1.7.1 (Rambaut et al., 2018) to check for convergence between runs and we discarded the first 25% of the trees as burn-in. UCEs are non-coding regions but are likely involved in controlling gene expression (Marcovitz et al., 2016). However, their function is still an area of research (Faircloth et al., 2012, 2015). In addition, its rate of evolution is still not well understood (Faircloth et al., 2015; Tangliacollo and Lanfear, 2018) and therefore, we did not partition this dataset.

We used the mitogenome data to infer the phylogenetic relationships of *P. mekisturus* in relation to other peromyscine rodents. We included samples generated in this work, as well as samples published previously by Sullivan et al. (2017) and a grasshopper mouse sample, *Onychomys leucogaster* (GenBank KU168563), which was used as an outgroup by Castañeda-Rico et al. (2014). We aligned sequences using MAFFT 7.4 (Katoh and Standley, 2013) in Geneious. We analyzed the mitogenome data without partitions and we estimated the best evolutionary model of nucleotide substitution in jModelTest 2.1.1 (Guindon and Gascuel, 2003; Darriba et al., 2012) using the Akaike Information Criterion (AIC). The GTR + I + G model was selected as the best fitting model with the following parameters: base frequencies $A = 0.3580$, $C = 0.2801$, $G = 0.1106$, $T = 0.2315$; nst = 6; rates = gamma with shape parameter ($\alpha$) = 0.7600; and portion of invariant sites = 0.4740. We also performed model and partition selection using PartitionFinder 2.1.1 (Lanfear et al., 2016), with linked, corrected Akaike Information Criterion (AICc) and greedy parameters, on the SI/HPC cluster. We analyzed two different partitions: (1) by gene and codon position and (2) by codon position, tRNA, rRNA and *D-loop*. The PartitionFinder analysis detected 34 partitions for the by gene and codon position selection and 6 partitions for the by codon position, tRNA,

[4] http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[5] https://github.com/FelixKrueger/TrimGalore

[6] https://github.com/lh3/seqtk

rRNA and *D-loop* selection, both results were incorporated in the phylogenetic reconstructions (**Supplementary Table S2**). We performed a BI analysis in MrBayes 3.2.6 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) on the CIPRES infrastructure (Miller et al., 2010). We used 50 million generations sampling every 1,000 generations (conditions as mentioned above for the UCE dataset). We visualized output parameters using Tracer v1.7.1 (Rambaut et al., 2018) to check for convergence between runs and we discarded the first 25% of the trees as burn-in. We also performed a ML analysis using RAxML 8.12 (Stamatakis, 2014) with a GTRGAMMA site rate substitution model. Clade support was assessed by bootstrapping with 1,000 replicates.

We analyzed *Cytb* sequences extracted from the mitogenomes that we generated here and from the mitogenomes published by Sullivan et al. (2017) in order to evaluate the phylogenetic position of *P. mekisturus* with respect to the genera *Peromyscus*, *Habromys*, *Megadontomys*, *Neotomodon*, *Osgoodomys*, *Podomys*, *Isthmomys*, *Onychomys*, and *Reithrodontomys*. We also used all the sequences published by Bradley et al. (2007) and Platt et al. (2015). By including data from these previous studies, we were able to include representatives of the genera *Neotoma*, *Ochrotomys*, *Baiomys*, *Ototylomys*, *Tylomys*, *Nyctomys*, *Oryzomys*, and *Sigmodon* to be used as outgroups.

We re-evaluated the phylogeny proposed by Castañeda-Rico et al. (2014) in order to compare the *P. mekisturus* sample sequenced by those authors using Sanger sequencing versus the sequence that we obtained here using NGS. We used complete mitochondrial genes *ND3*, *tRNA-Arg*, *ND4L*, and partial *ND4* (hereinafter referred as "multiple mitochondrial genes") sequences published by Castañeda-Rico et al. (2014). We also included sequences of these genes extracted from the mitogenomes that we generated in this study and from the mitogenomes published by Sullivan et al. (2017). Furthermore, we did not use partitions for this dataset so that we could reproduce the methods used by Castañeda-Rico et al. (2014).

We analyzed the *Cytb* dataset and multiple mitochondrial genes separately as follows: we performed alignment using MAFFT 7.4 (Katoh and Standley, 2013) in Geneious. We estimated the best evolutionary model of nucleotide substitution in jModelTest 2.1.1 (Guindon and Gascuel, 2003; Darriba et al., 2012) using the Akaike Information Criterion (AIC). The TVM + I + G model was selected as the best fitting model for *Cytb* with the following parameters: base frequencies $A = 0.3968$, $C = 0.3304$, $G = 0.0467$, $T = 0.2261$; nst = 6; rates = gamma with shape parameter ($\alpha$) = 0.5968, and portion of invariant sites = 0.4040. The TIM2 + I + G model was recognized as the best fitting model for the multiple mitochondrial genes with the following parameters: base frequencies $A = 0.3664$, $C = 0.2971$, $G = 0.0708$, $T = 0.2657$; nst = 6; rates = gamma with shape parameter ($\alpha$) = 1.0390, and portion of invariant sites = 0.4420. We used MrBayes 3.2.6 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) on CIPRES infrastructure (Miller et al., 2010) to reconstruct the phylogenetic trees. Each analysis included the appropriate model identified by jModelTest 2.1.1 (Guindon and Gascuel, 2003; Darriba et al., 2012), 50 million generations, and a sample frequency of every 1,000 generations. We used Tracer v1.7.1 (Rambaut et al., 2018) to check for convergence between runs, and the first 25% of the trees were discarded as burn-in. We also performed a ML analysis for both datasets, using RAxML 8.12 (Stamatakis, 2014) with a GTRGAMMA site rate substitution model. Clade support was assessed by bootstrapping with 1,000 replicates. All phylogenetic trees were visualized in FigTree 1.4.4[7].

We aligned and compared sequences obtained by Castañeda-Rico et al. (2014) [*P. mekisturus*_KF885810 and *P. perfulvus*_KF885791] and one unpublished sequence (*P. melanophrys*_MQ1229 – GenBank accession number MT078 814) using Sanger sequencing versus the multiple mitochondrial genes extracted from the mitogenome of *P. mekisturus*_UMMZ8 8967, *P. perfulvus*_MCP119, and *P. melanophrys*_MQ1229 obtained in this study to corroborate the sequences. We compared sequences derived from the same individuals, where the sample of *P. melanophrys* was an ethanol-preserved internal organ, and for *P. perfulvus* and *P. mekisturus*, the sample was dried skin.

## RESULTS

We successfully sequenced UCEs from all samples processed. Illumina reads were archived in GenBank under BioProject: PRJNA606805. On average, we generated 3.5 million reads per UCE-enriched library, yielding 2,545 UCE loci per sample (**Table 1**). The 75% matrix of aligned loci contained 2,436 UCEs of 3,512 total with an average length of 547 bp and 1.03 informative sites, with the minimum number of taxa per locus $n = 4$. The 95% matrix contained 1,010 UCEs of 3,512 with an average length of 514 bp, and 1.6 informative sites with the minimum number of taxa per locus $n = 5$. The unrooted trees obtained with both datasets showed the same topology and bootstrap values, where *P. melanophrys* and *P. perfulvus* cluster in a monophyletic group, and *P. mexicanus* is the sister species of this group. *P. mekisturus* is the sister species to the clade encompassing *P. mexicanus*, *P. perfulvus*, and *P. melanophrys* (**Figure 1**). All of these phylogenetic relationships are strongly supported, with high bootstrap values of 100. In this phylogenetic analysis, we included members of three out of the 13 *Peromyscus* species groups. *P. mekisturus* yielded long branch lengths using both 75 and 95% matrices (**Figure 1**). This suggests this long branch is not the result of missing data. However, it could be related to the high heterogeneity of *P. mekisturus* with respect to the other taxa included in the tree. This is most likely caused by limited taxon sampling and likely could be resolved by including more taxa in the analysis, as it has been suggested in other studies (Bergesten, 2005; Phillippe et al., 2005; Wiens, 2005). The BI tree (**Figure 1**) showed the same topology as the ML analyses with high posterior probability values of 1.

For several samples, we obtained mitogenome sequences from "bycatch" from the UCE enrichment with no need for mtDNA-specific baits. We were able to recover near-complete (>95% of the reference genome covered) mitogenome sequences for

[7] http://tree.bio.ed.ac.uk/software/figtree/

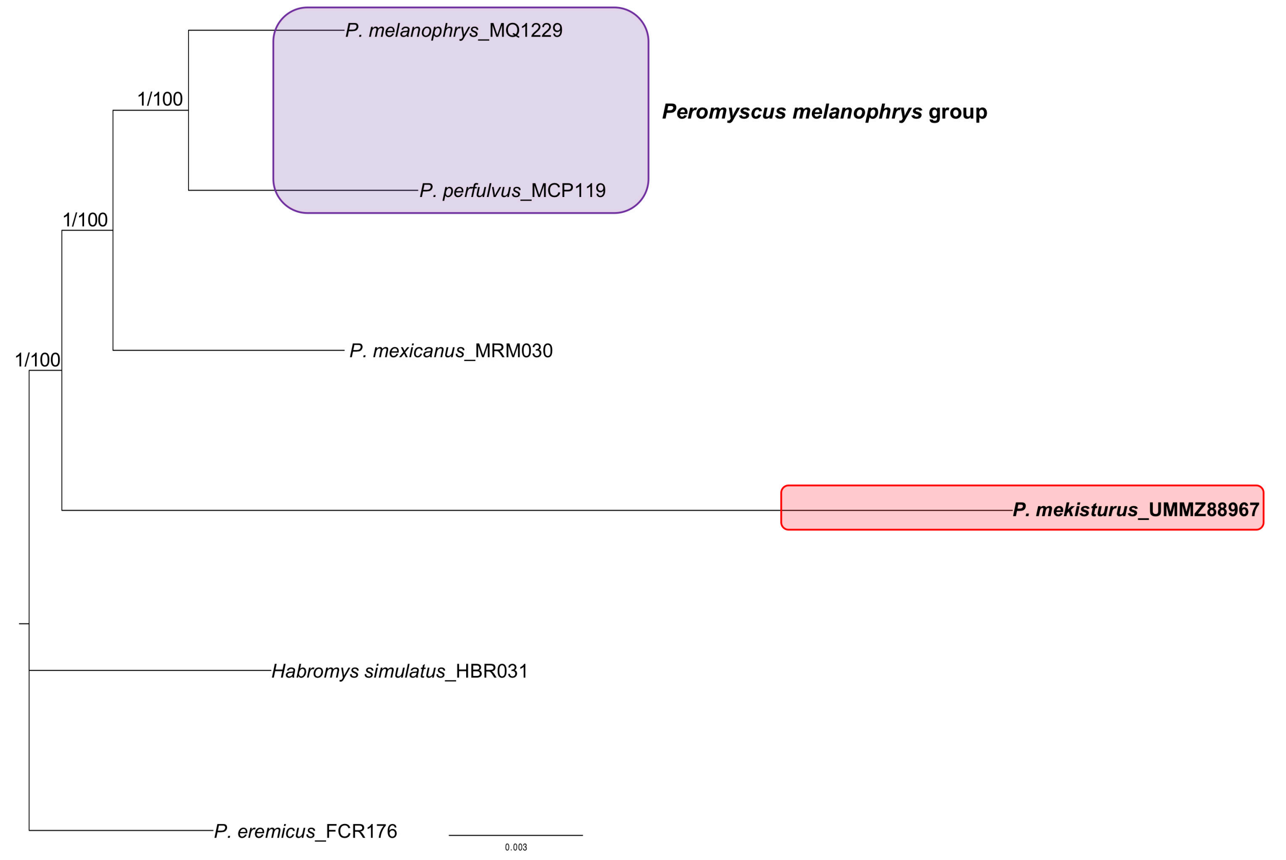


**FIGURE 1 |** Maximum Likelihood unrooted phylogeny based on a 75% UCE matrix with 2,436 loci and Maximum Likelihood and Bayesian Inference unrooted phylogenies based on a 95% UCE matrix with 1,010 loci from six species. Nodal support is provided with posterior probability/bootstrap values. The purple block shows the composition of the *Peromyscus melanophrys* species group and the red block highlights the phylogenetic position of *P. mekisturus*.

*P. melanophrys*, *P. perfulvus*, and *P. eremicus*. We did not include the rest of the samples in the analyses because of the low percentage of mitogenome obtained (35–50%). We also obtained the complete mitogenome of *P. mekisturus* (GenBank accession number MT078818) using the myBaits® Mito kit designed for *M. musculus*. The number of reads mapped and average coverage are shown in **Table 1**. Comparison of the *P. mekisturus* partial mitogenome (ca. 5,200 bp and mean coverage = 13×) obtained as off-target sequences and the complete mitogenome (15,975 bp and mean coverage = 581.7×) obtained using myBaits kit showed a perfect sequence match distributed along 2 rRNA, 7 tRNA, 9 protein coding regions and the *D-loop* region.

The comparison between the sequences from the same specimen obtained by Sanger sequencing in Castañeda-Rico et al. (2014) versus NGS showed that of a total of 1,307 bp, only 2 bp were different for *P. perfulvus* and 20 bp for *P. melanophrys*. For *P. perfulvus*, changes were found in *ND3* and *ND4L* genes whereas for *P. melanophrys* in *ND4L* and *ND4* genes. These represent a pairwise identity of 99.85% for *P. perfulvus* and 98.5% for *P. melanophrys*. However, for *P. mekisturus* we found 162 bp differences distributed along the *ND3*, *tRNA-Arg*, *ND4L*, and partial *ND4* genes. The pairwise identity between the two *P. mekisturus* sequences was 87.61%. To resolve the discrepancy between the previous published sequences and those out the NGS analysis, we Sanger sequenced five fragments from Castañeda-Rico et al. (2014) for the *P. mekisturus* sample, but only three were successfully amplified. The total length of the sequenced fragment was 545 bp (*tRNA-Gly* – 39 bp, *ND3* – 329 bp, and *ND4L* – 177 bp; GenBank accession number MT078814). The 545 bp fragment sequenced matched 100% with the homologous gene sequences obtained with NGS. On the other hand, when we compared these newly generated Sanger sequences with the ones generated by Castañeda-Rico et al. (2014) we found that only 489 bp (89.7%) matched.

Furthermore, the mitogenome obtained using a *de novo* assembly (mean coverage = 22.3×) matched completely with the one obtained mapping to reference genome. However, the *de novo* mitogenome assembly had ca. 9% missing data. Based on completeness and higher coverage (**Table 1**), we decided to use the mitogenome obtained from mapping for all the analyses. MEGAN analysis allowed us to identify three taxonomic levels within the reads analyzed as follows: family Cricetidae (53.1%), subfamily Neotominae (40.1%), and genus *Peromyscus* (6.8%). The result of mapDamage2 analysis showed a weak signal of damage patterns typical of ancient DNA (**Supplementary Figure S1**).

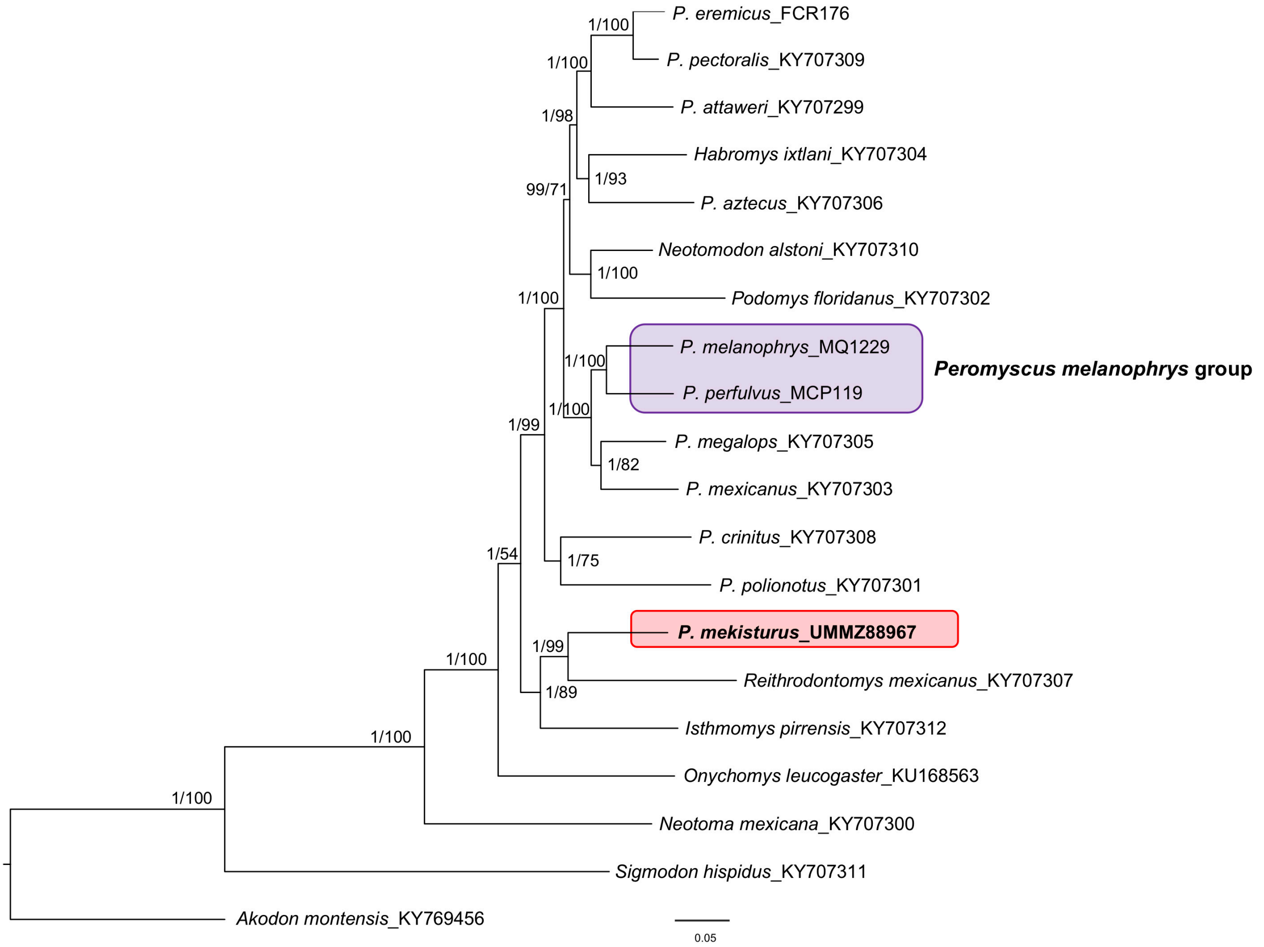


**FIGURE 2 |** Bayesian Inference and Maximum Likelihood phylogeny based on mitogenomes from 20 species. Nodal support is provided with posterior probabilities/bootstrap values. The purple block shows the composition of the *Peromyscus melanophrys* species group and the red block highlights the phylogenetic position of *P. mekisturus*.

The final dataset for our mitogenome phylogenetic analysis contained 20 mitogenomes from different species including members of the subfamily Neotominae, with at least one representative of eight out of the 13 *Peromyscus* species groups, and Sigmodontinae. The phylogenetic tree (**Figure 2**) showed that *P. melanophrys* and *P. perfulvus* form a monophyletic group which is the sister clade of the *P. mexicanus* + *P. megalops* clade. However, *P. mekisturus* is sister to *Reithrodontomys mexicanus* and it is more closely related to *Isthmomys pirrensis* than to any other member of the genus *Peromyscus*. The closer phylogenetic relationship of *P. mekisturus* to *R. mexicanus* is strongly supported (posterior probability value = 1 and bootstrap value = 99). The split between *P. mekisturu*s and other members of the genus *Peromyscus* is strongly supported with BI analysis (posterior probability value = 1) but not with ML analysis (bootstrap value = 54). The BI mitogenome phylogenetic trees obtained using: (1) no partitions, (2) by gene and codon position, and (3) by codon position, tRNA, rRNA and *D-loop* showed a similar topology with high posterior probability values. The only differences among them were the position of *O. leucogaster*, *P. crinitus*, *P. polionotus*, and the *P. floridanus* + *N. alstoni* clade. However, the phylogenetic relationship between *P. mekisturus* + *R. mexicanus* and *I. pirrensis* remained highly supported (**Figure 2** and **Supplementary Figures S2, S3**). The BI tree with no partitions and the ML tree showed the same topology (**Figure 2**).

The phylogenetic analyses of the *Cytb* gene allowed us to include representatives from all 13 *Peromyscus* species groups in order to test the position of *P. mekisturus* with respect to a denser taxon sampling of the group, as well as additional members of the subfamilies Neotominae and Sigmodontinae. We analyzed a total of 138 sequences from a total of 67 different rodent taxa. We found some minor changes between the topology obtained from BI and ML analyses. However, the result of both of these analyses confirmed that *P. mekisturus* is more closely related to the genera *Reithrodontomys* and *Isthmomys* than to *Peromyscus* (**Figures 3**, **4**). Although the relationships among the genera *Reithrodontomys* and *Isthmomys* and *P. mekisturus* are not well resolved here (low posterior probability and bootstrap values), the *Cytb* topology agrees with the phylogeny based on

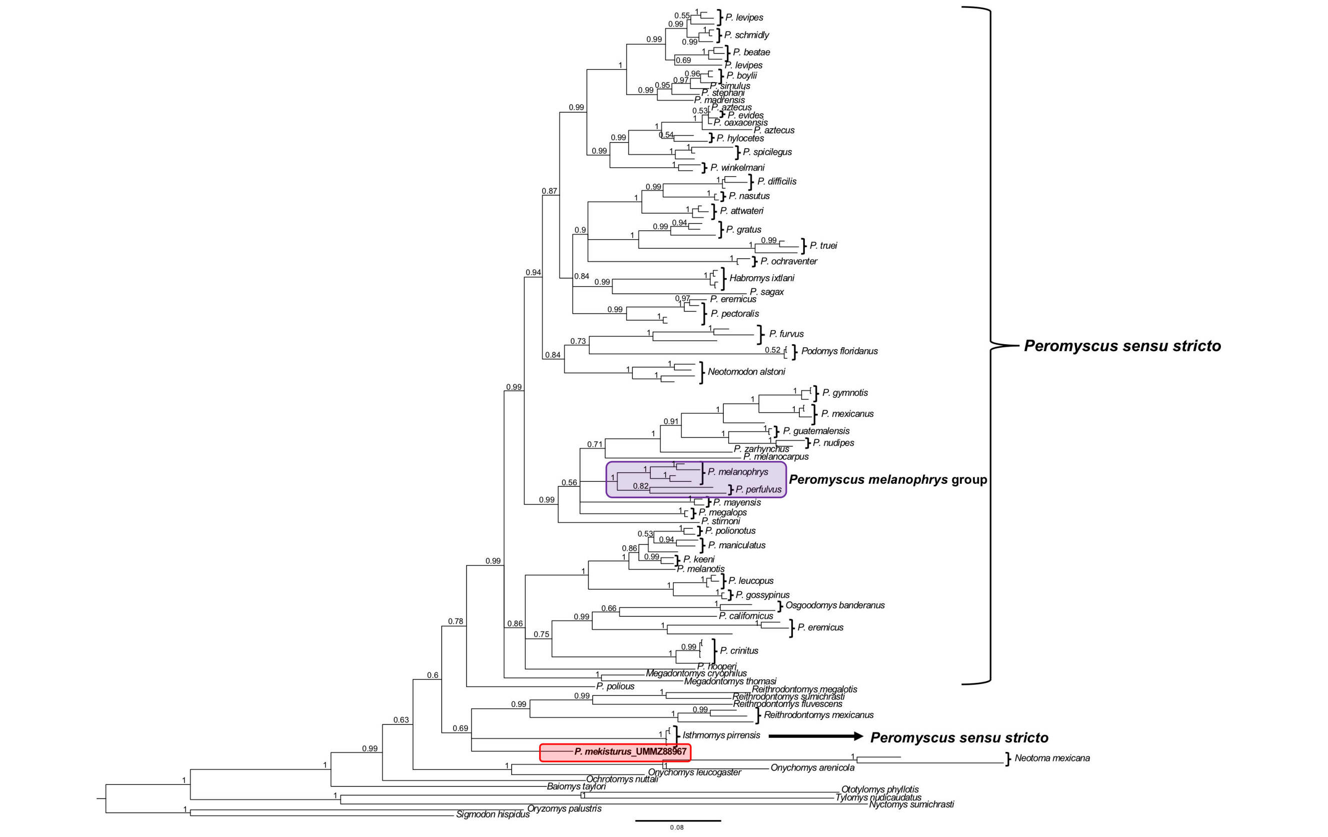


**FIGURE 3 |** Bayesian Inference phylogeny based on the *Cytb* gene from 64 species. Nodal support is provided with posterior probabilities values. The purple block shows the composition of the *Peromyscus melanophrys* species group and the red block highlights the phylogenetic position of *P. mekisturus*.

mitogenomes, and places *P. mekisturus* as the sister species of the genus *Reithrodontomys*.

The BI and ML analyses using multiple mitochondrial genes included a total of 119 sequences from a total of 22 different rodent taxa (**Figures 5**, **6**), including the *P. mekisturus* sequence generated using Sanger sequencing (*P. mekisturus*_KF885810, Castañeda-Rico et al., 2014) and the sequence generated in this study with NGS (*P. mekisturus*_UMMZ88967). This dataset also included at least one individual from eight out of the 13 *Peromyscus* species groups, and additional samples of the subfamilies Neotominae and Sigmodontinae. The phylogenetic tree (**Figures 5**, **6**) confirmed the position of *P. mekisturus*_KF885810 within one of the clades of *P. melanophrys* as Castañeda-Rico et al. (2014) previously showed. However, the new sample generated with NGS was placed as the sister species of *Reithrodontomys mexicanus*. Although the posterior probability value (0.8628, **Figure 5**) and the bootstrap value (39, **Figure 6**) of this node are not high, *P. mekisturus*_UMMZ88967 is more closely related to the genera *Reithrodontomys* and *Isthmomys* than to any other member of the genus *Peromyscus*. The ML tree shows, in general, low support values and slightly different topologies compared to the BI tree. However, the phylogenetic relationship between *P. mekisturus* and *Reithrodontomys* is the same.

# DISCUSSION

## Phylogenetic Inferences

Here, we used recent developments in NGS technology and ancient DNA protocols to help resolve the controversial phylogenetic position of the critically endangered Puebla deer mouse *P. mekisturus*. In doing so, we provide the first genomic study to generate nuclear data for *P. mekisturus*, and the first to include all members in the *Peromyscus melanophrys* species group recognized to date. We analyzed the same sample of *P. mekisturus* that Castañeda-Rico et al. (2014) used in their study (Hooper's record collected in 1947), where they sequenced mitochondrial genes by traditional Sanger sequencing. We did not perform a total evidence analysis, using UCEs and mitogenomes, because the rate of evolution of UCEs is still not well understood and it remains an area of research (Faircloth et al., 2015; Tangliacollo and Lanfear, 2018). In addition, the rate of evolution of between UCEs and mitogenomes would likely be very different and consequently these data would be partitioned anyway. Therefore, we analyzed and discussed nuclear and mitochondrial data independently. Our nuclear phylogenetic analysis using UCEs and mitogenomes (**Figures 1**, **2**) confirmed the monophyly of *P. melanophrys* and *P. perfulvus*, which had been previously proposed based on morphological characters

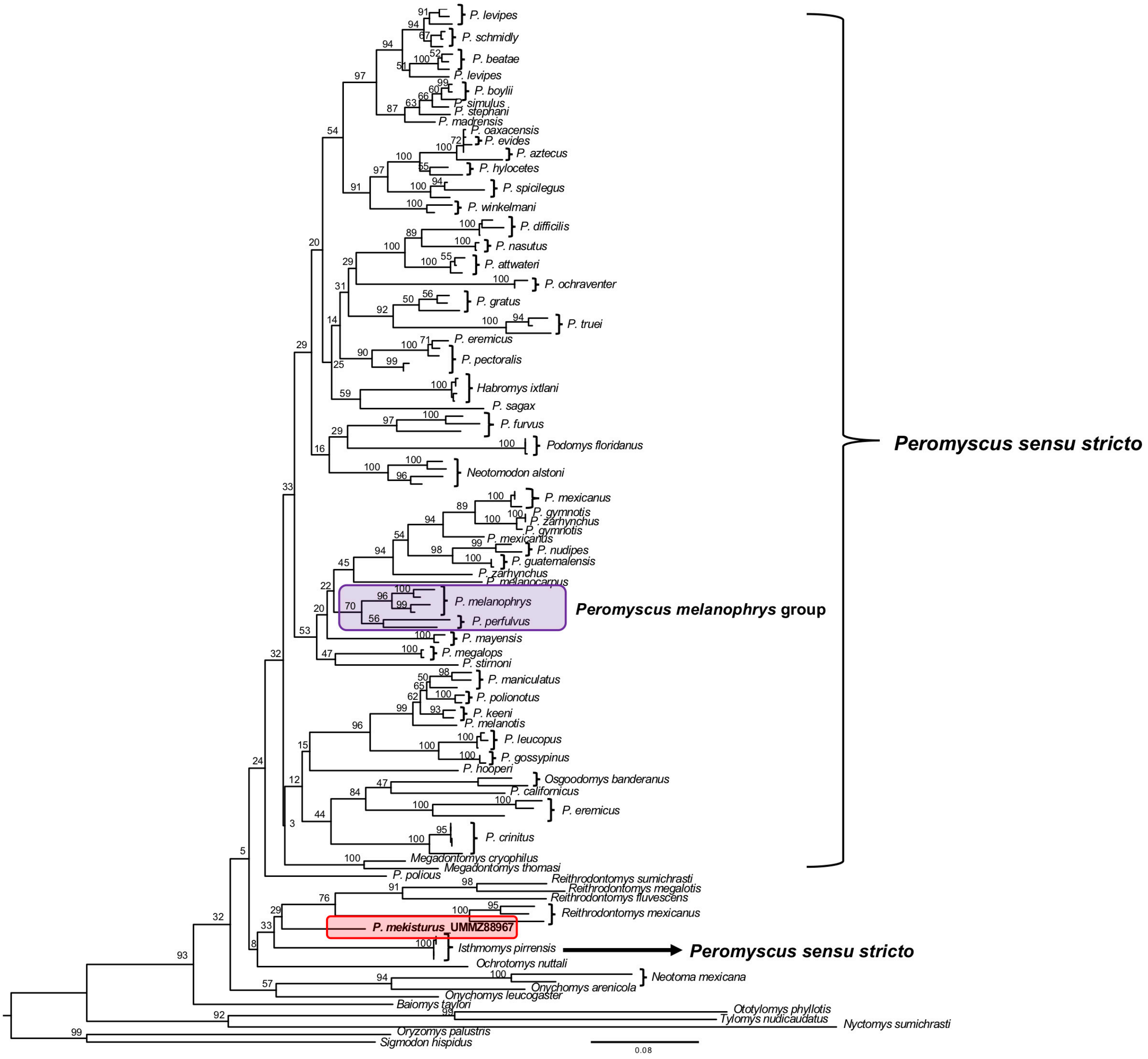


**FIGURE 4 |** Maximum Likelihood phylogeny based on the *Cytb* gene from 64 species. Nodal support is provided with bootstrap values. The purple block shows the composition of the *Peromyscus melanophrys* species group and the red block highlights the phylogenetic position of *P. mekisturus*.

and a few mitochondrial genes (Osgood, 1909; Hall and Kelson, 1952; Hooper and Musser, 1964; Hooper, 1968; Carleton, 1989; Bradley et al., 2007; Castañeda-Rico et al., 2014). However, contrary to previous hypotheses placing *P. mekisturus* in the *Peromysus melanophrys* species group (Osgood, 1909; Carleton, 1989; Musser and Carleton, 1993, 2005; Castañeda-Rico et al., 2014), we found that both UCEs and mitogenomes (**Figures 1**, **2**) place *P. mekisturus* outside of the monophyletic clade containing the other putative members of the *Peromyscus melanophrys* species group (*P. melanophrys* and *P. perfulvus*). Furthermore, our more densely taxonomically sampled mitochondrial DNA (mtDNA) phylogenies (**Figures 2–6**) support the placement of *P. mekisturus* as the sister species of *Reithrodontomys* and closely related to *Isthmomys*. We conclude that *P. mekisturus* is not part of the *Peromyscus melanophrys* species group (supported by nuclear and mitochondrial data) and it is placed as the sister species of the genus *Reithrodontomys* and closely related to the genus *Isthmomys* (supported by mitochondrial data) contrary to the findings of previous morphological and genetic studies (Osgood, 1909; Hall and Kelson, 1952; Hooper and Musser, 1964; Hooper, 1968; Carleton, 1989; Musser and Carleton, 1993, 2005; Bradley et al., 2007; Castañeda-Rico et al., 2014).

Given that some *Peromyscus* species groups were somewhat underrepresented in the mitogenome dataset (we were only able to include members of eight of 13 groups), we sought to further confirm the phylogenetic position of *P. mekisturus*

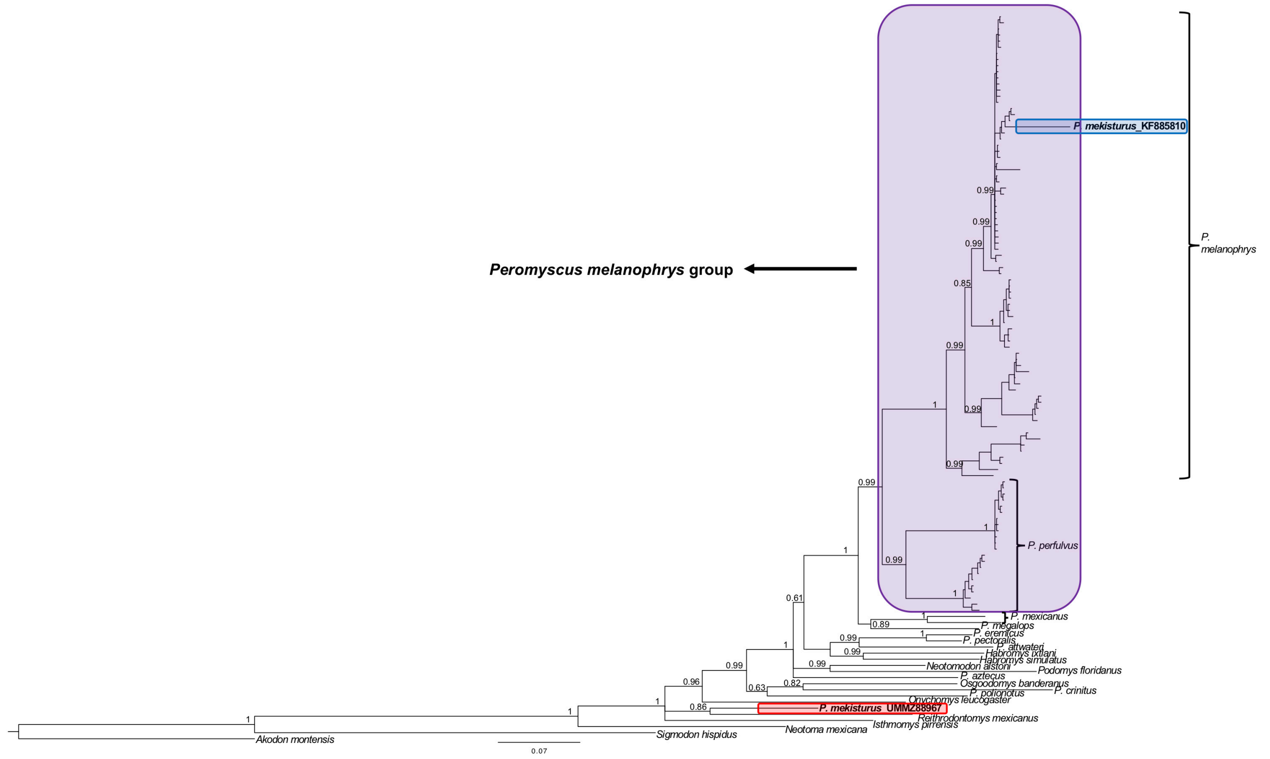


**FIGURE 5 |** Bayesian Inference phylogeny based on *ND3, tRNA-Arg, ND4L* and *partial ND4* genes (referred in the text as "multiple mitochondrial genes") from 23 species. Nodal support is provided with posterior probabilities values. The purple block shows the composition of the *Peromyscus melanophrys* species group, the red block highlights the phylogenetic position of *P. mekisturus*_UMMZ88967 (sequence obtained by next-generation sequencing), and the blue block shows the placement of *P. mekisturus*_KF885910 (sequence obtained by Sanger sequencing).

using the *Cytb* gene to reconstruct a more complete phylogeny including members of all 13 *Peromyscus* species groups as well as representatives of the more distantly related neotomine and sigmodontine rodents (**Figures 3**, **4**). Although more weakly supported (i.e., with low posterior probabilities and bootstrap values for some clades), the *Cytb* phylogeny confirmed our well-supported mitogenome phylogeny, resulting in the placement of *P. mekisturus* as more closely related to the genera *Reithrodontomys* and *Isthmomys* than to any other species of *Peromyscus*. Furthermore, we note that our near-complete mitogenome phylogeny and *Cytb* phylogeny are consistent with previous published phylogenies that did not sample *P. mekisturus* (e.g., mitogenome phylogeny: Sullivan et al., 2017; Sanger phylogenies using a few genes: Bradley et al., 2007; Miller and Engstrom, 2008; Platt et al., 2015).

This is the first time that any kind of published evidence suggests that *P. mekisturus* is the sister species of the genus *Reithrodontomys*. However, more analyses are needed to determine how the *P. mekisturus* + *Reithrodontomys* and *Isthmomys* clade should be classified, and where it sits in the phylogeny. It is important to mention that no other phylogenetic analysis has ever suggested that the genus *Reithrodontomys* should be nested within *Peromyscus* (Sullivan et al., 2017) despite the close relationship between *Reithrodontomys* and *Isthmomys* (Hooper and Musser, 1964; Bradley et al., 2007; Miller and Engstrom, 2008; Platt et al., 2015; Sullivan et al., 2017). However, *Isthmomys* is recognized at the generic (*sensu stricto*) or subgeneric (*sensu lato*) level within *Peromyscus* (**Figures 5**, **6**). Therefore, *P. mekisturus* could be in the same position as *Isthmomys*, i.e., still considered part of the genus *Peromyscus* (*sensu lato* or *sensu stricto*) but suggesting paraphyly as several previous studies have shown (Bradley et al., 2007; Miller and Engstrom, 2008; Platt et al., 2015; Sullivan et al., 2017).

A taxonomic revision of *P. mekisturus* is clearly warranted. We recommend that nuclear genetic data (UCEs) from more representatives of the subfamily Neotominae be incorporated in order to conclusively resolve the phylogenetic position of *P. mekisturus*. Given our striking and unexpected results, we also recommend that the taxonomic revision incorporates a morphological re-evaluation for *P. mekisturus*.

## Sanger Sequencing and Next-Generation Sequencing of Ancient DNA: The Case of *Peromyscus mekisturus*

Phylogenetic results obtained for *P. mekisturus* based on Sanger sequencing in a previous study (*P. mekisturus*_KF885810: Castañeda-Rico et al., 2014) and NGS sequencing in this study (*P. mekisturus*_UMMZ88967: this study) were strikingly different. Our re-analysis using only the mitochondrial genes

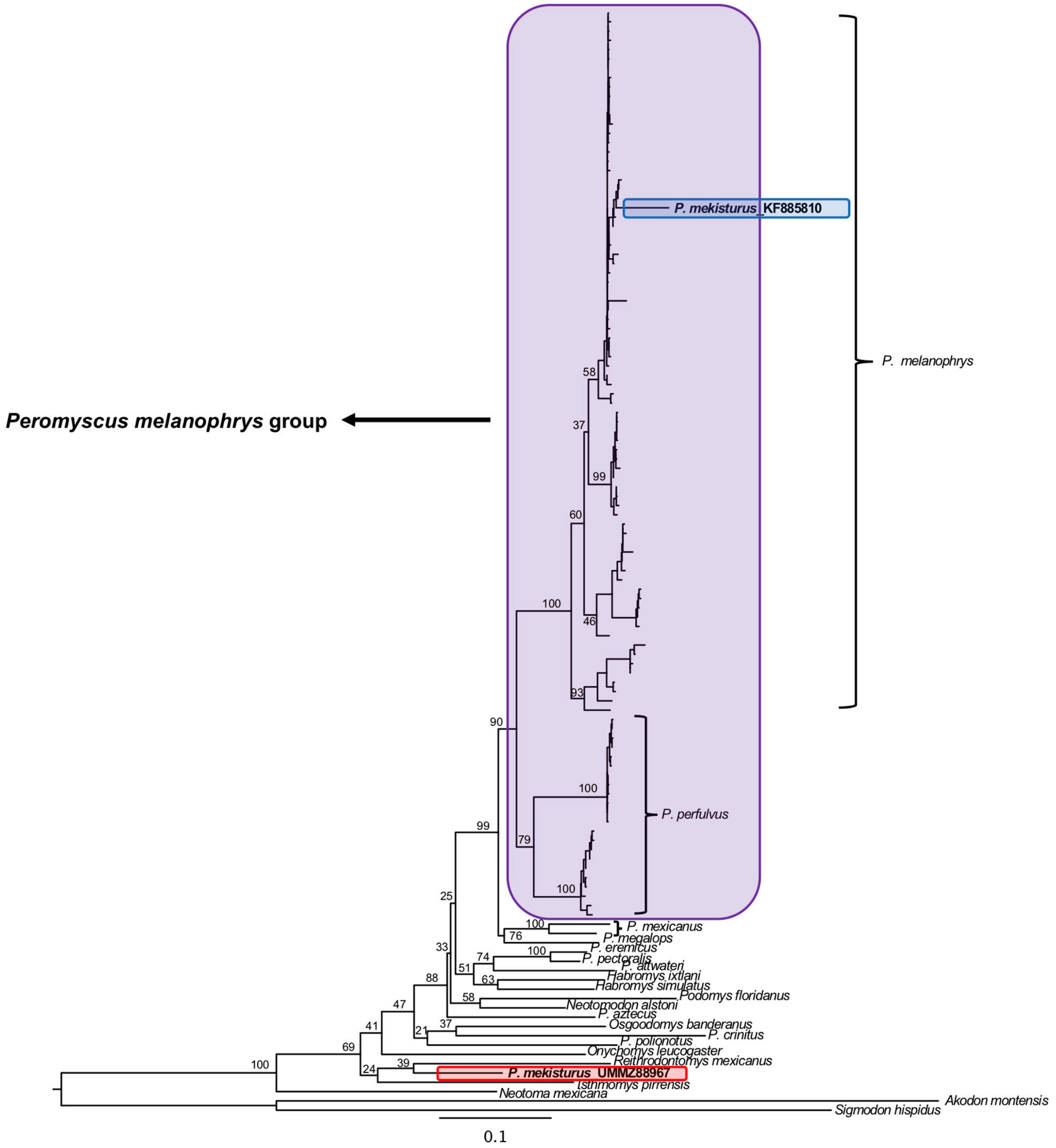


**FIGURE 6 |** Maximum Likelihood phylogeny based on *ND3, tRNA-Arg, ND4L* and *partial ND4* genes (referred in the text as "multiple mitochondrial genes") from 23 species. Nodal support is provided with bootstrap values. The purple block shows the composition of the *Peromyscus melanophrys* species group, the red block highlights the phylogenetic position of *P. mekisturus*_UMMZ88967 (sequence obtained by next-generation sequencing), and the blue block shows the placement of *P. mekisturus*_KF885910 (sequence obtained by Sanger sequencing).

sequenced by Castañeda-Rico et al. (2014) (**Figures 5**, **6**) but including additional members of the subfamily Neotominae clearly showed that the Sanger-based and NGS-based sequences of *P. mekisturus* do not cluster together. The previously published Sanger sequence (*P. mekisturus*_KF885810) was again placed within a clade where all the other samples are considered to be *P. melanophrys* (same result as Castañeda-Rico et al., 2014), while our novel NGS sequence (*P. mekisturus*_UMMZ88967) was placed as the sister species of *R. mexicanus*, in agreement with the mitogenome and *Cytb* phylogenies (**Figures 2–4**). Thus, we can reject the limited number of members of the subfamily Neotominae in Castañeda-Rico et al. (2014) as the cause for the differences between the Sanger- and NGS-based phylogenetic hypotheses.

Instead, we propose that this discrepancy is most likely due to the different protocols and conditions used to obtain sequences from each sample of *P. mekisturus* (i.e., the use of a modern DNA facility vs. a dedicated ancient DNA facility and Sanger sequencing vs. NGS) better explain the phylogenetic discrepancies that we found. To clarify this, we investigated how the sequences (ca. 1,315 bp) in Castañeda-Rico et al. (2014) were generated. The *P. perfulvus* sample (*P. perfulvus*_MCP119) was an ethanol-preserved internal organ, therefore DNA was of sufficient quality and quantity to sequence it in two fragments, each with an average length of 700 bp (Castañeda-Rico et al., 2014). The two *P. perfulvus* sequences, from this study and from Castañeda-Rico et al. (2014) were nearly identical (*P. perfulvus*_KF885791 and *P. perfulvus*_MCP119). Samples from *P. melanophrys* and *P. mekisturus* were obtained from dried skin and dried skin + turbinate bones, respectively. Because of DNA degradation, sequences were amplified using specific primers designed to amplify an average of 217 bp for each fragment, from a total of eight separate PCR reactions (Castañeda-Rico et al., 2014). In this case, the comparison of the sequences yielded by the two sequencing methods showed more differences between the sequences (20 bp for *P. melanophrys* and 162 bp for *P. mekisturus*). Although the samples of *P. melanophrys* and *P. mekisturus* were both dried skins (in this study), the *P. melanophrys* specimen was collected in 1984 and the *P. mekisturus* in 1947, almost 40 years earlier. We were not able to determine whether the relative ages of the specimens affected the resulting sequence data or if it was due to other characteristics of the specimens (e.g., preservation method). Nevertheless, the few differences detected in *P. melanophrys* were not enough to influence the phylogenetic signal, but the *P. mekisturus* sequences had enough differences to affect the phylogenetic reconstruction resulting in conflicting hypotheses. Additionally, the fact that the partial mitochondrial sequence (545 bp) that we generated here using Sanger sequencing was a perfect match to the same fragments generated by NGS allowed us to further corroborate that the correct sequence for *P. mekisturus* was generated in this study (*P. mekisturus*_UMMZ88967). Therefore, we can conclude that the discrepancies found between the Sanger sequence (*P. mekisturus*_KF885810) from Castañeda-Rico et al. (2014) and the NGS sequences generated in this study can be attributed primarily to the use of an ancient DNA facility and rigorous protocols designed to control contamination which were not followed in Castañeda-Rico et al. (2014), and not to the use of Sanger sequencing vs. NGS methodologies. However, we should point out that it is much more challenging to obtain sequences using traditional Sanger method from museum specimens, and the use of specialized aDNA extraction and sequencing protocols should be carefully considered.

For decades, Sanger sequencing has been the gold standard sequencing technology (Berglund et al., 2011; Liu et al., 2012). However, in recent years the use of NGS technologies has dramatically increased owing to its higher throughput, reduced cost, and benefits of obtaining sequences from highly degraded ancient DNA samples (Berglund et al., 2011; Liu et al., 2012). Both Sanger and NGS are accurate, though some studies have disagreed on which is more accurate for the detection of low frequency mutations (e.g., Ihle et al., 2014; Arsenic et al., 2015; Beck et al., 2016). Most of the studies comparing sequences derived from the two methods show a high correlation between the two methodologies (Gunnarsdottir et al., 2011; Arsenic et al., 2015; Arias et al., 2018). Critically, all of these studies have been focused on clinical research using fresh samples (e.g., blood, organ tissues, etc.) and thus do not assess the additional difficulties and peculiarities that can affect the validity and quality of ancient DNA sequences obtained from both Sanger and NGS protocols. Ancient DNA samples are characterized and defined by their low yield and poor quality associated with damage accumulated over time. This results in the progressive fragmentation of DNA molecules into shorter fragments, and cytosines deaminating to uracils. There is also a high risk of contamination of samples with modern DNA (Pääbo et al., 2004; Fortes and Paijmans, 2015) and a low ratio of endogenous versus environmental and contaminant DNA (Knapp and Hofreiter, 2010; Fortes and Paijmans, 2015). The analysis of ancient samples is therefore particularly challenging, and for many years has been more focused on mtDNA which is expected to be better preserved and to provide higher endogenous content than nuclear DNA because of its higher copy numbers (Rowe et al., 2011; Fortes and Paijmans, 2015). Further contributing to these issues is the unpredictable preservation of DNA in ancient DNA sources. Several studies have proposed that the quality and quantity of the DNA is related to specimen age (Pääbo et al., 1990; Ellegren, 1994), the manner in which a specimen was prepared and stored during and after collection (Wandeler et al., 2007; Mason et al., 2011; Guschanski et al., 2013; McCormack et al., 2016; McDonough et al., 2018), and the difference of tissue type used to extract ancient DNA (Casas-Marce et al., 2010; McDonough et al., 2018). An additional challenge is that preservation is frequently intended to ensure long-term integrity of the museum specimen rather than its DNA (McDonough et al., 2018), thus affecting the success of genetic and genomic studies.

Sanger sequencing DNA from museum specimens is not an easy process. The degraded fragments of ancient DNA require numerous independent PCR amplifications of overlapping short fragments (Knapp and Hofreiter, 2010; Rowe et al., 2011), and many sets of novel primers need to be designed. Because they are designed to target variable regions to the organism under study, the primers often lose their universality, which is one of their principal benefits (McCormack et al., 2017). Additional problems include that ancient DNA extract is often limited in quantity. Sanger approaches using PCR will only target a tiny fraction of the molecules at the very top end of the fragment length distribution, thereby greatly increasing the risk of starting amplifications from single molecules. This effect can introduce contamination of the resulting sequences and even small overall amounts of contamination can lead to erroneous sequences (Gilbert et al., 2005; Knapp and Hofreiter, 2010). Traditional PCR amplification and Sanger sequencing of historical specimens require considerable resources and strict protocols and controls to obtain reliable results. Even if PCR is successful, it often succeeds from only one or a small number of DNA template copies and thus can propagate postmortem mutations into the resulting sequence trace (Rowe et al., 2011). Overall, NGS technology represents the best option for the analysis of ancient DNA (Rizzi et al., 2012; Lemmon and Lemmon, 2013; McCormack et al., 2017; Webster, 2018) as it favors short DNA fragments and genomic library preparation targets the entire DNA extract by ligating adapters to each end of a single DNA fragment. This means that species-specific primers are no longer necessary, effectively allowing all fragments present in the extraction to be sequenced (Fortes and Paijmans, 2015). Additionally, NGS capture or target enrichment methods allow us to target specific regions within the whole genome (Burrell et al., 2015), increasing the chances to target genes or regions of interest for specific studies.

We have shown with our own data that while NGS and Sanger methods can provide accurate and consistent results for modern tissue samples (here, an ethanol preserved organ sample of *P. perfulvus*) independent of the kind of DNA facility used for the study. However, our results suggest that there is a poor correlation between Sanger and NGS sequences obtained from ancient DNA samples (i.e., our dried skin and bones from prepared museum specimens of *P. melanophrys* and *P. mekisturus*) that were Sanger sequenced in a facility with other modern tissue samples. In contrast, we found that NGS and Sanger methods are accurate for ancient DNA samples, when the study is performed following strict protocols to control for contamination in a dedicated ancient DNA facility. In particular, it is apparent that the Sanger sequence *P. mekisturus*_KF885810 sample obtained in Castañeda-Rico et al. (2014) was affected by the inherent problems associated with Sanger sequencing of ancient DNA (i.e., low DNA quantity and quality, several independent PCR amplifications in small fragments, use of specific primers, high susceptibility to modern DNA contamination, etc.) magnified by the lack of an ancient DNA facility and protocols. Specifically, we conclude that some of the factors that caused the erroneous *P. mekisturus*_KF885810 Sanger sequence were: (i) cross-contamination of other *Peromyscus* samples processed in that same lab during extraction and/or PCR steps, (ii) a chimera sequence formation by recombining different template molecules (jumping PCR) during PCR reactions, and (iii) a high risk of contamination from the environment when not working in a dedicated ancient DNA facility. Notably, some of our historical samples are not affected by contamination nor show excessive discrepancies between sequences obtained using different methods of sequencing (e.g., *P. melanophrys*_MQ1229). This could be due to better quality and quantity of DNA compared to other samples (e.g., *P. mekisturus*_UMMZ88967). All of these issues have been suggested in other studies as problems when using Sanger sequencing to work with ancient DNA (Meyerhans et al., 1990; Pääbo et al., 1990; Lahr and Katz, 2009; Kircher et al., 2012).

Here, we reiterate that extreme caution and validation are necessary to ensure accurate results when sequencing ancient DNA derived from museum specimens (dried skin, bones, cartilage, osteoclasts, hair, teeth, claws, etc.). Controlling for contamination can be problematic, particularly without comparing sequences obtained by different sequencing methods or based on independent replicates preferably obtained in different laboratories. Even when negative controls appear to be free of contamination, as was the case of *P. mekisturus*_KF885810 in Castañeda-Rico et al. (2014), a chimeric sequence could be obtained. Therefore, we strongly recommend the use of a dedicated ancient DNA facility with strict protocols for dealing with contamination when working with museum specimens, especially with those for which few or no modern samples exist in scientific collections or can no longer be found in the wild.

## Conservation Implications

A large part of the earth's biodiversity is still unknown to science, and a frequent misconception of the discovery process is that new species are only recognized as new when they are discovered in the field (Fontaine et al., 2012). However, this is not always the case. In fact, scientific collections can act as a reservoir of potential new species that are waiting to be discovered and described (Green, 1998; Bebber et al., 2010; Fontaine et al., 2012). An additional challenge is that most of these new species are cataloged as rare and they are represented by singletons (species only known from a single specimen), uniques (species that have only been collected once), or doubletons (species with a new singleton specimen being discovered in the process of additional sampling) (Fontaine et al., 2012; Lim et al., 2012). Rarity is not a new phenomenon and its commonness has been demonstrated in the description of many singleton species in the literature (Lim et al., 2012). The description of the olinguito (*Bassaricyon neblina*) by Helgen et al. (2013) is an example of the description of a species new to science using ancient DNA and museum specimens.

With a biodiversity crisis that predicts massive extinctions and an increase in time between discovery and description of new species, taxonomists will increasingly be describing species that are already extinct in the wild from museum collections (Fontaine et al., 2012). The study of extinct or highly endangered species, and extirpated populations or species, is an important application of NGS to museum specimens because no high-quality DNA will likely ever be available for many of these species (Rowe et al., 2011; McCormack et al., 2017), and their knowledge can contribute to studies of biodiversity loss, conservation and population genetics (Roy et al., 1994; Pichler et al., 2001; Martinez-Cruz et al., 2007; Peery et al., 2010; Rowe et al., 2011).

Our study demonstrates that ancient DNA approaches with appropriate rigorous protocols using museum specimens combined with high-throughput sequencing offers an opportunity to re-evaluate previous phylogenetic hypotheses. These new studies can validate previous results with additional data (e.g., complete mitogenomes and nuclear data) or reveal new conclusions, resulting in different phylogenetic hypotheses. This new evidence could offer a robust and reliable resolution of the evolutionary history and taxonomic status of species, resulting in the re-evaluation of previous conservation strategies or the establishment of new conservation programs in order to better protect and manage biodiversity. Information used to decide if a species is at risk of extinction and its threat category are usually based on ecological and demographic data such as the number of known individuals, current or projected declines in population size and the extent of occurrence or distribution (IUCN, 2014; Carneiro Muniz et al., 2019). However, for rare species where little additional information is available, genetic data is extremely important (Carneiro Muniz et al., 2019).

The case of the Puebla deer mouse, *P. mekisturus*, is very surprising. This species has always been recognized as a valid taxon since it was first described (Merriam, 1898; Osgood, 1909; Hooper, 1947; Carleton, 1989; Musser and Carleton, 1993, 2005; Álvarez-Castañeda and González-Ruiz, 2008) and there is no available evidence to suggest that the pattern that we observed was the result of hybridization or any other artifact. Furthermore, this species is only known from two specimens and two localities – Merriam's (1898) holotype from Chalchicomula, and Hooper's (1947) record from Tehuacán, both in the state of Puebla, Mexico. Therefore, little is known about its biology and information relevant to determining its conservation status remains incomplete. Despite this lack of information, this species falls into the Critically Endangered (CR) category of extinction risk in the IUCN and the Threatened category (A) in the Mexican Official Norm NOM-059-SEMARTNAT-2010 (Secretaría de Medio Ambiente y Recursos Naturales [SEMARNAT], 2010). Sánchez-Cordero et al. (2005) estimated that only ca. 41.66% of its habitat (pine-oak forest and arid zones) remains untransformed. That result was calculated more that 14 years ago; therefore, we suspect that the current remaining habitat is even smaller in area. Additionally, this species has not been detected or recorded in the wild for more than 70 years. The placement of this species within an Extinct (EX) category would definitely represent a loss of biodiversity. However, if our results are confirmed with morphological and additional nuclear data, we would be facing the extinction of a unique lineage of rodents. This would be a tremendous loss of evolutionary uniqueness and distinctiveness. *P. mekisturus* is a great example of how new efforts and studies are needed in order to discover and preserve our biodiversity. These results could be used by conservation managers and policymakers to minimize the impact of anthropogenic development on Earth's biodiversity and help design urgent conservation strategies for pine-oak forest and arid zones.

## DATA AVAILABILITY STATEMENT

The raw data generated for this study can be found in the GenBank under BioProject: PRJNA606805 and under GenBank accession numbers given in **Supplementary Table S1**.

## ETHICS STATEMENT

Ethical review and approval was not required for the animal study because we used only museum specimens deposited in scientific collections for this study. We obtained DNA from those specimens. Also, we used some published data available on GenBank.

## AUTHOR CONTRIBUTIONS

SC-R, CE, and JM secured funding. SC-R, LL-P, CE, and JM designed the study. SC-R performed the specimen sampling and lab experiments, analyzed and archived the data, and produced the figures, and wrote the manuscript with contributions from all co-authors. All authors read and approved the submitted version.

## FUNDING

SC-R received a fellowship from the Consejo Nacional de Ciencia y Tecnología (CONACyT) for a postdoctoral stay at the Smithsonian Conservation Biology Institute and a Fellowship from Smithsonian-Mason School of Conservation and George Mason University.

## ACKNOWLEDGMENTS

We thank the specimen collectors, collection managers, and the curators of the Museo de Zoología "Alfonso L. Herrera", Facultad de Ciencias, UNAM (MZFC) and the University of Michigan Museum of Zoology (UMMZ) that granted sample loans. We specially thank Philip Myers (UMMZ) for the loan of *P. mekisturus* sample to the first author during the fulfillment of her Ph.D. dissertation research. We thank Rob C. Fleischer, Nancy R. McInerney, Lillian D. Parker, and Michael G. Campana for support of this research at the CCG; Sabina M. Lindley from the FDA-Center for Food Safety and Nutrition for donated MiSeq kits used in this study; and Lillian D. Parker, Madhvi X. Venkatraman, Anna M. Kearns, and the CCG writing group for comments that greatly improved this manuscript. This research was conducted using the Smithsonian Institution High Performance Cluster (SI/HPC). We wish to thank the Associate Editor Michael Knapp and two reviewers for their careful and insightful comments that greatly improved the quality of our manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.00094/full#supplementary-material

## REFERENCES

Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L., Hale, M. L., et al. (2012). The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. B* 279, 4724–4733. doi: 10.1098/rspb.2012.1745

Andrews, S. (2010). *FastQC: A Quality Control Tool For High Throughput Sequence Data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Arias, A., López, P., Sánchez, R., Yamamura, Y., and Rivera-Amill, V. (2018). Sanger and next-generation sequencing approaches to evaluate HIV-1 virus in blood compartments. *Int. J. Environ. Res. Public Health* 15:1697. doi: 10.3390/ijerph15081697

Arsenic, R., Treue, D., Lehmann, A., Hummel, M., Dietel, M., Denkert, C., et al. (2015). Comparison of targeted next-generation sequencing and Sanger sequencing for the detection of PIK3CA mutations in breast cancer. *BMC Clin. Pathol.* 15:20. doi: 10.1186/s12907-015-0020-6

Álvarez-Castañeda, S., and González-Ruiz, N. (2008). "Análisis preliminar de las relaciones filogenéticas entre los grupos de especies del género *Peromyscus*," in *Avances en el estudio de los mamíferos de México II*, eds C. Lorenzo, E. Espinoza, and J. Ortega (México: CIB. Universidad Veracruzana), 5–26.

Bebber, D. P., Carine, M. A., Wood, J. R. I., Wortley, A. H., Harris, D. J., Prance, G. T., et al. (2010). Herbaria are a major frontier for species discovery. *Proc. Natl. Acad. Sci. U.S.A.* 107, 22169–22171. doi: 10.1073/pnas.1011841108

Beck, T. F., Mullikin, J. C., and Biesecker, L. G. (2016). Systematic evaluation of Sanger validation of next-generation sequencing variants. *Clin. Chem.* 62, 647–654. doi: 10.1373/clinchem.2015.249623

Bergesten, J. (2005). A review of long-branch attraction. *Cladistics* 21, 163–193. doi: 10.1111/j.1096-0031.2005.00059.x

Berglund, E. C., Kiialainen, A., and Syvänen, A. C. (2011). Next-generation sequencing technologies and applications for human genetic history and forensics. *Ivestig. Genet.* 2:23. doi: 10.1186/2041-2223-2-23

Bradley, R. D., Durish, N., Rogers, D., Millar, J., Engstrom, M., and Kilpatrick, W. (2007). Toward a molecular Phylogeny for *Peromyscus*: evidence from mitochondrial cytochrome-b sequences. *J. Mamm.* 88, 1146–1159. doi: 10.1644/06-mamm-a-342r.1

Buerki, S., and Baker, W. J. (2016). Collections-based research in the genomic era. *Biol. J. Linn. Soc.* 117, 5–10. doi: 10.1111/bij.12721

Burrell, A. S., Disotell, T. R., and Bergey, C. M. (2015). The use of museum specimens with high-throughput DNA sequencers. *J. Hum. Evol.* 79, 35–44. doi: 10.1016/j.jhevol.2014.10.015

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Carleton, M. D. (1989). "Systematics and evolution," in *Advances in the study of Peromyscus (Rodentia)*, eds G. L. Kirkland and J. Layne (Lubbock, TX: Texas Tech University Press), 7–141.

Carneiro Muniz, A., Lemos-Filho, J. P., Santiago de Oliveira Buzatti, R., Correa Ribeiro, P. C., Moreira Fernandes, F., and Lovato, M. B. (2019). Genetic data improve the assessment of the conservation status based only on herbarium records of a Neotropical tree. *Sci. Rep.* 9:5693. doi: 10.1038/s41598-019-41454-0

Carøe, C., Gopalakrishnan, S., Vinner, L., Mark, S. S. T., Sinding, M. H. S., Samaniego, J. A., et al. (2017). Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* 9, 410–419. doi: 10.1111/2041-210x.12871

Casas-Marce, M., Revilla, E., and Godoy, J. A. (2010). Searching for DNA in museum specimens: a comparison of sources in a mammal species. *Mol. Ecol. Resour.* 10, 502–507. doi: 10.1111/j.1755-0998.2009.02784.x

Castañeda-Rico, S., León-Paniagua, L., Vázquez-Domínguez, E., and Navarro-Sigüenza, A. G. (2014). Evolutionary diversification and speciation in rodents of the Mexican lowlands: the *Peromyscus* melanophrys species group. *Mol. Phylogenet. Evol.* 70, 454–463. doi: 10.1016/j.ympev.2013.10.004

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552. doi: 10.1093/oxfordjournals.molbev.a026334

Chevreux, B., Wetter, T., and Suhai, S. (1999). Genome sequence assembly using trace signals and additional sequence information. *Comput. Sci. Biol.* 99, 45–56.

Church, G. M. (2006). Genomes for all. *Sci. Am.* 294, 46–54.

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods.* 9:772. doi: 10.1038/nmeth.2109

Dawson, W. (2005). "Peromyscine biogeography, Mexican topography and pleistocene climatology," in *Contribuciones Mastozooloìgicas en Homenaje a Bernardo Villa*, eds V. Saìnchez-Cordero and R. Medelliìn (Meìxico: UNAM-CONABIO), 145–156.

Ellegren, H. (1994). "Genomic DNA from museum bird feathers," in *Ancient DNA*, eds B. Herrmann and two methodologies S. Hummel (New York, NY: Springer), 211–217. doi: 10.1007/978-1-4612-4318-2_15

Faircloth, B. C. (2013). Illumiprocessor: a Trimmomatic wrapper for parallel adapter and quality trimming. doi: 10.6079/J9ILL

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* 32, 786–788. doi: 10.1093/bioinformatics/btv646

Faircloth, B. C., Branstetter, M. G., White, N. D., and Brady, S. G. (2015). Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Mol. Ecol. Resour.* 15, 489–501. doi: 10.1111/1755-0998.12328

Faircloth, B. C., and Glenn, T. C. (2012). Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. *PLoS One* 7:e42543. doi: 10.1371/journal.pone.0042543

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., and Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* 61, 717–726. doi: 10.1093/sysbio/sys004

Fontaine, B., Perrard, A., and Bouchet, P. (2012). 21 years of shelf life between discovery and description of new species. *Curr. Biol.* 22, 943–944. doi: 10.1016/j.cub.2012.10.029

Fortes, G. G., and Paijmans, J. L. A. (2015). "Analysis of whole mitogenomes from ancient samples," in *Whole Genome Amplification: Methods and Protocols*, ed. T. Kroneis (US: Human Press), 1–17.

Gilbert, M. T. P., Bandelt, H. J., Hofreiter, M., and Barnes, I. (2005). Assessing ancient DNA studies. *Trends Ecol. Evol.* 20, 541–544. doi: 10.1016/j.tree.2005.07.005

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Green, S. V. (1998). The taxonomic impediment in orthopteran research and conservation. *J. Insect Conserv.* 2, 151–159.

Guindon, S., and Gascuel, O. (2003). A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. *Syst. Biol.* 52, 696–704. doi: 10.1080/10635150390235520

Gunnarsdottir, E. D., Li, M., Bauchet, M., Finstermeier, K., and Stoneking, M. (2011). High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome Res.* 21, 1–11. doi: 10.1101/gr.107615.110

Guschanski, K., Krause, J., Sawyer, S., Valente, L. M., Bailey, S., Finstermeier, K., et al. (2013). Next-generation museomics disentangles one of the largest primate radiations. *Syst. Biol.* 62, 539–554. doi: 10.1093/sysbio/syt018

Hall, E. R., and Kelson, K. R. (1952). Comments on the taxonomy and geographic distribution of some North American rodents. *Univ. Kans. Publ. Mus. Nat. Hist.* 5, 343–371.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/TN. *Nucleic Acids Symp. Ser.* 41, 95–98. doi: 10.1007/0-387-27604-1_8

Hawkins, M. T. R., Hofman, C. A., Callicrate, T., McDonough, M. M., Tsuchiya, M. T. N., Gutiérrez, E. E., et al. (2016a). In-solution hybridization for mammalian mitogenome enrichment: pros, cons and challenges associated with multiplexing degraded DNA. *Mol. Ecol. Resour.* 16, 1173–1188. doi: 10.1111/1755-0998.12448

Hawkins, M. T. R., Leonard, J. A., Helgen, K. M., McDonough, M. M., Rockwood, L. L., and Maldonado, J. E. (2016b). Evolutionary history of endemic Sulawesi squirrels constructed from UCEs and mitogenomes sequenced from museum specimens. *BMC Evol. Biol.* 16:80. doi: 10.1186/s12862-016-0650-z

Helgen, K. M., Pinto, C. M., Kays, R., Helgen, L. E., Tsuchiya, M. T. N., Quinn, A., et al. (2013). Taxonomic revision of the olingos (Bassaricyon), with description of a new species, the Olinguito. *ZooKeys* 324, 1–83. doi: 10.3897/zookeys.324.5827

Hogan, K., Heding, M., Koh, S., Davis, K., and Greenbaum, I. (1993). Systematic and taxonomic implications of karyotypic, electrophoretic and mitochondrial DNA variation in *Peromyscus* from the Pacific Northwest. *J. Mamm.* 74, 819–831. doi: 10.2307/1382420

Hooper, E. T. (1947). Notes on Mexican mammals. *J. Mamm.* 28, 40–57.

Hooper, E. T. (1968). "Classification," in *Biology of Peromyscus (Rodentia)*, ed. J. A. King (USA: American Society of Mammalogist Special Publication), 27–74.

Hooper, E. T., and Musser, G. G. (1964). Notes on classification of the rodent genus *Peromyscus*. *Occas. Pap. Mus. Zool. Univ. Mich.* 635, 1–13.

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: bayesian inference of phylogeny. *Bioinformatics* 17, 754–755. doi: 10.1093/bioinformatics/17.8.754

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., et al. (2016). MEGAN Community Edition - Interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput. Biol.* 12:e1004957. doi: 10.1371/journal pcbi.1004957

Ihle, M. A., Fassunke, J., König, K., Grünewald, I., Schlaak, M., Kreuzberg, N., et al. (2014). Comparison of high resolution melting analysis, pyrosequencing, next generation sequencing and immunohistochemistry to conventional Sanger sequencing for the detection of p.*V600E* and non-p.V600E BRAF mutations. *BMC Cancer.* 14:13. doi: 10.1186/1471-2407-14-13

IUCN (2014). *Guidelines for using the IUCN Red List Categories and Criteria. Version 11*.

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 13, 1682–1684. doi: 10.1093/bioinformatics/btt193

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 40:e3. doi: 10.1093/nar/gkr771

Knapp, M., and Hofreiter, M. (2010). Next generation sequencing of ancient DNA: requirements, strategies and perspectives. *Genes* 1, 227–243. doi: 10.3390/genes1020227

Lahr, D. J., and Katz, L. A. (2009). Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques* 47, 857–866. doi: 10.2144/000113219

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. (2016). PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* 34, 772–773. doi: 10.1093/molbev/msw260

Lemmon, E. M., and Lemmon, A. R. (2013). High-Throughput genomic data in systematics and phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 44, 99–121. doi: 10.1146/annurev-ecolsys-110512-135822

Li, H. (2013). *Seqtk: A Fast and Lightweight Tool for Processing FASTA or FASTQ Sequences*. Available online at: https://github.com/lh3/seqtk/

Lim, G., Balke, M., and Meier, R. (2012). Determining species boundaries in a world full of rarity: singletons, species delimitation methods. *Syst. Biol.* 61, 165–169. doi: 10.1093/sysbio/syr030

Liu, L., Li, Y., Li, S., Hu, N., He, R., et al. (2012). Comparison of Next-Generation Sequencing Systems. *J. Biomed. Biotechnol.* 2012, 1–11. doi: 10.1155/2012/251364

Marcovitz, A., Jia, R., and Bejerano, G. (2016). "Reverse genomics" predicts function of human conserved noncoding elements. *Mol. Biol. Evol.* 33, 1358–1369. doi: 10.1093/molbev/msw001

Martinez-Cruz, B., Godoy, J. A., and Negro, J. J. (2007). Population fragmentation leads to spatial and temporal genetic structure in the endangered Spanish imperial eagle. *Mol. Ecol.* 16, 477–486. doi: 10.1111/j.1365-294x.2007.03147.x

Mason, V. C., Li, G., Helgen, K. M., and Murphy, W. J. (2011). Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. *Genome Res.* 21, 1695–1704. doi: 10.1101/gr.120196.111

McCormack, J. E., Rodríguez-Gómez, F., Tsai, W. L. E., and Faircloth, B. C. (2017). "Transforming museum specimens into genetic resources," in *The Extended Specimen: Emerging Frontiers in Collections-Based Ornithological Research*, ed. M. S. Webster (Boca Raton, FL: CRC Press), 143–156.

McCormack, J. E., Tsai, W. L., and Faircloth, B. C. (2016). Sequence capture of ultraconserved elements from bird museum specimens. *Mol. Ecol. Resour.* 16, 1189–1203. doi: 10.1111/1755-0998.12466

McDonough, M. M., Parker, L. D., Rotzel, N., Campana, M. G., and Maldonado, J. E. (2018). Performance of commonly requested destructive museum samples for mammalian genomic studies. *J. Mamm.* 99, 789–802. doi: 10.1093/jmammal/gyy080

Merriam, C. H. (1898). Description of twenty new species and a subgenus of *Peromyscus* from Mexico and Guatemala. *Proc. Biol. Soc. Wash.* 12, 115–125.

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 2010:pdb.prot5448. doi: 10.1101/pdb.prot5448

Meyerhans, A., Vartanian, J. P., and Wain-Hobson, S. (1990). DNA recombination during PCR. *Nucleic Acids Res.* 18, 1687–1691. doi: 10.1093/nar/18.7.1687

Miller, J. R., and Engstrom, M. D. (2008). The relationships of major lineages within peromyscine rodents: a molecular phylogenetic hypothesis and systematic reappraisal. *J Mamm.* 89, 1279–1295. doi: 10.1644/07-mamm-a-195.1

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). "Creating the CIPRES Science Gateway for inference of large phylogenetic trees," in *Proceedings of the Gateway Computing Environments Workshop (GCE)*, New Orleans, LA, 1–8.

Musser, G., and Carleton, M. D. (1993). "Family muridae," in *Mammal Species of the World: A Taxonomic and Geographic Reference*, eds D. E. Wilson and M. Reeder (Washington DC: Smithsonian Institution Press), 501–755.

Musser, G., and Carleton, M. D. (2005). "Superfamily muridae," in *Mammal Species of the World: A Taxonomic and Geographic Reference*, eds D. E. Wilson and M. Reeder (Baltimore, MD: Johns Hopkins University Press), 894–1531.

Osgood, W. (1909). Revision of the mice of the American genus *Peromyscus*. *North Am. Fauna* 28, 1–285. doi: 10.3996/nafa.28.0001

Pääbo, S. (1989). Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proc. Natl. Acad. Sci. U.S.A.* 86, 1939–1943. doi: 10.1073/pnas.86.6.1939

Pääbo, S., Irwin, D. M., and Wilson, A. C. (1990). DNA damage promotes jumping between templates during enzymatic amplification. *J. Biol. Chem.* 265, 4718–4721.

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rhohland, N., et al. (2004). Genetic analyses from ancient DNA. *Annu. Rev. Genet.* 38, 645–679.

Peery, M. Z., Hall, L. A., Sellas, A., Beissinger, S. R., Moritz, C., Bérubé, M., et al. (2010). Genetic analyses of historic and modern marbled murrelets suggest decoupling of migration and gene flow after habitat fragmentation. *Proc. R. Soc. B Biol. Sci.* 277, 697–706. doi: 10.1098/rspb.2009.1666

Phillippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., and Delsuc, F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC Evol. Biol.* 5:50. doi: 10.1186/1471-2148-5-50

Pichler, F. B., Dalebout, M. L., and Baker, C. S. (2001). Nondestructive DNA extraction from sperm whale teeth and scrimshaw. *Mol. Ecol. Notes.* 1, 106–109. doi: 10.1046/j.1471-8278.2001.00027.x

Platt, R. N. II., Amman, A. M., Keith, M. S., Thompson, C. W., and Bradley, R. D. (2015). What is *Peromyscus*? Evidence from nuclear and mitochondrial DNA sequences suggests the need for a new classification. *J Mamm.* 96, 708–719. doi: 10.1093/jmammal/gyv067

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Riddle, B., Hafner, D., and Alexander, L. (2000). Phylogeography and systematics of the *Peromyscus* eremicus species group and the historical biogeography of North American Warm Regional deserts. *Mol. Phylogenet. Evol.* 17, 145–160. doi: 10.1006/mpev.2000.0841

Rizzi, E., Lari, M., Gigli, E., De Bellis, G., and Caramelli, D. (2012). Ancient DNA studies: new perspectives on old samples. *Genet. Sel. Evol.* 44:21. doi: 10.1186/1297-9686-44-21

Rohland, N., and Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* 22, 939–946. doi: 10.1101/gr.128124.111

Ronquist, F., and Huelsenbeck, J. P. (2003). MRBAYES 3: bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574. doi: 10.1093/bioinformatics/btg180

Rowe, K. C., Singhal, S., Macmanes, M. D., Ayroles, J. F., Morelli, T., Rubidge, M., et al. (2011). Museum genomics: low-cost and high-accuracy genetic data from historical specimens. *Mol. Ecol. Resour.* 11, 1082–1010.

Roy, M. S., Girman, D. J., Taylor, A. C., and Wayne, R. K. (1994). The use of museum specimens to reconstruct the genetic-variability and relationships of extinct populations. *Experientia* 50, 551–557. doi: 10.1007/bf01921724

Sánchez-Cordero, V., Illoldi-Rangel, P., Linaje, M., Sarkar, S., and Peterson, T. A. (2005). Deforestation and extant distribution of Mexican endemic mammals. *Biol. Conserv.* 126, 465–473. doi: 10.1016/j.biocon.2005.06.022

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One* 7:e34131. doi: 10.1371/journal.pone.0034131

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi: 10.1093/bioinformatics/btr026

Secretaría de Medio Ambiente y Recursos Naturales [SEMARNAT] (2010). *Norma Oficial Mexicana NOM-059-SEMARNAT-2010. Protección Ambiental, Especies Nativas de Flora y Fauna Silvestres de México, Categorías de Riesgo y Especificaciones Para su Inclusión, Exclusión o Cambio, y Lista de Especies en Riesgo. Diario Oficial de la Federación, 30 de diciembre de 2010, Segunda Sección*. México: SEMARNAT.

Stamatakis, A. (2014). RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Sullivan, K. A. M., Platt, R. N. II., Bradley, R. D., and Ray, D. A. (2017). Whole mitochondrial genomes provide increased resolution and indicate paraphyly in deer mice. *BMC Zool.* 2:11. doi: 10.1186/s40850-017-0020-3

Tangliacollo, V. A., and Lanfear, R. (2018). Estimating improved partitioning schemes for ultraconserved elements. *Mol. Biol. Evol.* 35, 1798–1811. doi: 10.1093/molbev/msy069

Wandeler, P., Hoeck, P. E., and Keller, L. F. (2007). Back to the future: museum specimens in population genetics. *Trends Ecol. Evol.* 22, 634–642. doi: 10.1016/j.tree.2007.08.017

Webster, M. S. (2018). "The extended specimen," in *The Extended Specimen: Emerging Frontiers in Collections-Based Ornithological Research*, ed. M. S. Webster (Boca Raton, FL: CRC Press), 1–9.

Wiens, J. J. (2005). Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? *Syst. Biol.* 54, 731–742. doi: 10.1080/10635150500234583

Willerslev, E., and Cooper, A. (2005). Ancient DNA. *Proc. R. Soc. Lond. B Biol.* 272, 3–16.

Zimmermann, J., Hajibabaei, M., Blackburn, D. C., Hanken, J., Cantin, E., Posfai, J., et al. (2008). DNA damage in preserved specimens and tissue samples: a molecular assessment. *Front. Zool.* 5:18. doi: 10.1186/1742-9994-5-18

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

ORIGINAL RESEARCH
published: 06 May 2020
doi: 10.3389/fevo.2020.00105

# Assessing DNA Sequence Alignment Methods for Characterizing Ancient Genomes and Methylomes

*Marine Poullet*[1] *and Ludovic Orlando*[1,2]*

[1] *Laboratoire d'Anthropobiologie et d'Imagerie de Synthèse, CNRS UMR 5288, Faculté de Médecine de Purpan, Toulouse, France,* [2] *GLOBE Institute, University of Copenhagen, Copenhagen, Denmark*

Applying high-throughput DNA sequencing technologies to the ancient DNA molecules preserved in subfossil material can provide genetic information from past individuals, populations, and communities at the genomic scale. The combination of dedicated statistical techniques and specific molecular tools aimed at reducing the impact of post-mortem DNA damage can also help recover epigenetic data from ancient individuals. However, the capacity of different sequence aligners to identify ultrashort and deaminated ancient DNA templates and their impact on the characterization of ancient methylomes remain overlooked. In this study, we use both simulated and real ancient DNA sequence data to benchmark the performance of the read alignment tools most commonly used in ancient DNA research. We identify a read alignment strategy making use of the Bowtie2 aligner that substantially reduce computational times but shows increased sensitivity relative to previous recommendations based on the BWA aligner. This strategy significantly improves the genome coverage especially when DNA templates are shorter than 90 bp, as is typically the case for ancient DNA. It also impacts on ancient DNA methylation estimates as it maximizes coverage improvement within CpG dinucleotide contexts, which hold the vast majority of DNA methylation marks in mammals. Our work contributes to improve the accuracy of DNA methylation maps and to maximize the amount of recoverable genetic information from archeological and subfossil material. As the molecular complexity of ancient DNA libraries is generally limited, the mapping strategy recommended here is essential to limit both sequencing costs and sample destruction.

**Keywords: ancient DNA, DNA methylation, DNA damage, alignment, mapping, coverage, genome, methylome**

OPEN ACCESS

***Edited by:***
*Michael Knapp,*
*University of Otago, New Zealand*

***Reviewed by:***
*Kieren James Mitchell,*
*University of Adelaide, Australia*
*Katharina Dulias,*
*University of York, United Kingdom*
*Peter D. Heintzman,*
*UiT The Arctic University of Norway, Norway*

****Correspondence:***
*Ludovic Orlando*
*ludovic.orlando@univ-tlse3.fr*

***Specialty section:***
*This article was submitted to Paleoecology, a section of the journal Frontiers in Ecology and Evolution*

***Received:*** *28 November 2019*
***Accepted:*** *31 March 2020*
***Published:*** *06 May 2020*

***Citation:***
*Poullet M and Orlando L (2020) Assessing DNA Sequence Alignment Methods for Characterizing Ancient Genomes and Methylomes. Front. Ecol. Evol. 8:105. doi: 10.3389/fevo.2020.00105*

## INTRODUCTION

The first genome from an ancient human individual was sequenced in 2010 (Rasmussen et al., 2010) and was immediately followed by the genome sequencing of a Neanderthal (Green et al., 2010) and Denisovan (Reich et al., 2010) individual, two extinct archaic hominins. Since then, hundreds of ancient genomes have been characterized across many branches of the tree of life, including humans, horses, dogs, pigs, cattle, goats, wooly mammoths, but also many human pathogens and crops such as maize, sorghum, and barley (see Marciniak and Perry, 2017 and Brunson and Reich, 2019 for reviews). Ancient genome time series have made it possible to chart migration, admixture, and selection through space and time at unprecedented resolution. They have provided many opportunities to revisit evolutionary scenarios developed from patterns of cultural variation among archeological sites (e.g., the spread of steppe-related ancestry during the Eneolithic and early Bronze Age, see Allentoft et al., 2015; Haak et al., 2015; Damgaard et al., 2018a; Narasimhan et al., 2019; Wang et al., 2019) and from patterns of genetic variation in present-day populations (e.g., the temporal and geographic rise of lactose tolerance in western Eurasia, Mathieson et al., 2015; Ségurel and Bon, 2017).

The variation present in ancient DNA sequences does not only inform us about the genetic affinities of past individuals, populations, and species. It can also provide insights into ancient epigenetic landscapes, which play a crucial role in the regulation of gene expression (Lea et al., 2018) in response to infection (Smith et al., 2014; Pacis et al., 2015) as well as social (Laubach et al., 2019; Santos et al., 2019; Sanz et al., 2019; Snyder-Mackler et al., 2019) and environmental (Fagny et al., 2015) cues. It can, thus, help predict individual phenotypes in the past (see Pedersen et al., 2014 and Hanghøj et al., 2016 for age predictions on ancient individuals, or Gokhman et al., 2019 for morphological predictions).

Although methods have been developed to infer nucleosome maps in ancient tissues (Pedersen et al., 2014; Hanghøj et al., 2016), most of ancient epigenetic work thus far has focused on detecting DNA methylation within CpG dinucleotide (CpG) contexts. While molecular tools such as bisulfite sequencing (Llamas et al., 2012; Smith et al., 2015) or immunoprecipitation (Seguin-Orlando et al., 2015) have been used, genome-wide DNA methylation maps have been mostly produced through statistical inference leveraging the differential sequence footprint of post-mortem DNA damage at methylated and unmethylated sites, in particular at CpGs (see Hanghøj and Orlando, 2018 for a review). When molecular tools are used to prevent the sequencing of those unmethylated CpGs that have been degraded into UpGs, ancient methylated CpGs can indeed be revealed through CpG→TpG mis-incorporations in the sequence data (see "Materials and Methods"). Most recent methodologies have been proposed to mitigate the impact of evolutionary divergence and/or sequence variation at CpG sites on the calculation of DNA methylation scores (Hanghøj et al., 2019).

High-quality DNA sequence alignments against a reference genome are essential to make accurate predictions of past genetic and epigenetic variation. Yet, the vast majority of ancient DNA studies make use of read aligner software that were developed for mapping short read sequences produced from rather long DNA molecules extracted from fresh tissues. They are, thus, not optimized for the ultra-short and degraded nature of ancient DNA templates (Dabney et al., 2013). Several studies have contrasted a range of mapping conditions to identify those most specific and most sensitive (e.g., Schubert et al., 2012; Cahill et al., 2018) or to mitigate the extent of reference bias (Günther and Nettelblad, 2019; Martiniano et al., 2019). Yet, the sensitivity and specificity of other read aligners for ancient DNA data, such as Bowtie2 (Langmead and Salzberg, 2012), as well as the impact of different read alignment strategies on ancient methylation inference, remain untested. However, since the latter relies on patterns of CpG→TpG mis-incorporations introduced in the genome sequence data by post-mortem DNA damage, the read-to-reference edit distance is expected to be increased at methylated sites. This may affect the alignment sensitivity at such sites and, in turn, impact the accuracy of DNA methylation inference for ancient individuals. It is, thus, essential to investigate the possible sensitivity of read alignment methods at CpG dinucleotides so as to not underestimate DNA methylation levels along the genome and to accurately identify differentially methylated regions between individuals showing different levels of post-mortem DNA damage.

In this study, we assess the performance of 11 read alignment strategies for mapping ancient DNA sequence data against reference genomes and their impact on the inference of ancient DNA methylation. Our main purpose is not to carry out an exhaustive investigation about the impact of mapping parameters on an entire array of sequence data reflecting various post-mortem DNA decay conditions (for such studies, please refer to, e.g., Schubert et al., 2012; Cahill et al., 2018; Renaud et al., 2018). We instead focus on identifying those parameters and factors with potential impact on DNA methylation, using publicly available sequence data carefully selected from the literature to have been generated both in the absence and presence of USER treatment of the same ancient DNA extracts. The latter currently provide the best source of information to estimate CpG methylation level from patterns of CpG→TpG mis-incorporations (see Hanghøj and Orlando, 2018 for a review). Overall, we uncover that the end-to-end alignment mode of Bowtie2 shows better performance than all other commonly used alternatives. Simulated and real ancient human DNA sequence data reveal that the coverage can be increased by up to 2.1–9.4% for a given sequencing effort. The gain in recovered read alignments is particularly important within CpGs and significantly impacts the inference of regional DNA methylation levels. Applying such alignment procedures thus, improves both the quality of the genome and epigenetic data produced from ancient individuals and extinct species.

## MATERIALS AND METHODS

### Ancient DNA Sequence Datasets

Previously published raw sequence data from four ancient human individuals were downloaded from the European Nucleotide Archive (**Table 1**). For three of the four ancient humans (SI, SIII, and SIV, Sikora et al., 2017), Illumina DNA sequences were generated for libraries prepared with and without treatment with the USER enzymatic mix (Rohland et al., 2015). The USER treatment makes use of a first enzymatic activity, the uracil DNA glycosylase, to eliminate uracil residues (U) accumulated in ancient DNA templates due to post-mortem deamination of cytosine residues (C). This leaves abasic sites as targets for a second enzymatic reaction in which the Endonuclease VIII cleaves the DNA backbone 3′ of the abasic site. As a result, the fraction of DNA library templates containing U residues is reduced, which limits the number of C→T

**TABLE 1** | Sample and sequence information.

| Name | Experimental conditions | Age (years ago) | Location | Bone | Mean fragment length | Read pairs | Libraries | ENA accession number | Publication |
|---|---|---|---|---|---|---|---|---|---|
| Sunghir SI | USER+ | 33,875–31,770 | Russia | Molar root | 50.636 | 15,069,820 | SI_388_USER_14_CGTATA | PRJEB22592 | Sikora et al., 2017 |
| Sunghir SI | USER− | 33,875–31,770 | Russia | Molar root | 55.162 | 20,168,236 | SI_388_NOT_USER_13_CTATCA | PRJEB22592 | Sikora et al., 2017 |
| Sunghir SIII | USER+ | 35,154–33,031 | Russia | Molar root | 64.873 | 12,419,661 | SIII_386_USER_39_CGACCT | PRJEB22592 | Sikora et al., 2017 |
| Sunghir SIII | USER− | 35,154–33,031 | Russia | Molar root | 70.630 | 13,336,119 | SIII_386_NOT_USER_2_CGATGT | PRJEB22592 | Sikora et al., 2017 |
| Sunghir SIV | USER+ | 34,485–33,499 | Russia | Femur | 60.055 | 12,029,755 | SIV_392_USER_23_AGCATG | PRJEB22592 | Sikora et al., 2017 |
| Sunghir SIV | USER− | 34,485–33,499 | Russia | Femur | 64.448 | 10,100,614 | SIV_392_NOT_USER_10_TAGCTT | PRJEB22592 | Sikora et al., 2017 |
| NE1 | USER− | 5,070–5,310 | Hungary | Petrous bone | 69.762 | 63,774,886 | NE1_SRR1186790 | PRJNA240906 | Gamba et al., 2014 |

*The name, ages, and location of each ancient DNA specimen considered in this study are provided with respect to the original publication reporting the DNA sequence data. The numbers of read sequencing pairs considered in the analyses are provided and represent only a subset of the overall data available for download at the European Nucleotide Archive (ENA). The experimental conditions indicate whether raw ancient DNA extracts where treated (USER+) or not (USER−) with the USER enzymatic mix prior to DNA library construction. Ages are calibrated radiocarbon ages. The mean fragment length corresponds to BWA ds.*

nucleotide mis-incorporations introduced during sequencing (Briggs et al., 2010). USER treatment is, however, inefficient for those C residues that were methylated but deaminated post-mortem, as the uracil DNA glycosylase shows no activity on the resulting thymine (T) residues. Therefore, C→T nucleotide mis-incorporations are mostly restricted to methylated loci in the presence of USER treatment (Pedersen et al., 2014). In these conditions, the read-to-genome edit distance can be expected to be inflated at such sites, which may affect the performance of read alignment software. In the absence of USER treatment, this effect is, however, expected to impact all C residues that were deaminated post-mortem, be methylated or not. Contrasting sequence data generated from raw or USER-treated ancient DNA extracts provided, thus, an opportunity to assess the performance of read alignment software at methylated loci. The sequence data underlying the ancient genomes of Sunghir Upper Paleolithic individuals originate from similar preservation conditions and were generated both in the presence and in the absence of USER treatment. These data thus provided us with an opportunity to assess whether mapping conditions could affect regional methylation prediction.

## Data Simulation

Quantifying the sensitivity and predicted positive value of DNA alignment software requires the identification of the fraction of reads correctly mapped (true positives), those not correctly mapped (false positives), and those not mapped at all (false negatives). In order to assess those performance statistics, we simulated DNA sequence data using the human (hg19, The Genome Sequencing Consortium, 2001) reference genome and Gargammel (Renaud et al., 2017). This software returns DNA sequences of a selected size and can include sequencing errors typical of Illumina DNA sequencing instruments and, optionally, DNA mis-incorporations reflecting post-mortem DNA damage. A total of 3.3 million read pairs were simulated both in the presence and in the absence of ancient DNA damage for an entire size range of DNA templates overlapping typical ancient DNA size distributions. This included 100,000 read pairs for each size increment of one nucleotide within the 25–45 bp range, as well as 100,000 read pairs for each size increment of five nucleotides within the 45–90 bp range, and finally 100,000 read pairs for each size increment of 10 nucleotides within the 90–120 bp range. DNA damage was simulated using the DNA mis-incorporation of sample SIII produced by mapDamage2 (Jónsson et al., 2013). The alignment file that was used as input for mapDamage2 was generated by Paleomix (version 1.2.13.2, Schubert et al., 2014) using the same human reference genome as above and the default end-to-end alignment mode of Bowtie2, sensitive.

## Read Processing and Alignment

Both simulated and real ancient DNA sequence data were processed using Paleomix. This automated computational pipeline carries out a number of read processing steps, including adapter trimming, pair collapsing, mapping, quality/size filtering, duplicate removal, and local realignment. Mapping was performed using both BWA (Li and Durbin, 2009) and Bowtie2 (Langmead and Salzberg, 2012), which represent the two most commonly used read alignment software in ancient DNA research. BWA version 0.7.17 was used in this study, together with two main alignment modes (backtrack and mem). The backtrack algorithm was applied both using seed or disabling seeding with default parameters (-n 0.04), as recommended by Schubert et al. (2012) for ancient DNA data. Version 2.3.5.1 of the Bowtie2 read mapper was used, applying both the local and end-to-end alignment modes and the four sensitivity options provided (very fast, fast, sensitive, and very sensitive). Combined, this represented a total of 11 read alignment conditions. Read pairs were automatically collapsed as single reads when showing sufficient sequence overlap and the base quality was recalculated according to sequence match at those overlapping positions, following the default procedure implemented in the AdapterRemoval2 software (Schubert et al., 2016). Reads shorter than 25 bp post-trimming and/or collapsing were disregarded except for BWA mem where reads shorter than 30 bp were disregarded. Computational running times were recorded using the time bash command.

## Coverage and DNA Methylation Calculations

Binary Alignment Map (BAM) read alignment files and summary files obtained from Paleomix were processed for a number of analyses. First, average depth-of-coverage was calculated disregarding alignments showing quality scores strictly lower than 30. This corresponded to the estimated endogenous coverage provided in the Paleomix summary file. Second, average depth-of-coverage estimates were calculated at CpG, CpA, CpC, and CpT dinucleotides using the coverage option of Bedtools (Quinlan and Hall, 2010), conditioning on the bed coordinates of each dinucleotide type present in the human reference genome (-d option). The coordinates were obtained using Seqkit Version 0.3.1.1 (Shen et al., 2016). Third, we repeated the previous calculations after soft-clipped bases present in the read alignments were masked using the Jvarkit Biostar84452 tool (Lindenbaum, 2015). All previous analyses were carried out on both the read and simulated DNA sequence data. For simulated data, we also estimated the sensitivity and positive predicted value of each alignment condition. The alignment sensitivity was measured by dividing the number of true-positive alignments by their sum with the number of false-negative alignments [i.e., true positives/(true positives + false negatives] (Schubert et al., 2016). The alignment positive predictive value was estimated as the fraction of all simulated reads that were correctly mapped [i.e., true positives/(true positives + false positives)] (Schubert et al., 2016). Reads were considered as true positives if they showed a minimum of 80% of their length overlapping the known genomic coordinates used for simulation. Reads were considered as false negative when not mapping and false positives otherwise. These three categories were identified using python version 2.7.5 and the pysam library (Li et al., 2009). Additionally, DNA methylation analyses were carried out using the recently developed DamMet package (version 1.0.1, Hanghøj et al., 2019), in which the fraction of DNA methylation, f, can be estimated for a given genomic region including a pre-selected number of CpG dinucleotides. In this study, we selected 22,845 regions showing a total of 100 CpG dinucleotides in the human reference genome as the amount of sequence data considered was not sufficient to retrieve genuine estimates in regions of smaller sizes (data not shown). The corresponding genomic coordinates were provided to DamMet in the form of a BED coordinate file using the -B option. The f DNA methylation values were directly retrieved for each genomic window from the DamMet output. For consistency, coverage estimates within each window were calculated using the approach described above. All plots were generated using RStudio Version 1.1.463 (RStudio Team, 2016) and the ggplot2 library (Wickham, 2016).

# RESULTS

## Overall Alignment Performance

BWA (Li and Durbin, 2009) represents the most common software for aligning ancient DNA data against a reference genome. Previous work has established that disabling seeding in BWA increased mapping sensitivity for ancient DNA data, owing to the presence of inflated mis-incorporation rates at read ends (Schubert et al., 2012). Additional work investigated the specificity and sensitivity of Bowtie2 for ancient DNA data (Cahill et al., 2018). The performance of both aligners has, however, not been benchmarked on ancient DNA data with the specific aim to assess their possible impact on the inference of ancient DNA methylation. We, thus, compared their overall alignment performance on previously published ancient DNA data from four ancient humans, consisting of three Upper Paleolithic individuals excavated at Sunghir (SI, SIII, and SIV) and one Neolithic individual from Hungary (NE1) (**Table 1**). This represented a total of 11 mapping conditions, including 3 for BWA (with/out seeding, and mem) and 8 for Bowtie2 (very fast, fast, sensitive, and very sensitive options for both the local and end-to-end alignment modes). Alignment performance was calculated by normalizing the genome coverage obtained in one mapping condition relative to that obtained when disabling seeding in BWA, after quality filtering and duplicate removal (**Figure 1**).

We first confirmed previous work reporting reduced BWA performance when seeding, corresponding to a loss of 0.19–0.51% coverage across all four ancient DNA sequence datasets investigated in the absence of USER treatment (**Figure 1**, USER−). BWA mem was found to show increased performance in all three Sunghir individuals, in which a gain of 1.66–3.63% coverage was obtained. However, the performance was reduced (1.25%) for the NE1 individual. This indicates that the individual features of ancient DNA datasets, which reflect different post-mortem DNA preservation conditions, can have both a positive and a negative impact on the performance of the mem alignment procedure. The same was found for the four sensitivity options (very fast, fast, sensitive, and very sensitive) of the Bowtie2 local alignment mode, in which up to 2.63% coverage could be gained and up to 1.75% could be lost depending on the procedure considered. In these conditions, the very fast sensitivity option was the only one associated with a performance drop in all four

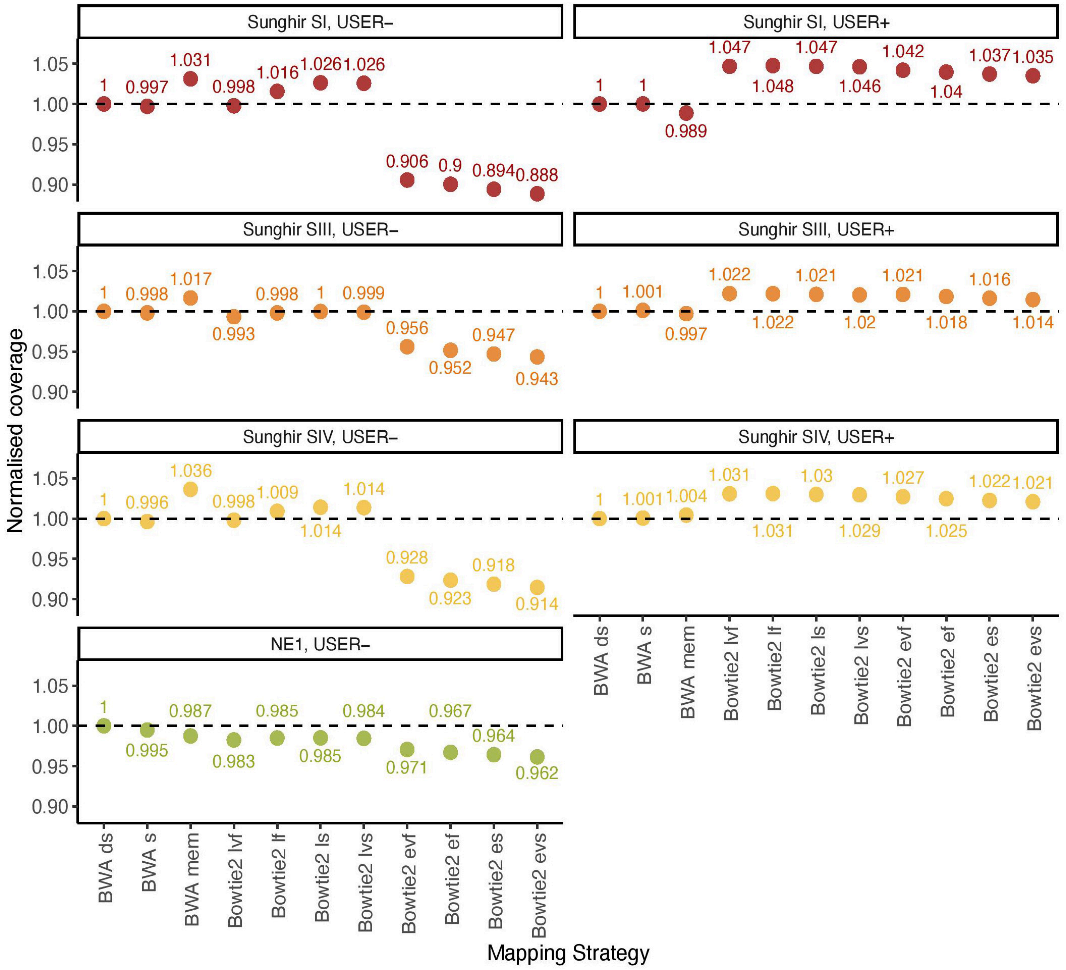


**FIGURE 1 |** Normalized coverage across all 11 mapping conditions investigated (real data). The average depth of coverage was estimated by filtering alignments for minimal mapping quality scores of 30 (MQ ≥ 30) and removing PCR duplicates. All coverage estimates are reported relative to those obtained with the BWA read aligner when disabling seeding (BWA ds). The other mapping conditions investigated in BWA correspond to seeding (BWA s) and mem (BWA mem). Two alignment modes (local, l; and end-to-end, e) were tested in Bowtie2, together with four options (fast, f; very fast, vf; sensitive, s; and, very sensitive, vs). The mean fragment length corresponding to BWA ds is indicated in **Table 1**.

samples tested (0.20–1.75%). In contrast to what was observed for the local alignment mode, the end-to-end mode in Bowtie2 was consistently found to show reduced performance, representing a loss of 2.93%–11.15% coverage.

We next assessed the mapping performance of ancient DNA data generated following USER treatment of raw DNA extracts (USER+). This treatment was developed to reduce the amount of DNA mis-incorporations resulting from post-mortem Cytosine deamination, which represents the most common DNA degradation reaction taking place after death (Briggs et al., 2010; Dabney et al., 2013). USER-treated ancient DNA data were available for the three Sunghir individuals but not for the NE1 individual. We found marginal coverage gain (0.01–0.11%) when disabling seeding in BWA, and different performance for the mem alignment procedure, in which a fraction of coverage could be gained (0.43%) or lost (1.12%) (**Figure 1**, USER+). Interestingly, all eight alignment modes tested for Bowtie2 were associated with increased performance, corresponding to a gain of 1.45–4.76% coverage relative to what was obtained when disabling seeding in BWA. This is in striking contrast with the reduced performance observed for the end-to-end alignment mode in the absence of USER treatment and indicates that the USER treatment modified the properties of ancient DNA data sufficiently enough to positively impact on the alignment performance.

To further gain insights into the alignment consequences of USER treatment, we simulated ancient DNA sequence data of increasing size (25–120 bp) and assessed the fraction of true positives, false positives, and false negatives obtained for each of the 11 alignment procedures tested (**Figure 2** and **Supplementary Figures S1, S2**). We found that the fraction of false-negative alignments was minimal when using the BWA aligner, except for the mem alignment mode and DNA templates of sizes inferior to 35 bp. The end-to-end alignment mode in Bowtie2 also led to virtually no false-negative alignments across all size categories investigated, including for DNA templates of 25–26 bp in which a detectable proportion of false-negative alignments was obtained when using the local alignment mode (albeit more limited than that observed in BWA mem for larger sizes, **Figure 2A**). Applying strict mapping quality thresholds of 30 was found appropriate for eliminating all false-positive alignments obtained in all mapping conditions investigated (**Figure 2B**). Interestingly, the fraction of true-positive alignments showing mapping quality scores strictly inferior to 30 increased in BWA rather than in Bowtie2 for DNA templates of limited sizes (25–70 bp) (**Figure 2C**). This fraction increased for larger sizes when using the Bowtie2 local alignment mode (70 bp) or the Bowtie2 end-to-end alignment mode (90 bp). This indicates that the mapping quality scores returned by BWA for short size categories such as those generally observed with ancient DNA data are more conservative than those returned by Bowtie2. Moreover, when applying a strict mapping quality threshold of 30 (as commonly practiced, e.g., Sikora et al., 2017), the sensitivity (i.e., the fraction of true positives relative to both true positives and false negatives) of BWA was more limited in the mem alignment mode than when seeding or disabling seeding for short DNA templates. It, however, returned to ∼100% for size categories superior or equal to 40 bp (**Supplementary Figure S3**). Maximal sensitivity was observed in Bowtie2 using the end-to-end alignment mode (Damgaard et al., 2018b), resulting in a significant loss of true-positive alignments in BWA compared to Bowtie2 (**Supplementary Figure S3**). This effect is reversed for size categories larger than 70 bp (local mode) or 90 bp (end-to-end mode), but this is expected to minimally impact ancient DNA datasets due to the generally extensive DNA fragmentation that takes place post-mortem (**Figure 2C**).

We next tested this prediction by measuring the overall alignment performance of the 11 procedures investigated by calculating the total coverage achieved after applying a strict mapping quality threshold of 30 and removing PCR duplicates (**Figure 3**). We confirmed that Bowtie2 showed an increased performance relative to BWA for DNA templates of size inferior to 70 bp when running with the local mode and for templates of size inferior to 90 when running the end-to-end mode.

Altogether, our USER-treated read simulations revealed that across all size categories. The sensitivity of the local alignment mode was reduced for DNA templates of size strictly inferior to 38 bp, but was generally larger than that observed with BWA mem. The quality scores returned by BWA in the short size range were found to be conservative, leading to the loss of a significant fraction of true positives (9.4–10.0%) when applying strict quality thresholds (**Figure 2C**).

## Alignment Performance at CpG Sites and DNA Methylation Inference

The most commonly used strategy available for estimating ancient methylation maps leverages patterns of C→T mis-incorporations at CpG dinucleotide sites as identified from BAM alignment files providing ancient DNA sequence alignment against a reference genome. We next investigated if the different mapping conditions investigated above showed different performance at CpG dinucleotide sites and could lead to different estimates of ancient DNA methylation levels. We first calculated the coverage achieved at each CpN dinucleotide context (i.e., CpA, CpC, CpG, and CpT) when applying the 11 mapping conditions to the sequencing data available for the three Sunghir individuals (**Figure 4** and **Supplementary Figure S4**). This revealed results largely consistent with those obtained when measuring coverage genome-wide, in which the Bowtie2 end-to-end mode showed the poorest performance when considering data generated in the absence of USER treatment (**Figure 4**, USER−). The performance drop was more pronounced at CpG dinucleotides. This is most likely due to the faster cytosine deamination rates reported at such sites when methylated (Seguin-Orlando et al., 2015; Smith et al., 2015), which increases the read-to-reference edit distance and, thus, limits the alignment sensitivity.

In striking contrast, Bowtie2 showed increased performance for all eight alignment conditions investigated when considering data generated following USER treatment (**Figure 4**, USER+). The performance gain was generally found to be especially pronounced within CpG dinucleotide contexts. This indicates that the USER treatment restored a fraction of reads that could not be previously aligned by reducing the read-to-reference edit distance. Since USER treatment is inefficient on those CpG dinucleotides that are methylated (Briggs et al., 2010; Pedersen et al., 2014; Hanghøj et al., 2016), we deduce that the Bowtie2 alignment conditions tested are more prone to result in gain of coverage at unmethylated CpG dinucleotides, which could have important consequences when deriving estimates of ancient DNA methylation levels.

The sensitive option of the end-to-end Bowtie2 alignment mode was found to show favorable running performance speed (**Supplementary Figure S5**). It also returned maximal sensitivity using simulated sequence data (**Supplementary Figure S3**) and maximal coverage gain at CpG dinucleotides on ancient DNA sequence data generated following USER treatment (**Figure 4**). We, thus, next compared the impact of this mapping condition on DNA methylation estimates relative to that used in all previous ancient DNA data work and consisting of disabling seeding in BWA (Gokhman et al., 2014, 2019; Pedersen et al., 2014; Hanghøj et al., 2016). To achieve this, we divided the human reference genome in windows comprising a total of 100 CpG dinucleotides and counted the number of such windows covered by at least one sequencing read in both alignment conditions. Although both alignment conditions identified read alignments in the vast majority of such genomic windows, we found that the total number of windows returning non-null coverage was larger for Bowtie2 than for BWA (**Figure 5A**), in line with the increased coverage observed with both simulated and real data with this mapper. This demonstrates that Bowtie2 retrieves data within regions for which no sequencing data could be aligned with BWA, thereby extending the genomic contexts into which DNA methylation can be estimated. Additionally, the distribution of sequencing depth obtained across genomic windows of 100 CpG dinucleotides was shifted toward larger values when applying

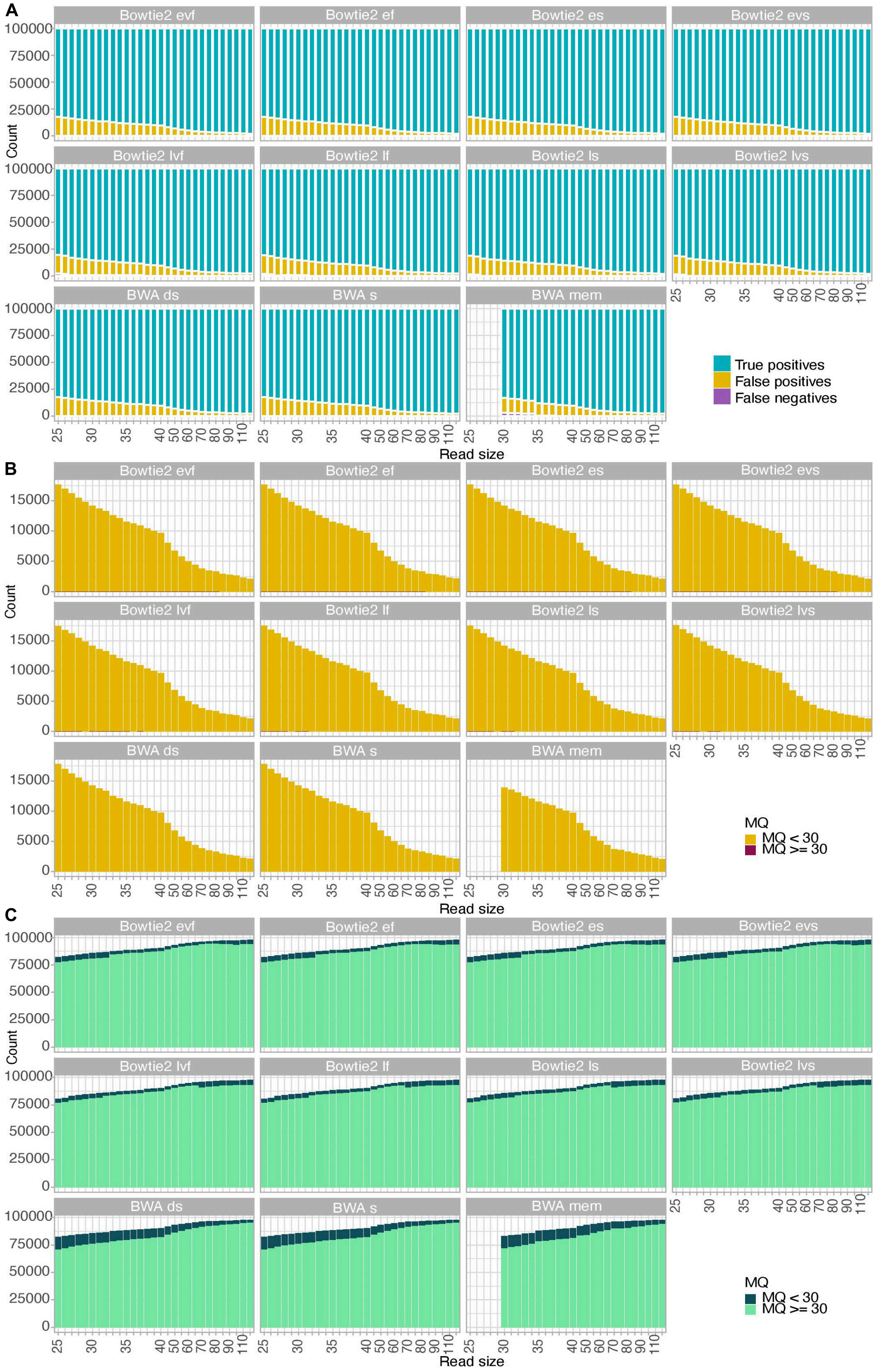


**FIGURE 2 |** Alignment performance of simulated data. A total of 100,000 reads were simulated for each size category considered in the presence of typical Illumina sequencing errors as well as nucleotide mis-incorporations remaining following USER treatment. MQ refers to the mapping quality scores of the read alignments. **(A)** Fractions of true-positive, false-positive, and false-negative alignments. **(B)** Mapping quality scores of false-positive alignments. **(C)** Mapping quality scores of true-positive alignments.

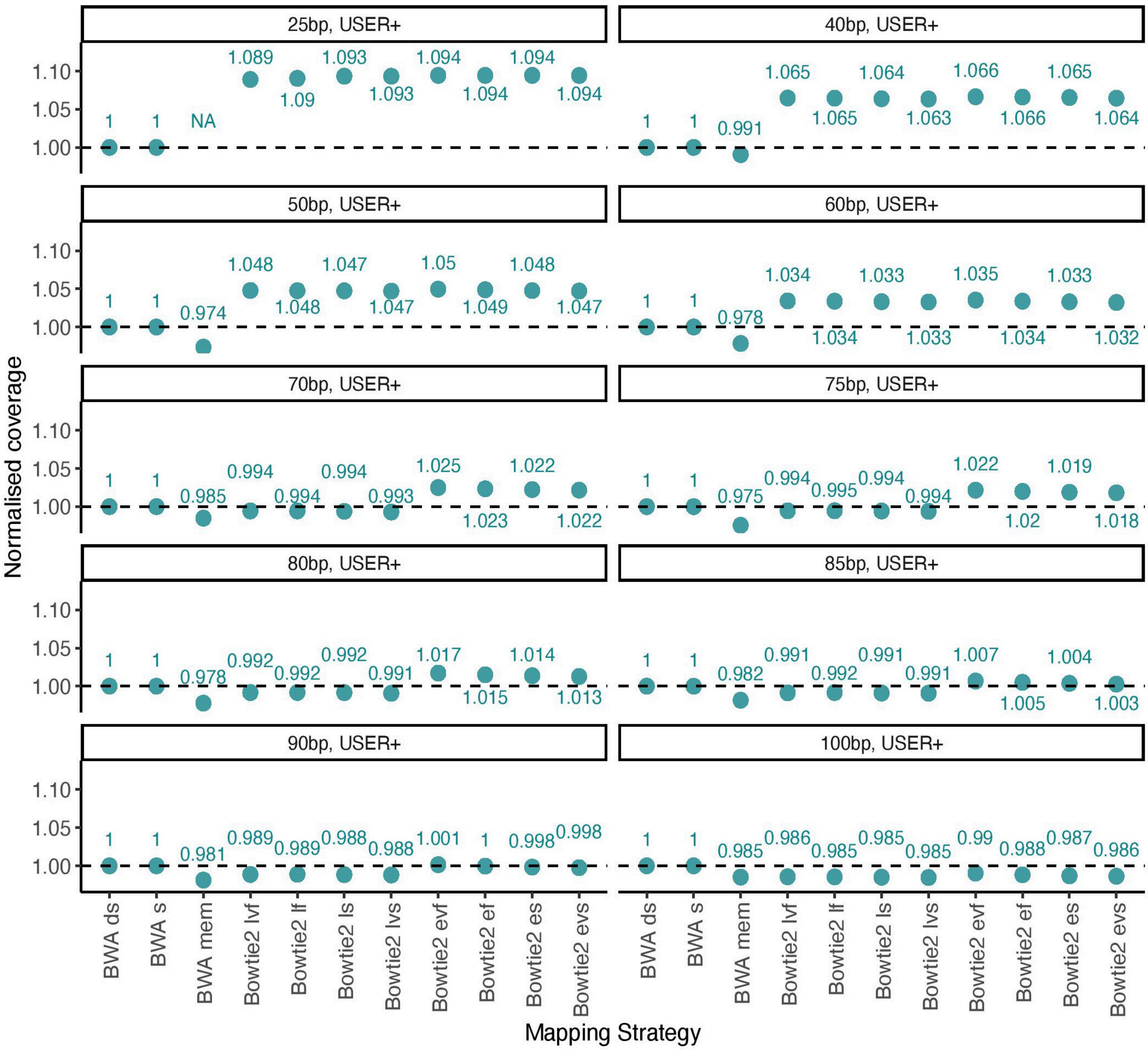


**FIGURE 3 |** Normalized coverage across all 11 mapping conditions investigated (simulated data). The average depth of coverage was estimated by filtering alignments for minimal mapping quality scores of 30 (MQ ≥ 30) and removing PCR duplicates. All coverage estimates are reported relative to those obtained with the BWA read aligner when disabling seeding (BWA ds). The other mapping conditions investigated in BWA correspond to seeding (BWA s) and mem (BWA mem). Two alignment modes (local, l and end-to-end, e) were tested in Bowtie2, together with four options (fast, f; very fast, vf; sensitive, s, and; very sensitive, vs). A total of 100,000 reads were simulated for each size category considered in the presence of typical Illumina sequencing errors as well as nucleotide mis-incorporations remaining following USER treatment. Each panel provides the coverage for each size category (25, 40, 50, 60, 70, 75, 80, 85, 90, and 100 bp).

Bowtie2 instead of BWA (**Figure 5B**, in which dashed lines indicate mean depth of coverage). This indicates that regional DNA methylation inference based on Bowtie2 alignments can build on more data than when based on BWA. This is important as the inference accuracy for ancient DNA methylation levels was previously shown to improve with sequencing depth (Hanghøj et al., 2019).

We next used DamMet (Hanghøj et al., 2019) to calculate in both alignment conditions the DNA methylation levels, f, for those genomic windows encompassing 100 CpG dinucleotides (**Figure 5C**). We found that the distributions of differences between the *f* values returned from Bowtie2 and BWA read alignments were centered around zero, indicating that the two mapping conditions resulted in similar regional methylation estimates. However, a fraction of the windows considered returned *f* values of one (i.e., full methylation) when using the sequence data aligned with BWA and values of zero (i.e., full demethylation) when using Bowtie2 alignments. This represented a fraction of 0.040–0.103% of the windows across the three ancient individuals investigated. Reciprocally, a fraction of the windows considered returned *f* values of zero when using the sequence data aligned with BWA and

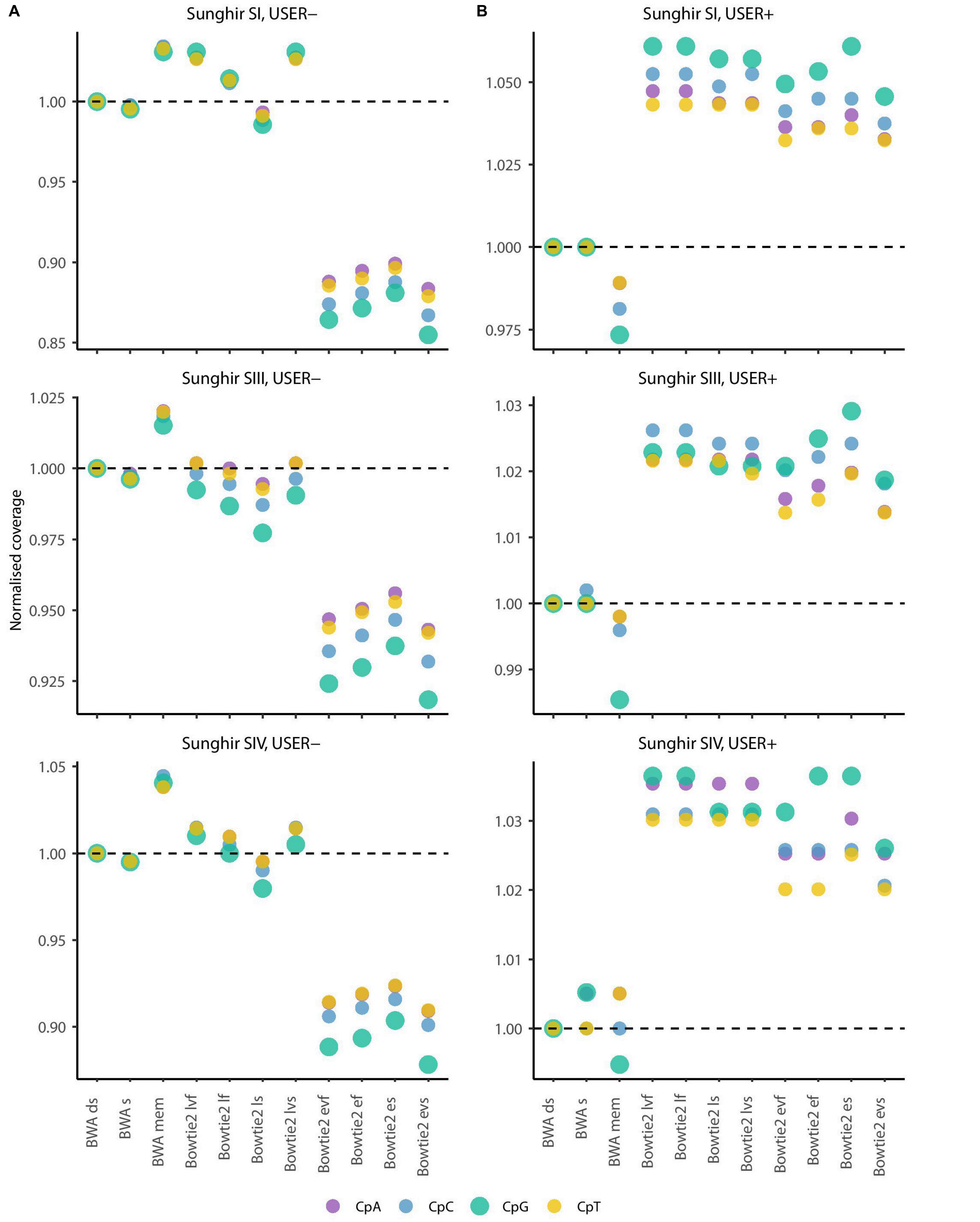


**FIGURE 4 |** Average depth of coverage in four dinucleotide contexts (real data). **(A)** Average depth of coverage when real data are generated in the absence of USER treatment. **(B)** Average depth of coverage when real data are generated following USER treatment. The average depth of coverage was estimated by filtering alignments for minimal mapping quality scores of 30 (MQ ≥ 30) and removing PCR duplicates. Coverage values are calculated in the dinucleotide sequence context most affected by DNA methylation (CpG), as well as the three other dinucleotides potentially affected by post-mortem cytosine deamination at the same position (i.e., CpA, CpC, and CpT). The differences observed are not due to soft-clipped bases as the values returned in the presence or not of soft-clipping masking are identical (**Supplementary Figure S4**).

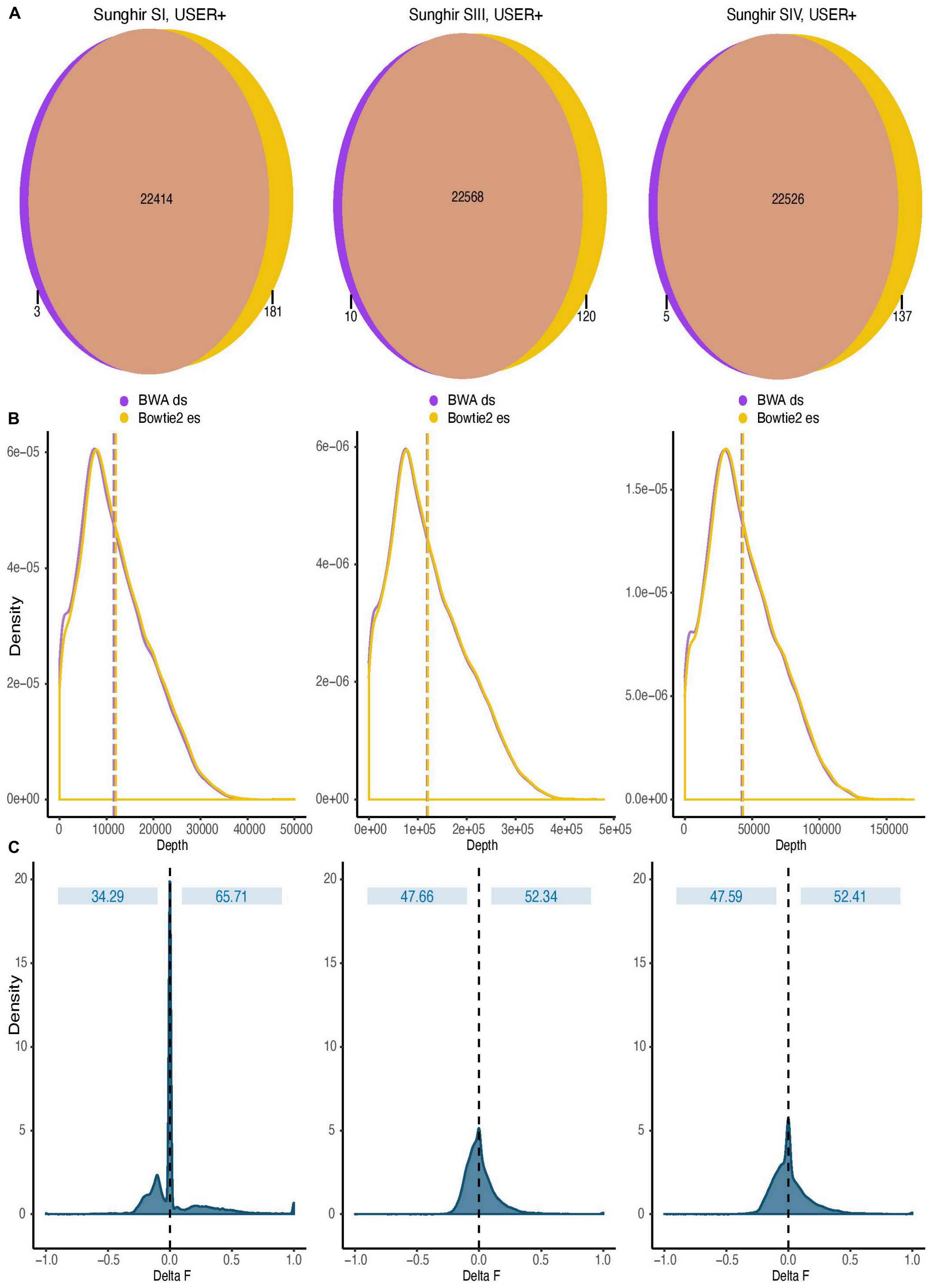


**FIGURE 5 |** Impact of read alignment conditions on ancient DNA methylation inference. The analyses were carried out using the sequence data generated following USER treatment of the raw DNA extracts of the three ancient Sunghir individuals (SI, left; SIII, center, and; SIV, right). The consequences of two mapping conditions on DNA methylation inference are investigated (BWA disabling seeding, BWA ds versus Bowtie2 end-to-end sensitive, Bowtie2 es). **(A)** Venn diagram of genomic windows showing non-null sequence coverage. Numbers indicate the total of genomic windows comprising 100 CpG dinucleotides and for which non-null sequence coverage was observed. Most windows are covered in both mapping conditions, but a fraction was exclusively identified by only one read aligner. Each circle is not scaled proportionally to increase readability. **(B)** Coverage distribution of genomic windows containing 100 CpG dinucleotides. Dashed lines represent the mean depth, respectively. **(C)** Distribution of the difference observed between the DNA methylation values inferred by two mapping conditions (Delta F). The difference (Delta F) reported corresponds to the difference between the *f* values estimated from Bowtie2 es alignments and those estimated from BWA ds alignments (f.Bowtie2 es - f.BWA ds).

values of one when using Bowtie2 alignments. This represented a larger fraction of the windows across the three ancient individuals investigated (0.316–1.392%), which demonstrates the increased sensitivity of the Bowtie2 aligner for those reads carrying CpG→TpG substitutions and informing on regional methylation levels. This demonstrates that the alignment procedures can significantly impact on the inference of regional DNA methylation levels.

## DISCUSSION

In this study, we report that Bowtie2 shows a higher performance than BWA when aligning ancient DNA data generated following USER treatment. This effect is especially pronounced within the shorter size range (25–70 bp), due to the combined effects of a higher sensitivity for the Bowtie2 read aligner, and more conservative mapping quality scores for the BWA aligner. Moreover, in the absence of USER treatment, BWA mem was found to impact positively the coverage estimates for some samples, but negatively for others. This may be related to the respective representation of ultrashort templates among the sequencing data, as the most positive impact is found for those libraries showing the shortest average sizes. Although this remains to be tested systematically, it suggests that individual post-mortem DNA preservation conditions will significantly affect the performance of this alignment procedure. Nonetheless, the vast majority of DNA fragments retrieved from archeological and paleontological remains are of limited sizes; the mapping conditions investigated with Bowtie2 can be expected to substantially improve genome coverage estimates and, hence, data quality. Using the sequence data obtained from three Upper Paleolithic Sunghir individuals, we found that Bowtie2 could improve the genome average depth of coverage by up to 1.62–3.72%. This improvement may appear modest at first glance but represents a significant improvement for ancient DNA research as the material available for destructive DNA extraction is finite and not replaceable. Additionally, improving sequencing depth is not always possible due to the limited molecular complexity of the DNA libraries available for sequencing. Importantly, read alignment conditions were found to impact not only depth of coverage but also the inference of regional ancient DNA methylation levels.

While ancient DNA methylation has received increasing scholar attention over the last 5 years, and while several statistical inference methods have been developed (e.g., BindDB, Livyatan et al., 2015; epiPaleomix, Hanghøj et al., 2016; and DamMet, Hanghøj et al., 2019), how much different read alignment methods could impact on methylation predictions had not been investigated. This study reveals that the Bowtie2 mapping conditions recommended (sensitive option, end-to-end mode) returns with larger numbers of read alignments that increase the number of genomic windows available for inference as well as their sequence coverage, which improves accuracy of the predictions. This has important consequences for the nascent field of ancient epigenomics, which, to the best of our knowledge, based all previous predictions on BWA DNA alignments. The fact that different read alignment conditions significantly impact the inference of ancient DNA methylation levels also implies that strictly identical alignment procedures are used when comparing DNA methylation levels in different ancient remains, including when groups showing different evolutionary distances to the reference genome used for alignments are considered (e.g., archaic hominins and anatomically modern humans, Gokhman et al., 2019).

Recent work has revealed that mapping ancient, ultrashort, and damaged ancient DNA reads against a linear reference genome can introduce substantial reference bias in the data, with possible impact on downstream population genetics inference (Günther and Nettelblad, 2019). Alignment procedures including a variation graph recapitulating the known genetic variation within a panel of modern individuals further mitigated this bias and helped effectively recover non-reference variants (Martiniano et al., 2019). Future work should focus on assessing the impact of such alignment procedures on ancient DNA methylation inference.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the ENA Accession number Sunghir: PRJEB22592, ENA Accession number NE1: PRJNA240906.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## AUTHOR CONTRIBUTIONS

LO conceived the study and provided material and infrastructure, and wrote the manuscript. MP carried out the analyses, with significant input from LO and plotted the figures.

## FUNDING

This work was supported by the Initiative d'Excellence Chaires d'attractivité, Université de Toulouse (OURASI), AMADEUS (CNRS LIA), and the Villum Fonden miGENEPI research project. LO has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program (grant agreement no. 681605-PEGASUS).

## ACKNOWLEDGMENTS

We are grateful to all members of the AGES research group (Archaeology, Genomics, Evolution and Societies) for discussions.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.00105/full#supplementary-material

## REFERENCES

Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., et al. (2015). Population genomics of Bronze Age Eurasia. *Nature* 522:167. doi: 10.1038/nature14507

Briggs, A. W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., and Pääbo, S. (2010). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* 38:e87. doi: 10.1093/nar/gkp1163

Brunson, K., and Reich, D. (2019). The promise of paleogenomics beyond our own species. *Trends Genet.* 35, 319–329. doi: 10.1016/j.tig.2019.02.006

Cahill, J. A., Heintzman, P. D., Harris, K., Teasdale, M. D., Kapp, J., Soares, A. E. R., et al. (2018). Genomic evidence of widespread admixture from polar bears into brown bears during the last ice age. *Mol. Biol. Evol.* 35, 1120–1129. doi: 10.1093/molbev/msy018

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., et al. (2013). Complete mitochondrial genome sequence of a middle pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15758–15763. doi: 10.1073/pnas.1314445110

Damgaard, P. D. B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., et al. (2018a). 137 ancient human genomes from across the Eurasian steppes. *Nature* 557, 369–374. doi: 10.1038/s41586-018-0488-1

Damgaard, P. D. B., Martiniano, R., Kamm, J., Moreno-Mayar, J. V., Kroonen, G., Peyrot, M., et al. (2018b). The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* 360:eaar7711. doi: 10.1126/science.aar7711

Fagny, M., Patin, E., MacIsaac, J. L., Rotival, M., Flutre, T., Jones, M. J., et al. (2015). The epigenomic landscape of African rainforest hunter-gatherers and farmers. *Nat. Commun.* 6:10047. doi: 10.1038/ncomms10047

Gamba, C., Jones, E. R., Teasdale, M. D., McLaughlin, R. L., Gonzalez-Fortes, G., Mattiangeli, V., et al. (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* 5:5257. doi: 10.1038/ncomms6257

Gokhman, D., Lavi, E., Prüfer, K., Fraga, M. F., Riancho, J. A., Kelso, J., et al. (2014). Reconstructing the DNA Methylation maps of the neandertal and the denisovan. *Science* 344, 523–527. doi: 10.1126/science.1250368

Gokhman, D., Mishol, N., de Manuel, M., de Juan, D., Shuqrun, J., Meshorer, E., et al. (2019). Reconstructing denisovan anatomy using DNA Methylation maps. *Cell* 179, 180–192. doi: 10.1016/j.cell.2020.01.020

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., et al. (2010). A draft sequence of the neandertal genome. *Science* 328, 710–722.

Günther, T., and Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet.* 15:e1008302. doi: 10.1371/journal.pgen.1008302

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522:207. doi: 10.1038/nature14317

Hanghøj, K., and Orlando, L. (2018). "Ancient epigenomics," in *Population Genomics*, ed. O. P. Rajora (Cham: Springer).

Hanghøj, K., Renaud, G., Albrechtsen, A., and Orlando, L. (2019). DamMet: ancient methylome mapping accounting for errors, true variants, and post-mortem DNA damage. *Gigascience* 8:giz02. doi: 10.1093/gigascience/giz025

Hanghøj, K., Seguin-Orlando, A., Schubert, M., Madsen, T., Pedersen, J. S., Willerslev, E., et al. (2016). Fast, accurate and automatic ancient nucleosome and methylation maps with epiPALEOMIX. *Mol. Biol. Evol.* 33, 3284–3298. doi: 10.1093/molbev/msw184

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., and Orlando, L. (2013). MapDamage2.0: fast approximate bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684. doi: 10.1093/bioinformatics/btt193

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Laubach, Z. M., Faulk, C. D., Dolinoy, D. C., Montrose, L., Jones, T. R., Ray, D., et al. (2019). Early life social and ecological determinants of global DNA methylation in wild spotted hyenas. *Mol. Ecol.* 28, 3799–3812. doi: 10.1111/mec.15174

Lea, A. J., Vockley, C. M., Johnston, R. A., Del Carpio, C. A., Barreiro, L. B., Reddy, T. E., et al. (2018). Genome-wide quantification of the effects of DNA methylation on human gene regulation. *eLife* 7:e37513. doi: 10.7554/eLife.37513

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lindenbaum, P. (2015). *JVarkit: Java-Based Utilities for Bioinformatics*. Available online at: https://github.com/lindenb/jvarkit (accessed July 22, 2019).

Livyatan, I., Aaronson, Y., Gokhman, D., Ashkenazi, R., and Meshorer, E. (2015). BindDB: an integrated database and webtool platform for "reverse-ChIP" epigenomic analysis. *Cell Stem Cell* 17, 647–648. doi: 10.1016/j.stem.2015.11.015

Llamas, B., Holland, M. L., Chen, K., Cropley, J. E., Cooper, A., and Suter, C. M. (2012). High-resolution analysis of cytosine Methylation in ancient DNA. *PLoS One* 7:e30226. doi: 10.1371/journal.pone.0030226

Marciniak, S., and Perry, G. H. (2017). Harnessing ancient genomes to study the history of human adaptation. *Nat. Rev. Genet.* 18:659. doi: 10.1038/nrg.2017.65

Martiniano, R., Garrison, E., Jones, E. R., Manica, A., and Durbin, R. (2019). Removing reference bias in ancient DNA data analysis by mapping to a sequence variation graph. *BioRxiv* [Preprint], doi: 10.1101/782755

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528:499. doi: 10.1038/nature16152

Narasimhan, V. M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., et al. (2019). The formation of human populations in South and Central Asia. *Science* 365:eaat7487. doi: 10.1126/science.aat7487

Pacis, A., Tailleux, L., Morin, A. M., Lambourne, J., MacIsaac, J. L., Yotova, V., et al. (2015). Bacterial infection remodels the DNA methylation landscape of human dendritic cells. *Genome Res.* 25, 1801–1811. doi: 10.1101/gr.192005.115

Pedersen, J. S., Valen, E., Velazquez, A. M. V., Parker, B. J., Rasmussen, M., Lindgreen, S., et al. (2014). Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Res.* 24, 454–466. doi: 10.1101/gr.163592.113

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., et al. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* 463, 757–762. doi: 10.1038/nature08835

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., et al. (2010). Genetic history of an archaic hominin group from denisova cave in Siberia. *Nature* 468, 1053–1060. doi: 10.1038/nature09710

Renaud, G., Hanghøj, K., Willerslev, E., and Orlando, L. (2017). Gargammel: a sequence simulator for ancient DNA. *Bioinformatics* 33, 577–579. doi: 10.1093/bioinformatics/btw670

Renaud, G., Petersen, B., Seguin-Orlando, A., Bertelsen, M. F., Waller, A., Newton, R., et al. (2018). Improved de novo genomic assembly for the domestic donkey. *Sci. Adv.* 4:eaaq0392. doi: 10.1126/sciadv.aaq0392

Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., and Reich, D. (2015). Partial uracil – DNA – glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20130624. doi: 10.1098/rstb.2013.0624

RStudio Team (2016). *RStudio: Integrated Development for R*. Boston, MA: RStudio, Inc.

Santos, H. P., Bhattacharya, A., Martin, E. M., Addo, K., Psioda, M., Smeester, L., et al. (2019). Epigenome-wide DNA methylation in placentas from preterm infants: association with maternal socioeconomic status. *Epigenetics* 14, 751–765. doi: 10.1080/15592294.2019.1614743

Sanz, J., Maurizio, P. L., Snyder-Mackler, N., Simons, N. D., Voyles, T., Kohn, J., et al. (2019). Social history and exposure to pathogen signals modulate social status effects on gene regulation in rhesus macaques. *Proc. Natl. Acad. Sci. U.S.A.* 201820846. doi: 10.1073/pnas.1820846116

Schubert, M., Ermini, L., Sarkissian, C. Der, Jónsson, H., Ginolhac, A., Schaefer, R., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* 9:1056. doi: 10.1038/nprot.2014.063

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., Al-Rasheid, K. A. S., Willerslev, E., et al. (2012). Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* 13:178. doi: 10.1186/1471-2164-13-178

Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* 9:88. doi: 10.1186/s13104-016-1900-2

Seguin-Orlando, A., Gamba, C., Sarkissian, C. Der, Ermini, L., Louvel, G., Boulygina, E., et al. (2015). Pros and cons of methylation-based enrichment methods for ancient DNA. *Sci. Rep.* 5:11826. doi: 10.1038/srep11826

Ségurel, L., and Bon, C. (2017). On the evolution of lactase persistence in humans. *Annu. Rev. Genomics Hum. Genet.* 18, 297–319.

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* 11:e0163962. doi: 10.1371/journal.pone.0163962

Sikora, M., Seguin-Orlando, A., Sousa, V. C., Albrechtsen, A., Korneliussen, T., Ko, A., et al. (2017). Ancient genomes show social and reproductive behavior of early upper paleolithic foragers. *Science* 358, 659–662. doi: 10.3389/fpsyg.2017.02247

Smith, O., Clapham, A. J., Rose, P., Liu, Y., Wang, J., and Allaby, R. G. (2014). Genomic methylation patterns in archaeological barley show de-methylation as a time-dependent diagenetic process. *Sci. Rep.* 4:5559. doi: 10.1038/srep05559

Smith, R. W. A., Monroe, C., and Bolnick, D. A. (2015). Detection of cytosine methylation in ancient DNA from five native american populations using bisulfite sequencing. *PLoS One* 10:e0125344. doi: 10.1371/journal.pone.0125344

Snyder-Mackler, N., Sanz, J., Kohn, J. N., Voyles, T., Pique-Regi, R., Wilson, M. E., et al. (2019). Social status alters chromatin accessibility and the gene regulatory response to glucocorticoid stimulation in rhesus macaques. *Proc. Natl. Acad. Sci. U.S.A.* 116, 1219–1228. doi: 10.1073/pnas.1811758115

The Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Wang, C.-C., Reinhold, S., Kalmykov, A., Wissgott, A., Brandt, G., Jeong, C., et al. (2019). Ancient human genome-wide data from a 3000-year interval in the caucasus corresponds with eco-geographic regions. *Nat. Commun.* 10:590. doi: 10.1038/s41467-018-08220-8

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, PH, declared a past collaboration with one of the author, LO, to the handling Editor.

frontiers in Ecology and Evolution

REVIEW
published: 16 June 2020
doi: 10.3389/fevo.2020.00040

# Palaeomicrobiology: Application of Ancient DNA Sequencing to Better Understand Bacterial Genome Evolution and Adaptation

*Luis A. Arriola[1]*, Alan Cooper[1,2] and Laura S. Weyrich[1,3,4]**

[1] Australian Centre for Ancient DNA, School of Biomedical Sciences, University of Adelaide, Adelaide, SA, Australia, [2] South Australian Museum, Adelaide, SA, Australia, [3] Centre for Australian Biodiversity and Heritage, School of Biological Sciences, University of Adelaide, Adelaide, SA, Australia, [4] Department of Anthropology and The Huck Institutes of Life Sciences, The Pennsylvania State University, University Park, PA, United States

OPEN ACCESS

**Edited by:**
*Nathan Wales,*
*University of York, United Kingdom*

**Reviewed by:**
*Alexander Herbig,*
*Max Planck Institute for the Science of Human History (MPI-SHH), Germany*
*Andrew Thomas Ozga,*
*Nova Southeastern University, United States*

***Correspondence:**
*Luis A. Arriola*
*luis.arriola@adelaide.edu.au*
*Laura S. Weyrich*
*lsw132@psu.edu*

**Specialty section:**
*This article was submitted to Paleoecology, a section of the journal Frontiers in Ecology and Evolution*

***Received:** 18 October 2019*
***Accepted:** 10 February 2020*
***Published:** 16 June 2020*

**Citation:**
*Arriola LA, Cooper A and Weyrich LS (2020) Palaeomicrobiology: Application of Ancient DNA Sequencing to Better Understand Bacterial Genome Evolution and Adaptation. Front. Ecol. Evol. 8:40. doi: 10.3389/fevo.2020.00040*

Next generation sequencing (NGS) has unlocked access to the wide range of non-cultivable microorganisms, including those present in the ancient past. The study of microorganisms from ancient sources (palaeomicrobiology) using DNA sequencing now provides a unique opportunity to examine ancient microbial genomic content, explore pathogenicity, and understand microbial evolution in greater detail than ever before. As a result, current studies have focused on reconstructing the evolutionary history of a number of human pathogens involved in ancient and historic pandemic events. These studies have opened the door for a variety of future palaeomicrobiology studies, which can focus on commensal microorganisms, species from non-human hosts, information from host-genomics, and the use of bacteria as proxies for additional information about past human health, behavior, migration, and culture. Here, we describe the origin and the historical and recent advances in the field of palaeomicrobiology, review some of the most notable ancient pathogenic microorganism studies, and provide perspectives on how NGS and whole genome information from ancient microorganisms contributes to our understanding of bacterial evolution on a broader scale. We conclude by exploring the application of newly developed tools in palaeomicrobiology and discussing how future studies can improve our current understanding of non-pathogenic microbes.

**Keywords: ancient DNA, pathogens, microbiome, microbiota, genomics, commensals, palaeomicrobiology, palaeomicrobiomics**

## INTRODUCTION

The advancement of DNA sequencing technologies has expanded significantly our understanding of the biology and evolution of microorganisms and drastically added to the body of literature using classical microbiological methodologies (culture-based assays, *in vitro* studies, microscopy, etc.). Today, more than 130,000 complete or near-complete bacterial genomes from >50 different phyla have been sequenced along with a variety of microorganisms, including archaea, fungi, and viruses (Genomes Online Database, GOLD v.7; Reddy et al., 2014; Land et al., 2015). Extensive genomic data from these microorganisms provided new insights into genome content, adaptation, and evolution, and opened up the opportunity to explore relationships between different related strains and across organisms from different kingdoms (Bryant et al., 2012; Bentley and Parkhill, 2015). However, we still know very little about microbial genome evolution over long time scales, i.e., over the course of anatomically modern human evolution, or the phenotypic, functional consequences of those long-term adaptations that do not result in disease.

Historically, microbial genomic sequencing was limited by the ability to culture pure strains; however, next-generation sequencing technologies have allowed researchers to rapidly access genomic information from uncultivable organisms (Riesenfeld et al., 2004; Schloss and Handelsman, 2005). Due to molecular decay and DNA fragmentation, microorganisms in ancient samples (i.e., hundreds to thousands of years old) are prime examples of taxa that cannot be cultivated. As a consequence, the use of ancient DNA (aDNA) sequencing in palaeomicrobiology – the study of ancient microorganisms – emerged with a focus on diagnosing and characterizing the pathogenic agents from past human pandemics using novel molecular techniques, such as hybridization enrichment and shotgun sequencing (Drancourt and Raoult, 2005). The use of genetics in paleomicrobiology now almost completely obscures historical methods, such as microscopy, to detect or characterize ancient microbes. The ability to interrogate a mixture of genetic material from all of the microorganisms present in a sample (the microbiome) using metagenomic approaches has also resulted in renewed interest and significant technical and analytical advances in the palaeomicrobiology field (Adler et al., 2013; Warinner et al., 2014; Harkins and Stone, 2015). By comparing ancient and modern microorganisms, researchers have now reconstructed the evolutionary history of several pathogens over extensive timescales and traced specific genomic changes that are linked to past diseases and epidemics (Anastasiou and Mitchell, 2013; Harkins and Stone, 2015; Bos et al., 2019).

In the following review, we examine how the field of palaeomicrobiology has been enhanced through the application of aDNA analysis and the use of next-generation sequencing technologies. We discuss the findings from several key ancient human bacterial pathogens, as well as several viral and eukaryotic pathogens, to explore how these techniques might be used on other microbes. We summarize how existing studies have revolutionized our understanding of microbial evolution and explore the existing pitfalls and barriers within this new research field. We conclude by describing how aDNA research can be improved in the future to address existing epidemiological and evolutionary questions, especially in non-pathogenic microbes.

## ORIGIN AND DEVELOPMENT OF ANCIENT MICROBIAL STUDIES USING ANCIENT DNA

Ancient DNA is the residual genetic record that can be found in historical and archaeological samples and has been successfully retrieved from archaeological sources, such as skin, teeth, soil, museum specimens, coprolites, calcified dental plaque, and bones (Tsangaras and Greenwood, 2012; Anastasiou and Mitchell, 2013; Burrell et al., 2015). By obtaining and sequencing aDNA, researchers can test evolutionary hypotheses with actual ancestral information from a wide variety of organisms (including plants, animals, and microorganisms) and provide critical insights into their evolutionary histories (Drancourt and Raoult, 2005; Roberts and Ingham, 2008; Burrell et al., 2015). There are a number of inherent complications when working with aDNA, including the damage and fragmentation of aDNA molecules, reduced endogenous DNA content from the organism of interest, and contamination of the samples by either modern or aDNA from other sources (Willerslev and Cooper, 2005; Tsangaras and Greenwood, 2012; Carpenter et al., 2013). Therefore, validation of the results from aDNA is critical and requires the strict implementation of certain basic standards, including the use of dedicated work areas, implementation of negative template controls, proper molecular behavior of DNA fragments, originality and consistency of the sequences, and reproducibility from different extracts of the sample in independent laboratories when using single gene analyses (Cooper and Poinar, 2000; Roberts and Ingham, 2008; Sarkissian et al., 2015). Additional validation may be required in certain circumstances, especially during palaeomicrobiological analysis; this includes sample decontamination, extraction blank control assessment, neighboring sample processing, and contaminant filtering (Weyrich et al., 2015). For example, aDNA studies that use next generation sequencing (NGS) can also assess the level of modern microbial DNA contamination within their extracts (Key et al., 2017; Weyrich et al., 2019). Ancient DNA research had a turbulent start due to these limitations, but the emergence of new DNA sequencing technologies and the application of more stringent standards have allowed researchers to directly examine how organisms adapt and evolve over time (Tsangaras and Greenwood, 2012; Hagelberg et al., 2015; Sarkissian et al., 2015). Nevertheless, the analysis of data generated with these new sequencing technologies also came with inherent issues and challenges, particularly in the study of ancient microbes.

Originally embedded in paleopathology from ancient human fossil remains, early palaeomicrobiology studies relied on morphological approaches and the use of basic biomolecular techniques (i.e., microscopy and immunodetection assays) to examine ancient samples (Swain, 1969; Tran et al., 2011). For example, Dobney and Brothwell (1986) performed one of the first microscopy-based explorations of ancient human-associated microbes on archaeological dental calculus in the late 1980s. Aided by scanning electron microscopy (SEM), they showed the presence of calcified rod shaped microorganisms along with other plant and animal tissues in preserved dental plaque (Dobney and Brothwell, 1986). Fornaciari and Marchetti (1986) and Fornaciari et al. (1989) also identified the presence of ancient pathogens in Italian mummies using immunochemical studies. Later, the invention of the polymerase chain reaction (PCR) drastically improved the ability of researchers to identify and examine ancient microbes via DNA amplification. This enabled researchers to minimally explore the genetics of ancient microbes by screening for specific DNA sequences without the need to culture. While the bulk of early PCR literature is expansive, key studies moved our understanding of ancient pathogens forward. For example, Spigelman and Lemma (1993) were able to use PCR to identify the presence of *Mycobacterium tuberculosis* DNA sequences in bone remains for the first time, allowing them to confirm the suspected diagnoses in ancient samples. This approach provided some power to identify specific microorganisms and trace their relationships with modern day taxa. However, PCR-based techniques are highly sensitive to contamination and can led to erroneous conclusions (Drancourt and Raoult, 2005; Roberts and Ingham, 2008). Several inaccurate publications arose from erroneous PCR results and brewed mistrust in the aDNA field (Fischman, 1995; Yousten and Rippere, 1997; Cooper and Poinar, 2000; Hazen and Roedder, 2001), highlighting the need to increase validation standards and develop more sensitive methods. At the end of the last decade, amplicon-based metagenomic approaches and NGS methods emerged as means to improve identifications and explore broader range of microbes, revolutionize palaeomicrobiology again. Whole genome shotgun metagenomics approaches performed by implementing new sequencing techniques at the time, such as Illumina 'sequencing by synthesis' technology, again provided improved efficiencies. The massively parallelized short read sequencing lengths provided by Illumina were a perfect match for the degraded nature of short aDNA fragments. This technology represented a revolution in sequence output and allowed researchers to complete detailed analyses of nearly whole genome sequences of ancient pathogens (e.g., *Yersinia pestis*, *Mycobacterium tuberculosis*, *M. leprae*, *Tanerella forsythia*, *Brucella melitensis*, *Helicobacter pylori*) (as reviewed in Tsangaras and Greenwood, 2012).

High-throughput DNA sequencing approaches greatly extended PCR methods and made the examination of a whole ancient microbial community possible. For example, NGS shotgun sequencing was used to explore microbial functional profiles and diversity of paleofeces (coprolites) (Tito et al., 2008). The ancient fecal samples showed resemblances with modern feces and gut microbiomes. However, there was an unexpected increase in diversity in ancient coprolites compared to modern samples, providing insight into the effects of geographic, temporal, and natural selection on ancient human microbiomes. Subsequently, whole-genome shotgun sequencing was applied to ancient permafrost samples, identifying a diverse and abundant array of ancient soil taxa and providing insight into the functional profiles of ancient soil microorganisms (Yergeau et al., 2010). Metagenomic approaches, such as DNA metabarcoding, have also been applied to explore the diversity of bacteria present in and on ancient skeletons (Adler et al., 2013; Warinner et al., 2014). Metabarcoding of calcified dental plaque (calculus; i.e., a preserved oral bacteria biofilm present in ancient teeth) revealed major alterations in the overall composition of oral microbial communities through time that were associated with historical, cultural, and dietary revolutions (Adler et al., 2013). Changes in the presence of two specific oral pathogens, *Streptococcus mutans* and *Porphyromonas gingivalis*, were also apparent. Despite the use of metabarcoding for targeted questions, this approach can produce microbial community compositions that are biased due to DNA fragment length and GC content (Ziesemer et al., 2015). Therefore, metagenomic shotgun sequencing was applied to reconstruct human microbiota and retrieve bacterial genomes (Warinner et al., 2014; Rasmussen et al., 2015; Weyrich et al., 2017) or reconstruct non-bacterial DNA in microbiome samples (Warinner et al., 2014). Despite the typically smaller size and number of archaeological samples, these studies have revealed the potential of metagenomic sequencing approaches to examine the evolution of microbial organisms, both individually and as a community. These studies have also provided insights into large alterations in human health, behavior, and culture through time, even if these changes were not linked to the presence or absence of a single pathogen, but rather collections of commensal microbial species.

Following the increased accessibility of NGS technology, hybridization capture or enrichment methods were designed and employed to examine ancient microbes. This approach involves creating RNA or DNA 'bait' sequences that are used to 'fish out' sequences of interest. These methods markedly extended the ability to obtain genomic information from a single species present within a complex 'soup' of bacterial and environmental DNA (Riesenfeld et al., 2004). Draft genome sequences have been reconstructed using hybridization capture from at least four key ancient bacterial pathogens, including *Vibrio cholerae*, *Yersinia pestis*, *Mycobacterium tuberculosis*, and *Mycobacterium leprae* (Bos et al., 2011, 2014; Schuenemann et al., 2013; Devault et al., 2014; Wagner et al., 2014). Whole genome comparisons of ancient microbial pathogens can be used to infer complex population dynamics, investigate environmental and host-pathogen interactions, track the origins and eruptions of past epidemics, and test hypotheses about the evolution of virulence through episodes of selection and extinction (Bryant et al., 2012; Vågene et al., 2018; Spyrou et al., 2019). In the following section, we summarize some of the most relevant research on these ancient pathogens, assess approaches that have been used for the study of their genomic evolution, and examine how future studies may continue to strengthen the field of palaeomicrobiology.

## PALAEOMICROBIOLOGY: FROM GENES TO GENOMES

### *Yersinia pestis*: The Infectious Agent of Ancient Plagues

Throughout history, at least three large "plague" pandemics were recorded: the First Pandemic, also known as the Plague of Justinian in the mid-6th century; the Second Plague pandemic, including the Black Death and the Great Plague of London over the 14th and 18th centuries; and the Third Pandemic, over the 19th and 20th centuries. Artistic and written documentation suggested a common etiological agent based on the descriptions of symptoms; however, direct evidence was only available for the most recent plague pandemic, where *Yersinia pestis* was identified as the causative agent of the disease by Alexandre Yersin (Zietz and Dunkelberg, 2004). The first retrospective diagnosis of skeletal remains from the Black Death (BD) used

PCR and primers specific for *Y. pestis* to amplify and sequence individual aDNA fragments from the teeth of people suspected to have died of plague (Drancourt et al., 1998; Drancourt and Raoult, 2005). Although this study seemed to confirm *Y. pestis* as the pathological agent of the BD, some studies challenged the results, as *Y. pestis* could not be amplified from plague victims following stringent aDNA protocols (Gilbert et al., 2004). To settle the debate, new attempts were made to obtain *Y. pestis* from plague specimens using NGS approaches (Bos et al., 2011; Schuenemann et al., 2011). These results were performed in two separate labs and followed strict validation guidelines to demonstrate aDNA authenticity. Using hybridization capture and enrichment for sequences similar to modern *Y. pestis* strains, both teams were able to retrieve and reconstruct the ancient genomic sequence of *Y. pestis* from plague victims, finally proving that this pathogen was one of the infectious agents of the BD.

Similarly, the etiology of the Justinianic Plague was debated for several years until an aDNA study was able to identify *Y. pestis* as the causative agent (Harbeck et al., 2013). Later, a low-coverage whole-genome sequence of a Justinianic strain from Southern Germany (Wagner et al., 2014) allowed the first comparison of several ancient genome sequences involved in different pandemic events. A phylogenetic analysis from these strains showed that the lineage responsible for the Justinianic outbreak was distinct from the lineages in the Second and Third Pandemics and was placed between two extant Chinese strains recovered from rodents. These results were recently confirmed with a high-coverage whole-genome sequence from a different burial (Feldman et al., 2016). Mutations were identified that could have influenced the infectivity of the Justinianic strain; however, the high-coverage genome identified a number of false positive SNPs that were reported in the previous low-coverage genome, highlighting the advantages of higher coverage sequences and the use of quality criteria to confirm the authenticity of reported substitutions. Keller et al. (2019) took a conservative approach to assay mutations in low-coverage genomes recovered from a wide-range of first pandemic genomes across Europe, examining the micro-diversity present in strains at the time. The origins of other pandemics have also now been identified; for example, the origin of the Black Death during the Second European Pandemic was traced back to Eastern Europe, although the molecular mechanisms of virulence were observed to be similar between pandemics (Spyrou et al., 2019). Further, Rasmussen et al. (2015) applied a unique approach, utilizing a metagenomic sequencing analysis method of deeply sequenced human genomic data to obtain *Y. pestis* genomes from teeth of Late Neolithic/Bronze Age humans (5,000–3,500 yBP) and explore the earlier origins of this disease. This study showed that ancient *Y. pestis* strains were already infecting ancient human populations at least 3,000 years earlier than the first recorded pandemic; however, these ancestral strains lacked many of the key virulence genes required to infect humans today and that were present in the strains from the three plague pandemics (Rasmussen et al., 2015). Further research was able to describe the virulence factors present in ancient strains, as well as specifically the bubonic form of the pathogen, and further supported the origins of plague in Europe to the Bronze Age (∼3,000 yBP; Spyrou et al., 2018), providing key insights into the evolutionary processes behind the high pathogenicity of this organism.

The studies on *Y. pestis* clearly demonstrated the effectiveness of hybridization capture/enrichment techniques in ancient microbial DNA studies. For the first time, researchers were able to trace bacterial evolution temporally and geographically in ancient civilizations, and we now have detailed maps and information about how past pandemics originated and moved across Europe. These studies represented the first successful approaches to sequencing ancient microbial genomes and revealed the potential of ancient genomic studies for understanding how bacteria co-evolved alongside humans. Furthermore, the latest studies highlight the power of metagenomic sequencing approaches and mining deeply sequenced human specimens as a novel source of ancient microbial DNA. Future studies can also likely examine the presence of *Y. pestis* outside of Europe; for example, exploring these strains in non-human mammals or insect vectors would also be an avenue to pursue to investigate host-adaptation and how its interactions with humans have shaped its evolutionary history.

## *Mycobacterium tuberculosis*: The Pathogenic Agent of TB

*Mycobacterium tuberculosis* is one of the most prevalent pathogens in human history (Müller et al., 2014) and remains the second largest cause of infectious disease deaths even today (Ottini and Falchetti, 2010; Dabernat et al., 2014). Along with other genetically similar *Mycobacterium* species, collectively known as the *Mycobacterium tuberculosis* complex (MTBC), it is the predominant etiological agent in the development of tuberculosis (TB). *M. tuberculosis* is the most widely studied pathogen in palaeomicrobiology (Zink et al., 2007), as it has been identified in hundreds of ancient individuals from various geographical regions and across several historical periods.

*Mycobacterium tuberculosis* was the first ancient bacterial pathogen to be studied using biomolecular methods. Initial palaeomicrobiological research on *M. tuberculosis* concentrated on the identification of the pathogen via PCR within ancient individuals who displayed osteological damage indicative of tuberculosis (Spigelman and Lemma, 1993; Tsangaras and Greenwood, 2012). Early PCR amplification studies focused on the identification of ribosomal proteins and insertion sequences specific to MTBC, and further characterization was performed using spoligotyping (spacer oligotyping (Taylor et al., 1999; Zink et al., 2002, 2007; Fletcher et al., 2003). However, these aDNA results were criticized for their unsuitability for phylogenetic analysis, poor reproducibility, and environmental contamination (Wilbur et al., 2009; Müller et al., 2014). More appropriate molecular approaches have since been applied, such as PCR methodologies to target single nucleotide polymorphisms (SNPs) and large sequence polymorphisms (LSPs) that provide power to infer phylogenetic relationships among ancient and modern samples (Bouwman et al., 2012; Müller et al., 2014).

PCR evidence of ancient *M. tuberculosis* was obtained from several prehistoric samples (before 3,000 BCE) (Salo et al., 1994; Nerlich et al., 1997; Crubézy et al., 1998; Hershkovitz et al., 2008), and suggested a long-term co-occurrence of the pathogen and humans. Most of these studies characterized a few loci within MTBC and provided evidence of the emergence of certain TB strains through time. However, they failed to provide information about particular genomic changes that differentiate specific strains or pathogenicity traits (Zink et al., 2007). Further, some loci utilized as markers of the MTBC, such as the insertional sequence IS6110, were widely used in ancient palaeomicrobiology studies (Tsangaras and Greenwood, 2012) and have since been shown to be conserved in both the human-infecting lineages of MTBC and soil isolates within the *Mycobacterium* family, making identification from buried ancient samples dubious (Wilbur et al., 2009; Müller et al., 2016). In light of this, hybridization enrichment techniques have been utilized to obtain whole genomic information to identify the origins of *M. tuberculosis* and to examine its recent history.

Ancient genomes from *Mycobacterium* species present a lower-than-average damage pattern among all the ancient bacterial genomes studied to date. This has been attributed to a robust cell wall and high G-C content which may help limit bacterial DNA degradation by protecting and stabilizing DNA molecules (Zink et al., 2002; Hershkovitz et al., 2008; Tsangaras and Greenwood, 2012). Recent studies have expanded the analysis of ancient *M. tuberculosis* to examine human migration and demographic changes by analyzing DNA sequences from a variety of geographical areas and across different times (Bouwman et al., 2012; Darling et al., 2014). While previous studies amplified a limited number of markers, new studies targeted multiple loci or whole genomes from multiple samples to obtain adequate resolution and disentangle complex relationships and migration histories. NGS analysis and enrichment capture methods improved phylogenetic resolution, allowing researchers to answer more complex questions, such as the origin of tuberculosis in ancient archaeological samples from the Americas (Bos et al., 2014). Whole genome sequences of modern and ancient *M. tuberculosis* genomes from South America and Europe revealed that South American human TB isolates were most closely related to TB strains from seals. This finding suggests that early South American TB strains may have resulted from a zoonotic transmission event, followed by a possible dispersal throughout humans in the Americas (Bos et al., 2014).

Research on ancient *M. tuberculosis* reflects the evolution of the field of palaeomicrobiology, from the single gene identification via PCR to whole genome reconstruction using NGS technologies. Ancient TB research has also provided insight into the dangers of assessing partial genomic sequences without proper environmental controls and highlights the need to consider mixtures of multiple related species within the same sample (e.g., environmental and human-associated *Mycobacterium*). Despite this, whole genome sequences have elucidated the evolutionary history, origins, and migrations of modern lineages. Further, whole genome sequences from ancient and modern *M. tuberculosis* strains, from different geographical sources and covering a large time-transect, will provide increased resolution of evolutionary events and adaptations that are elusive to our current models. Using ancient TB as a model system to examine specific evolutionary questions will undoubtedly improve our understanding of microbial evolution and distribution in human and animal populations throughout the world and across time.

## *Mycobacterium leprae*: The Evolution of an Obligate Intracellular Human Pathogen

*Mycobacterium leprae* is the infectious agent of leprosy, a chronic disease that has likely infected humans since prehistoric times (Donoghue, 2013; Mendum et al., 2014). Although nowadays it is most commonly found in undeveloped countries outside of Europe, archaeological and historical evidence, such as the establishment of several leper colonies, indicates leprosy was endemic in the medieval Europe, until it inexplicably vanished in the 16th century while increasing its prevalence in America and Africa (Schuenemann et al., 2013; Mendum et al., 2014; Andam et al., 2016; World Health Organization [WHO], 2016). *M. leprae* is an obligate intracellular pathogen with a slow growth rate (∼14 days generation time in humans), high genomic conservation between strains (i.e., <0.01% difference), a high number of pseudogenes (41% of its genome), and the loss of several housekeeping (conserved) genes (Cole et al., 2001; Monot et al., 2005; Schuenemann et al., 2013; Mendum et al., 2014). These features suggest that the *M. leprae* has undergone genome reduction or degradation similar to that reported during the evolution of other intracellular pathogens and symbionts (Cole et al., 2001; Ochman and Moran, 2001; Monot et al., 2005; Schuenemann et al., 2013). Initial palaeomicrobiological studies used PCR to characterize *M. leprae* from skeletal remains that exhibited osteological lesions indicative of leprosy (i.e., palate and nasal bone thinning) (e.g., Spigelman and Lemma, 1993); however, the majority of these studies did not comply with the criteria required for their authentication, rendering their results questionable.

Complete genomes of ancient *M. leprae* were recently recovered from medieval European samples (Schuenemann et al., 2013; Mendum et al., 2014). These studies obtained draft *M. leprae* genomes and found very few sites of differentiation in comparison to modern strains. In addition to following the basic guidelines for authentication of aDNA, ancient human mitochondrial DNA was also obtained by enrichment as an additional control for their authenticity. While the recovered mtDNA had distinctive damage patterns consistent with aDNA, the aDNA from *M. leprae* showed much lower levels of damage, raising further questions about decay kinetics of DNA in thick-outer-walled *Mycobacterium* species. Phylogenetic analyses of the ancient strains found them to cluster within a group containing modern Near East and Central Asia strains (Branch 2), as well as a group containing American strains and previously reported ancient European samples from a later period (Branch 3). This suggests that ancient strains from Branch 2 and 3 co-occurred in Europe, but eventually the Near Eastern and Asian strains were replaced by Branch 3, which later spread to the Americas (Schuenemann et al., 2013; Mendum et al., 2014). There also appeared to be little to no genetic evidence for virulence reduction in modern compared to ancient *M. leprae*. This is a critical finding and contradicts hypotheses that the declining incidence of leper cases in Europe through time was due to genomic changes in the pathogen. Instead, the reduction was likely caused by alternative factors, such as altered living conditions, changes in human immunity, novel protective factors, or alterations in co-infection dynamics (Schuenemann et al., 2013; Mendum et al., 2014).

Together, modern and ancient *M. leprae* DNA comparisons show a surprisingly low level of genomic variation since the divergence of the different lineages (∼3,000 years ago). Further work should explore and compare additional whole genomic sequences of various strains of all *Mycobacterium* species through time to explore the unique evolutionary history of this diverse genus (Harkins and Stone, 2015). Lastly, the low genetic variation and geographical associations revealed through the use of aDNA has highlighted the influence of human migrations in the propagation of this microbe in the past, suggesting trade paths such as the Silk Road as possible origins of transmission in the medieval era (Monot et al., 2009). This shows how palaeomicrobiology can do more than confirm past diagnoses, but rather contribute to the development of new hypotheses and other fields, such as history, epidemiology, and archaeology.

## *Vibrio cholerae*: Identification of Strains Involved in Ancient Pandemic Events Through Whole Genome Sequences

Cholera is an intestinal disease caused by infection with *Vibrio cholerae*. In the past 200 years, seven cholera pandemics (*V. cholerae* serotype O1) have been recorded worldwide, which were all attributed to two biotypes, distinguishable by several phenotypic markers: classical and El Tor (Harris et al., 2012). Cholera investigations based on disease etiology of early epidemics suggest that the O1 classical biotype may have been dominant before the last century. To test this hypothesis, Devault et al. (2014) reconstructed a near complete genome of an ancient *V. cholerae* sampled from a 19th-century preserved intestine of a patient deceased during the second cholera pandemic. This study identified the ancient strain as a “classical” biotype (serotype O1), confirming previous speculations that earlier pandemic events were caused by this strain (Devault et al., 2014). Whole genome sequence comparisons between modern and ancient strains also revealed high levels of genomic conservation, suggesting that the evolution of the modern pathogen has been under selective constraint during the past two centuries. However, it remains unclear how prevalent cholera was in more ancient periods.

Evidence suggests that the pathogenic strains involved in the last pandemics originated in Asia, so a more detailed exploration of ancient skeletons from that area could provide important insights into its origins. Sequencing aDNA from other sources, such as coprolites or mummified tissues, could also help identify *V. cholerae* in ancient specimens from populations with poor medical records, and also provide contextual information on how its infections may have impacted the gut microbiota of its inhabitants. Future studies on ancient *V. cholerae* could analyze dated human genomes or microbiomes that date to cholera pandemic events to investigate if ancient human populations were similarly affected by different strains of *V. cholerae*, and if genomic changes potentially supported the origination of this human pathogen. Moreover, additional dated ancient genomes could help to establish evolutionary rates of *V. cholerae* and add a temporal dimension to the studies.

## *Helicobacter pylori*: A Model Organism for Migration and Health

*Helicobacter pylori* is one of the most prevalent human-associated microorganisms and can play commensal or pathogenic roles. It is present in more than 50% of the global population, and several strains can be associated with development of chronic gastritis and gastric ulcers (Disotell, 2003; Secka et al., 2014). Modern day populations of *H. pylori* cluster into seven geographically associated prototypes, and the most common prototype in Europe (hpEurope) is known to carry a recombinant genomic sequence from Asian and African prototypes (Secka et al., 2014). Due to its long association with humans (∼100,000 years), wide global dispersal, and high sequence diversity, this microorganism has been used to trace past human migrations (Falush et al., 2003). Genomic sequences from ancient *H. pylori* from the “Iceman” (a 5,300-year-old European glacial mummy) were retrieved using hybridization capture enrichment (Maixner et al., 2016). Surprisingly, the genomic sequences of the ancient *H. pylori* resembled the Asian prototype, suggesting a migration of the African strain into Europe some time after 5,300 BP. The researchers also assessed the potential impact that these bacteria had in the ancient human gut and identified 22 proteins associated with inflammatory response similar to those expressed in modern humans infected by *H. pylori*.

Due to the long association with humans and a low mortality rate from *H. pylori* infection, studies of this microorganism allow us not only to obtain insights into its evolution and pathogenicity, but also the health and migrations of ancient human populations. In studies involving *H. pylori*, the availability of well-preserved samples from ancient guts is scarce; however, similar studies could be conducted on other commensal species not present in the gut. For example, an analysis of commensal microbes present in dental calculus could provide an opportunity to identify new marker species to track human migrations (Eisenhofer et al., 2019), and examine how truly commensal isolates evolve alongside humans, in contrast to pathogens with different selective pressures.

## *Salmonella enterica*: Interactions With Domesticated Animals

*Salmonella enterica* is commonly referred to as a food-borne illness that causes gastroenteritis (non-typhoidal salmonellosis) with over 93.8 million cases globally each year (Majowicz et al., 2010). However, certain *S. enterica* serovars, such as *S.* Typhi and *S.* Paratyphi, are human-specific pathogens that can cause enteric fevers (Typhoid and Paratyphoid fevers, respectively) and can have lethal consequences, both in the past and today. While over 21 million cases of Typhoid still occur annually today around the world (Crump et al., 2004), their origins remained poorly understood until recently (Achtman, 2016). Two aDNA studies have examined the origins of Typhoid in both the Old and New Worlds. Zhou et al. (2018) were able to compare modern *S.* enterica ser. Paratyphi C strains to a single ancient strain recovered from bones and teeth in Norway that date to 1,200 CE. Using updated genomic evolutionary rates and comparative analysis to modern strains in pigs, boars, and chickens, Zhou et al. (2018) conclude that *S.* Paratyphi C likely originated during porcine domestication about ∼4,000 years ago in Europe. In the New World, *S.* Typhi was linked to a cocoliztli outbreak from a 16th century site in modern Mexico, several years after the first Europeans arrived in the region in 1511 (Vågene et al., 2018).

While these and other studies seem to suggest an Old World origin for *S.* Typhi (Achtman et al., 2012), deeper origins cannot be ruled out (Vågene et al., 2018). The recent study by Zhou et al. (2018) concludes that origin of *S.* Paratyphi dates back 3,428 years ago, although early genetic dating estimates still suggest that *S.* Typhi may have arisen 10,000–43,000 years ago (Roumagnac et al., 2006). Further, little is known about the distribution of *S.* Typhi in historic and ancient times outside of Europe and where it may have arisen, especially considering its links to animal domestication. Studies examining both domesticated and non-domesticated animal microbiota may reveal commensal *Salmonella* strains in the past, providing insight into the evolution of this diverse genus. Further still, its presence in the New World and the potential effects, both social and biological, on New World Indigenous peoples needs further exploration; for example, understanding how resistance or survival of *S.* Paratyphi influenced downstream immune responses or microbiota may be critical to better understanding Indigenous health in the Americas (Skelly et al., 2018).

## Paleomicrobiological Examinations of Non-bacterial Species

Recently, paleomicrobiological tools have been applied to studies beyond bacteria to include viruses, parasites, and fungi, both from humans, animals, and the environment. Ancient viruses have emerged as accessible aDNA targets to better understand the origins of human disease in real-time. For example, the genome of the Variola virus, which causes small pox, was recovered both from 20th century strains and a child mummy from the middle of the 17th century (Duggan et al., 2016), suggesting that much of the variola viral evolution examined today has occurred in the past several hundred years (Duggan et al., 2016; Wertheim, 2017). Similar approaches were utilized to exclude a paleopathological diagnosed case of ancient smallpox (Ross et al., 2018). Additional ancient viral studies have explored the much deeper evolutionary history between viruses and humans, including hepatitis B (HBV) (Krause-Kyora et al., 2018; Mühlemann et al., 2018a; Ross et al., 2018) and parvovirus (Mühlemann et al., 2018b). Similar research projects are emerging in ancient parasites, namely in *Plasmodium* species to trace the origins of malaria. Marciniak et al. (2016) were able to verify the presence of *P. falciparum* DNA from a suspected Italian malarial outbreak in 1st–2nd century CE, while Gelabert et al. (2016) connected historic European *Plasmodium* strains to those circulating in the Americas, suggesting that Europeans may have introduced malaria into the Americas. These studies highlight the power of aDNA to track the origins and suspected histories of all pathogens, not only bacteria.

While many of these studies have reconstructed DNA from viral or eukaryotic pathogens in historic times, researchers can now examine the deeper evolutionary history and origins of these microbes. For example, tracking the long-term rates and directionality of viral evolution in Paleolithic humans and other hominids, such as Neanderthals, is now a possibility (Weyrich et al., 2017). Ancient DNA studies of viruses in other primates would also allow us to better calibrate and understand long-term viral evolution in hominids. Tracing back the environmental origins of ancient malarial strains, perhaps from preserved mosquitos or unique environmental samples, would also be important to better trace is distribution in the ancient past. Despite the exciting new questions that can be approaches using aDNA, it is also clear that strict aDNA methodologies must be performed during ancient viral and eukaryotic studies, as in bacterial studies, as the reliability of several ancient viral studies has already been questioned in the literature (Porter et al., 2017; Duchêne and Holmes, 2018).

## Overall Lessons From Whole Genome Palaeomicrobiology Studies

The current study of ancient microbial pathogens has been mainly driven by the progress and accessibility of DNA sequencing technologies and sample availability. The early use of aDNA methodologies provided the opportunity to identify ancient pathogens for the first time and add a temporal dimension to the evolutionary history of important human microbial pathogens. Early palaeomicrobiological studies also used single or a restricted number of loci and provided only limited descriptions of temporal and geographic distributions of pathogens within human populations. In some cases, early studies even mistakenly identified environmental or non-pathogenic species that shared sequence with their pathogenic counterparts (Müller et al., 2016). Further, these early studies did not provide insight into genome-wide changes that allow organisms to increase or reduce their pathogenicity and adapt to niches, nor did they allow researchers to determine where specific strains originated.

The invention of NGS technologies has had a major impact on the field of palaeomicrobiology, allowing researchers to amplify, target, and retrieve specific aDNA sequences from a variety of microorganisms and reconstruct partial or whole genomes. By comparing near-complete genomes of ancient and modern microbial pathogens, researchers have been able to describe differences between strains, interrogate the evolution and acquisition of virulence mechanisms, and the social and biological mechanisms that underpin pathogen origins. By identifying the genomic changes in pathogenic microorganisms through time, and comparing their effects on ancient and modern populations, we are beginning to track the origins of modern disease, trace the underlying causes of pandemics, and monitor microbial evolution in real time.

# THE FUTURE OF PALAEOMICROBIOLOGICAL STUDIES: NON-PATHOGENIC MICROBES

## Moving Beyond Pathogens to Non-infectious, Non-lethal, and Commensal Microbial Species

Because of their relevance to human health and the availability of the samples, the majority of palaeomicrobiological studies have concentrated on the diagnosis and identification of ancient pathogenic organisms involved in historical pandemics that typically cause acute, highly lethal diseases. However, symbiotic, non-pathogenic microorganisms also have a significant impact on human health by playing commensal or mutualistic roles. Many commensal microorganisms are evidence of a long-standing, co-evolutionary relationships; however, commensal bacteria can become opportunistic pathogens depending on the environmental conditions, albeit those conditions are poorly understood. Therefore, it is important to examine not only pathogens, but also the symbiotic microorganisms in the body and the micro-ecological context in which pathogens are successful. Infectivity and disease success can be highly influenced by competition (for similar nutritional resources or space) and cooperation between microorganisms, and this complex network of interactions has not yet been fully examined in palaeomicrobiological studies. For example, the dental caries causing pathogen, *Streptococcus mutans*, has competition from multiple other *Streptococcus* species that are routinely known to inhabit the oral cavity, but it remains unclear with other *Streptococcus* species were more dominant before the rise of *S. mutans* associated with the adoption of agriculture (Cornejo et al., 2013). For this purpose, future research should explore the connection and relationship between individual bacterial genomes of commensal species and the related microbiota to ascertain how genomic differences correlate with selective pressures or alterations in the microbial community. Correlations between the prevalence of certain microbial communities and a pathogen could provide information on disease susceptibility and a powerful new way to assess past population morbidity and health, especially in non-infectious diseases. Further still, non-lethal pathogens can provide information about the evolution of commensalism and serve as proxies to understand human interactions and movements. Examples of these endemic oral microbes include *Streptococcus*, *Neisseria*, and *Actinomyces* species (Avila et al., 2009). Future microbial community analysis may also reveal other species that would be better targets to answer these questions.

## Examination of Non-bacterial Species

Future palaeomicrobiology studies should include further research on non-bacterial microorganisms. Whole-genome sequences of ancient yeasts, viruses, fungi, and parasites are emerging, demonstrating how aDNA can provide, as in the case of bacterial pathogens, important insights into their evolution, both within humans and the environment. For example, the study of ancient yeast is an important issue, as they have played critical roles in disease development in humans, such as dental decay (Brighenti et al., 2014; Koo and Bowen, 2014), as well as in the fermentation and preservation of food products (Gibbons and Rinker, 2015). Future efforts could be applied to understand the impact of human selection and interaction on fermentation and preservation yeasts through time. An important advantage of yeasts is their ability to form spores, which are resistant to a variety of environmental stresses and may provide unique opportunities for preservation, although new extraction methodologies will likely need to be developed, as has been done for many other species (Shapiro et al., 2019). Furthermore, the interaction of human-associated yeasts and their products, such as fermentation, could have had an identifiable impact on human evolution and alterations of the human microbiome.

## Incorporation of Unique Sample Types

Two key human tissues have been utilized in many of the studies described in this review – bones and teeth. However, recent aDNA applications included microbiome containing specimens, such as dental calculus or coprolites, and have expanded our view of appropriate ancient specimens for palaeomicrobiological analysis (Tito et al., 2008; Adler et al., 2013). Further still, some recent samples have examined unique human-associated samples, such as calcified abscesses or kidney stones (Kay et al., 2014; Devault et al., 2017). Further still, well-preserved soft-tissue samples have also been shown to contain human microbiota, such as the "Iceman" mummy, although these are notably scarce. As our ability to accurately identify ancient microbes expands, novel sample types should be explored as sources of ancient bacteria, including pathogens, non-lethal pathogenic species, or commensal species. Unique sample types may include different human tissues, such as hair (Tobler et al., 2017), human associated archaeological materials, such as pots, ochre (Lenehan et al., 2017), or pitch (Jensen et al., 2019), or even human-associated microbes preserved in environmental specimens, such as ice (Turney et al., 2020), calcium carbonate (Frisia et al., 2017), or marine and terrestrial soils (Tobler et al., 2017). The use of new specimens will certainly further our understanding of the origins and evolution of many human-associated microbes, although attention to proper authentication criteria must be considered in all cases. Minimally, the integration of unique sample types will increase the available geographic and temporal range of ancient microbial genomes.

## Assessment of Non-human Associated Microbes

Most of the palaeomicrobiological studies to date have been conducted on human samples. However, domestic and wild animals are an important source of diseases, and many ancient pathogens may have been transferred between animals and humans, especially during the Neolithic Revolution when agricultural practices including animal husbandry were adapted (Wolfe et al., 2007). Where available, conserved remains from animals should be explored for pathogens. Plague is a good example of cases where animals played an important role in the transmission of disease to humans, although the roles that zoonotic vectors played in other ancient diseases remains unknown. Investigations into human pathogens with animal or insect reservoirs would also provide key information about pathogen-host co-evolution. As we expect rates of evolution to be variable across different microbes and/or strains in specific hosts, clear attention to the calculation of evolutionary rates and how these are influenced by mis-mapping or contamination will be key (Key et al., 2017; Bos et al., 2019). Studies exploring animal microbiomes could also provide information into the health and morbidity of ancient animals, help define their interactions with humans, and even provide details about their domestication, migration, and adaptation to past environments.

## CONCLUSION

The field of palaeomicrobiology has been revolutionized by NGS technologies, moving away from retrospective, single-gene diagnoses to whole genome analyses. Whole ancient genome sequences provide greater resolution to unequivocally identify the causes of ancient diseases and to obtain important insights into the evolution of pathogens, giving way to more complex hypotheses, models, and conclusions. Ancient DNA now offers the advantage to test evolutionary hypotheses across a wide-range of microbes with real, time-stamped data collected from the past. Deep co-evolutionary histories with non-pathogenic microbes need to be explored further, thus confirming, correcting or adding information uncovered in modern microbiome studies or inferred from models created with modern data. Most importantly, future palaeomicrobiological studies of past microorganisms will continue to provide invaluable insight into our own history, migrations, and evolution.

## AUTHOR CONTRIBUTIONS

LA, AC, and LW conceived of the manuscript. LA and LW wrote the manuscript. LA, AC, and LW edited the final manuscript.

## FUNDING

This research was funded by an Australian Research Council Discovery Early Career Fellowship Award (LW; DE150101574), Laureate Fellowship (AC; FL140100260), and an Australian Research Council Centre of Excellence award (AC and LW; CE170100015). This research is supported by an Australian Government Research Training Program (RTP) Scholarship (LA).

## ACKNOWLEDGMENTS

We would like to thank the members of the Metagenomics Research Cluster at the Australian Centre for Ancient DNA for insightful discussions on this manuscript.

## REFERENCES

Achtman, M. (2016). How old are bacterial pathogens? *Proc. Biol. Sci.* 283:1836.

Achtman, M., Wain, J., Weill, F.-X., Nair, S., Zhou, Z., and Sangal, V. (2012). Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog.* 8:e1002776. doi: 10.1371/journal.ppat.1002776

Adler, C. J., Dobney, K., Weyrich, L. S., Kaidonis, J., Walker, A. W., and Haak, W. (2013). Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the neolithic and industrial revolutions. *Nat. Genet.* 45, 450e–455e. doi: 10.1038/ng.2536

Anastasiou, E., and Mitchell, P. D. (2013). Palaeopathology and genes: investigating the genetics of infectious diseases in excavated human skeletal remains and mummies from past populations. *Gene* 528, 33–40. doi: 10.1016/j.gene.2013.06.017

Andam, C. P., Worby, C. J., Chang, Q., and Campana, M. G. (2016). Microbial genomics of ancient plagues and outbreaks. *Trends Microbiol.* 24, 978–990. doi: 10.1016/j.tim.2016.08.004

Armbrecht, L. H., Coolen, M. J. L., Lejzerowicz, F., George, S. C., Negandhi, K., Suzuki, Y., et al. (2019). Ancient DNA from marine sediments: precautions and considerations for seafloor coring, sample handling and data generation. *Earth Sci. Rev.* 196:102887.

Avila, M., Ojcius, D. M., and Yilmaz, Ö (2009). The oral microbiota: living with a permanent guest. *DNA Cell Biol.* 28, 405–411. doi: 10.1089/dna.2009.0874

Bentley, S. D., and Parkhill, J. (2015). Genomic perspectives on the evolution and spread of bacterial pathogens. *Proc. Biol. Sci.* 282:20150488. doi: 10.1098/rspb.2015.0488

Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., and Comas, I. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of new world human tuberculosis. *Nature* 514, 494–497. doi: 10.1038/nature13591

Bos, K. I., Kühnert, D., Herbig, A., Esquivel-Gomez, L. R., Valtueña, A. A., and Barquera, R. (2019). Paleomicrobiology: diagnosis and evolution of ancient pathogens. *Annu. Rev. Microbiol.* 73, 639–666. doi: 10.1146/annurev-micro-090817-062436

Bos, K. I., Schuenemann, V. J., Golding, G. B., Burbano, H. A., Waglechner, N., and Coombes, B. K. (2011). A draft genome of *Yersinia pestis* from victims of the black death. *Nature* 478, 506–510. doi: 10.1038/nature10549

Bouwman, A. S., Kennedy, S. L., Müller, R., Stephens, R. H., Holst, M., and Caffell, A. C. (2012). Genotype of a historic strain of mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. U.S.A.* 109, 18511–18516. doi: 10.1073/pnas.1209444109

Brighenti, F. L., Medeiros, A. C., Matos, B. M., Ribeiro, Z. E., and Koga-Ito, C. Y. (2014). Evaluation of caries-associated virulence of biofilms from Candida albicans isolated from saliva of pediatric patients with sickle-cell anemia. *J. Appl. Oral Sci.* 22, 484–489. doi: 10.1590/1678-775720130603

Bryant, J., Chewapreecha, C., and Bentley, S. D. (2012). Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. *Future Microbiol.* 7, 1283–1296. doi: 10.2217/fmb.12.108

Burrell, A. S., Disotell, T. R., and Bergey, C. M. (2015). The use of museum specimens with high-throughput DNA sequencers. *J. Hum. Evol.* 79, 35–44. doi: 10.1016/j.jhevol.2014.10.015

Carpenter, M. L., Buenrostro, J. D., Valdiosera, C., Schroeder, H., Allentoft, M. E., and Sikora, M. (2013). Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am. J. Hum. Genet.* 93, 852–864. doi: 10.1016/j.ajhg.2013.10.002

Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., and Wheeler, P. R. (2001). Massive gene decay in the leprosy bacillus. *Nature* 409, 1007–1011.

Cornejo, O. E., Lefébure, T., PavinskiBitar, P. D., Lang, P., Richards, V. P., Eilertson, K., et al. (2013). Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutans*. *Mol. Biol. Evol.* 30, 881–893.

Cooper, A., and Poinar, H. N. (2000). Ancient DNA: do it right or not at all. *Science* 289:1139.

Crubézy, E., Ludes, B., Poveda, J.-D., Clayton, J., Crouau-Roy, B., and Montagnon, D. (1998). Identification of mycobacterium DNA in an Egyptian Pott's disease of 5400 years old. *C. R. Acad. Sci. III* 321, 941–951. doi: 10.1016/s0764-4469(99)80009-2

Crump, J. A., Luby, S. P., and Mintz, E. D. (2004). The global burden of typhoid fever. *Bull. World Health Organ.* 82, 346–353.

Dabernat, H., Theves, C., Bouakaze, C., Nikolaeva, D., Keyser, C., and Mokrousov, I. (2014). Tuberculosis epidemiology and selection in an autochthonous Siberian population from the 16th-19th century. *PLoS One* 9:e89877. doi: 10.1371/journal.pone.0089877

Darling, M. I., Donoghue, H. D., Darling, M. I., and Donoghue, H. D. (2014). Insights from paleomicrobiology into the indigenous peoples of pre-colonial America – a review. *Mem. Inst. Oswaldo Cruz* 109, 131–139. doi: 10.1590/0074-0276140589

Devault, A. M., Golding, G. B., Waglechner, N., Enk, J. M., Kuch, M., and Tien, J. H. (2014). Second-pandemic strain of *Vibrio cholerae* from the philadelphia cholera outbreak of 1849. *N. Engl. J. Med.* 370, 334–340. doi: 10.1056/NEJMoa1308663

Devault, A. M., Mortimer, T. D., Kitchen, A., Kiesewetter, H., Enk, J. M., Golding, G. B., et al. (2017). A molecular portrait of maternal sepsis from Byzantine Troy (GH Perry, Ed). *eLife* 6:e20983.

Disotell, T. R. (2003). Discovering human history from stomach bacteria. *Genome Biol.* 4:213.

Dobney, K., and Brothwell, D. (1986). Dental calculus: its relevance to ancient diet and oral ecology. in: teeth and anthropology. *BAR Int. Ser.* 291, 55–81.

Donoghue, H. D. (2013). Insights into ancient leprosy and tuberculosis using metagenomics. *Trends Microbiol.* 21, 448–450. doi: 10.1016/j.tim.2013.07.007

Drancourt, M., Aboudharam, G., Signoli, M., Dutour, O., and Raoult, D. (1998). Detection of 400-year-old *Yersinia pestis* DNA in human dental pulp: an approach to the diagnosis of ancient septicemia. *PNAS* 95, 12637–12640. doi: 10.1073/pnas.95.21.12637

Drancourt, M., and Raoult, D. (2005). Palaeomicrobiology: current issues and perspectives. *Nat. Rev. Microbiol.* 3, 23–35. doi: 10.1038/nrmicro1063

Duchêne, S., and Holmes, E. C. (2018). Estimating evolutionary rates in giant viruses using ancient genomes. *Virus Evol.* 4:vey006. doi: 10.1093/ve/vey006

Duggan, A. T., Perdomo, M. F., Piombino-Mascali, D., Marciniak, S., Poinar, D., and Emery, M. V. (2016). 17th century variola virus reveals the recent history of smallpox. *Curr. Biol.* 26, 3407–3412. doi: 10.1016/j.cub.2016.10.061

Eisenhofer, R., Anderson, A., Dobney, K., Cooper, A., and Weyrich, L. (2019). Ancient microbial DNA in dental calculus: a new method for studying rapid human migration events. *J. Island Coast. Archaeol.* 14, 149–162.

Falush, D., Wirth, T., Linz, B., Pritchard, J. K., Stephens, M., and Kidd, M. (2003). Traces of human migrations in *Helicobacter pylori* populations. *Science* 299, 1582–1585. doi: 10.1126/science.1080857

Feldman, M., Harbeck, M., Keller, M., Spyrou, M. A., Rott, A., and Trautmann, B. (2016). A high-coverage *Yersinia pestis* genome from a 6th-century justinianic plague victim. *Mol. Biol. Evol.* 33, 2911–2923.

Fischman, J. (1995). Have 25-million-year-old bacteria returned to life? *Science* 268, 977–977. doi: 10.1126/science.7754393

Fletcher, H. A., Donoghue, H. D., Taylor, G. M., van der Zanden, G. M., and Spigelman, M. (2003). Molecular analysis of mycobacterium tuberculosis DNA from a family of 18th century Hungarians. *Microbiology* 149(Pt 1), 143–151. doi: 10.1099/mic.0.25961-0

Fornaciari, G., Castagna, M., Tognetti, A., Tornaboni, D., and Bruno, J. (1989). Syphilis in a renaissance Italian mummy. *Lancet* 2:614. doi: 10.1016/s0140-6736(89)90729-0

Fornaciari, G., and Marchetti, A. (1986). Intact smallpox virus particles in an Italian mummy of sixteenth century. *Lancet* 2:625. doi: 10.1016/s0140-6736(86)92443-8

Frisia, S., Weyrich, L. S., Hellstrom, J., Borsato, A., Golledge, N. R., and Anesio, A. M. (2017). The influence of antarctic subglacial volcanism on the global iron cycle during the last glacial maximum. *Nat. Commun.* 8:15425.

Gelabert, P., Sandoval-Velasco, M., Olalde, I., Fregel, R., Rieux, A., and Escosa, R. (2016). Mitochondrial DNA from the eradicated European *Plasmodium vivax* and *P. falciparum* from 70-year-old slides from the Ebro Delta in Spain. *PNAS* 113, 11495–11500. doi: 10.1073/pnas.1611017113

Gibbons, J. G., and Rinker, D. C. (2015). The genomics of microbial domestication in the fermented food environment. *Curr. Opin. Genet. Dev.* 35, 1–8. doi: 10.1016/j.gde.2015.07.003

Gilbert, M. T. P., Cuccui, J., White, W., Lynnerup, N., Titball, R. W., and Cooper, A. (2004). Absence of *Yersinia pestis*-specific DNA in human teeth from five European excavations of putative plague victims. *Microbiology* 150, 341–354. doi: 10.1099/mic.0.26594-0

Hagelberg, E., Hofreiter, M., and Keyser, C. (2015). Ancient DNA: the first three decades. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370:20130371. doi: 10.1098/rstb.2013.0371

Harbeck, M., Seifert, L., Hänsch, S., Wagner, D. M., Birdsell, D., and Parise, K. L. (2013). *Yersinia pestis* DNA from skeletal remains from the 6th Century AD reveals insights into justinianic plague. *PLoS Pathog.* 9:e1003349. doi: 10.1371/journal.ppat.1003349

Harkins, K. M., and Stone, A. C. (2015). Ancient pathogen genomics: insights into timing and adaptation. *J. Hum. Evol.* 79, 137–149. doi: 10.1016/j.jhevol.2014.11.002

Harris, J. B., LaRocque, R. C., Qadri, F., Ryan, E. T., and Calderwood, S. B. (2012). Cholera. *Lancet* 379, 2466–2476. doi: 10.1016/S0140-6736(12)60436-X

Hazen, R. M., and Roedder, E. (2001). Biogeology: how old are bacteria from the permian age? *Nature* 411, 155–155. doi: 10.1038/35075663

Hershkovitz, I., Donoghue, H. D., Minnikin, D. E., Besra, G. S., Lee, O. Y., and Gernaey, A. M. (2008). Detection and molecular characterization of 9000-year-old mycobacterium tuberculosis from a neolithic settlement in the eastern mediterranean. *PLoS One* 3:e3426. doi: 10.1371/journal.pone.0003426

Jensen, T. Z. T., Niemann, J., Iversen, K. H., Fotakis, A. K., Gopalakrishnan, S., Vågene, et al. (2019). A 5700 year-old human genome and oral microbiome from chewed birch pitch. *Nat. Commun.* 10:5520. doi: 10.1038/s41467-019-13549-9

Kay, G. L., Sergeant, M. J., Giuffra, V., Bandiera, P., Milanese, M., Bramanti, B., et al. (2014). Recovery of a medieval brucella melitensis genome using shotgun metagenomics (PS Keim, Ed). *mBio* 5:e1337-14.

Keller, M., Spyrou, M. A., Scheib, C. L., Neumann, G. U., Kröpelin, A., and Haas-Gebhard, B. (2019). Ancient *Yersinia pestis* genomes from across Western Europe reveal early diversification during the first pandemic (541–750). *PNAS* 116, 12363–12372. doi: 10.1073/pnas.1820447116

Key, F. M., Posth, C., Krause, J., Herbig, A., and Bos, K. I. (2017). Mining metagenomic data sets for ancient DNA: recommended protocols for authentication. *Trends Genet.* 33, 508–520. doi: 10.1016/j.tig.2017.05.005

Koo, H., and Bowen, W. H. (2014). Candida albicans and *Streptococcus mutans*: a potential synergistic alliance to cause virulent tooth decay in children. *Future Microbiol.* 9, 1295–1297. doi: 10.2217/fmb.14.92

Krause-Kyora, B., Susat, J., Key, F. M., Kühnert, D., Bosse, E., and Immel, A. (2018). Neolithic and medieval virus genomes reveal complex evolution of hepatitis B. *eLife* 7:e36666. doi: 10.7554/eLife.36666

Lenehan, C. E., Tobe, S. S., Smith, R. J., and Popelka-Filcoff, R. S. (2017). Microbial composition analyses by 16S rRNA sequencing: a proof of concept approach to provenance determination of archaeological ochre. *PLoS One* 12:e0185252. doi: 10.1371/journal.pone.0185252

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., and Ahn, T. H. (2015). Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics* 15, 141–161. doi: 10.1007/s10142-015-0433-4

Maixner, F., Krause-Kyora, B., Turaev, D., Herbig, A., Hoopmann, M. R., Hallows, J. L., et al. (2016). The 5300-year-old *Helicobacter pylori* genome of the Iceman. *Science* 351, 162–165. doi: 10.1126/science.aad2545

Marciniak, S., Prowse, T. L., Herring, D. A., Klunk, J., Kuch, M., Duggan, A. T., et al. (2016). Plasmodium falciparum malaria in 1$^{st}$-2$^{nd}$ century CE southern Italy. Curr. Biol. 26, R1220-R1222. doi: 10.1016/j.cub.2016.10.016

Majowicz, S. E., Musto, J., Scallan, E., Angulo, F. J., Kirk, M., O'Brien, S. J., et al. (2010). The global burden of nontyphoidal *Salmonella* gastroenteritis. *Clin. Infect. Dis.* 50, 882–889.

Mendum, T. A., Schuenemann, V. J., Roffey, S., Taylor, G. M., Wu, H., Singh, P., et al. (2014). Mycobacterium leprae genomes from a British medieval leprosy hospital: towards understanding an ancient epidemic. *BMC Genomics* 15:270. doi: 10.1186/1471-2164-15-270

Monot, M., Honoré, N., Garnier, T., Araoz, R., Coppée, J. Y., Lacroix, C., et al. (2005). On the origin of leprosy. *Science* 308, 1040–1042. doi: 10.1126/science/1109759

Monot, M., Honoré, N., Garnier, T., Zidane, N., Sherafi, D., Paniz-Mondolfi, A., et al. (2009). Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. *Nat. Genet.* 41, 1282–1289. doi: 10.1038/ng.477

Mühlemann, B., Jones, T. C., Damgaard, P., de, B., Allentoft, M. E., Shevnina, I., et al. (2018a). Ancient hepatitis B viruses from the Bronze age to the medieval period. *Nature* 557, 418–423. doi: 10.1038/s41586-018-0097-z

Mühlemann, B., Margaryan, A., Damgaard, P., de, B., Morten, A. E., Lasse, V., et al. (2018b). Ancient human parvovirus B19 in Eurasia reveals its long-term association with humans. *PNAS* 115, 7557–7562. doi: 10.1073/pnas.1804921115

Müller, R., Roberts, C. A., and Brown, T. A. (2014). Genotyping of ancient mycobacterium tuberculosis strains reveals historic genetic diversity. *Proc. R. Soc. B* 281:20133236. doi: 10.1098/rspb.2013.3236

Müller, R., Roberts, C. A., and Brown, T. A. (2016). Complications in the study of ancient tuberculosis: presence of environmental bacteria in human archaeological remains. *J. Archaeol. Sci.* 68, 5–11. doi: 10.1016/j.jas.2016.03.002

Nerlich, A. G., Haas, C. J., Zink, A., Szeimies, U., and Hagedorn, H. G. (1997). Molecular evidence for tuberculosis in an ancient Egyptian mummy. *Lancet* 350:1404. doi: 10.1016/s0140-6736(05)65185-9

Ochman, H., and Moran, N. A. (2001). Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* 292, 1096–1099. doi: 10.1126/science.1058543

Ottini, L., and Falchetti, M. (2010). [When history meets molecular medicine: molecular history of human tuberculosis]. *Med. Secoli.* 22, 611–632.

Porter, A. F., Duggan, A. T., Poinar, H. N., and Holmes, E. C. (2017). Comment: characterization of two historic smallpox specimens from a Czech museum. *Viruses* 9:276. doi: 10.3390/v9100276

Rasmussen, S., Allentoft, M. E., Nielsen, K., Orlando, L., Sikora, M., Sjögren, K. G., et al. (2015). Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell* 163, 571–582. doi: 10.1016/j.cell.2015.10.009

Reddy, T. B. K., Thomas, A. D., Stamatis, D., Bertsch, J., Isbandi, M., and Jansson, J. (2014). The genomes online database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucl. Acids Res.* 43, D1099–D1106.

Riesenfeld, C. S., Schloss, P. D., and Handelsman, J. (2004). METAGENOMICS: genomic analysis of microbial communities. *Annu. Rev. Genet.* 38, 525–552. doi: 10.1146/annurev.genet.38.072902.091216

Roberts, C., and Ingham, S. (2008). Using ancient DNA analysis in palaeopathology: a critical analysis of published papers, with recommendations for future work. *Int. J. Osteoarchaeol.* 18, 600–613. doi: 10.1002/oa.966

Ross, Z. P., Klunk, J., Fornaciari, G., Giuffra, V., Duchêne, S., and Duggan, A. T. (2018). The paradox of HBV evolution as revealed from a 16th century mummy. *PLoS Pathog.* 14:e1006750. doi: 10.1371/journal.ppat.1006750

Roumagnac, P., Weill, F.-X., Dolecek, C., Baker, S., Brisse, S., and Chinh, N. T. (2006). Evolutionary history of *Salmonella typhi*. *Science* 314, 1301–1304.

Salo, W. L., Aufderheide, A. C., Buikstra, J., and Holcomb, T. A. (1994). Identification of mycobacterium tuberculosis DNA in a pre-Columbian Peruvian mummy. *PNAS* 91, 2091–2094. doi: 10.1073/pnas.91.6.2091

Sarkissian, C. D., Allentoft, M. E., Ávila-Arcos, M. C., Barnett, R., Campos, P. F., and Cappellini, E. (2015). Ancient genomics. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370:20130387. doi: 10.1098/rstb.2013.0387

Schloss, P. D., and Handelsman, J. (2005). Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol.* 6:229.

Schuenemann, V. J., Bos, K., DeWitte, S., Schmedes, S., Jamieson, J., and Mittnik, A. (2011). Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the black death. *Proc. Natl. Acad. Sci. U.S.A.* 108, E746–E752. doi: 10.1073/pnas.1105107108

Schuenemann, V. J., Singh, P., Mendum, T. A., Krause-Kyora, B., Jäger, G., and Bos, K. I. (2013). Genome-wide comparison of medieval and modern *Mycobacterium leprae*. *Science* 341, 179–183. doi: 10.1126/science.1238286

Secka, O., Moodley, Y., Antonio, M., Berg, D. E., Tapgun, M., and Walton, R. (2014). Population genetic analyses of *Helicobacter pylori* isolates from Gambian adults and children. *PLoS One* 9:e109466. doi: 10.1371/journal.pone.0109466

Shapiro, B., Barlow, A., and Heintzman, PD. (Eds). (2019). *Ancient DNA: Methods and Protocols*. Totowa, NJ: Humana Press.

Skelly, E., Kapellas, K., Cooper, A., and Weyrich, L. S. (2018). Consequences of colonialism: a microbial perspective to contemporary Indigenous health. *Am. J. Phys. Anthropol.* 167, 423–437. doi: 10.1002/ajpa.23637

Spigelman, M., and Lemma, E. (1993). The use of the polymerase chain reaction (PCR) to detect mycobacterium tuberculosis in ancient skeletons. *Int. J. Osteoarchaeol.* 3, 137–143. doi: 10.1002/oa.1390030211

Spyrou, M. A., Keller, M., Tukhbatova, R. I., Scheib, C. L., Nelson, E. A., and Andrades Valtueña, A. (2019). Phylogeography of the second plague pandemic revealed through analysis of historical *Yersinia pestis* genomes. *Nat. Commun.* 10:4470. doi: 10.1038/s41467-019-12154-0

Spyrou, M. A., Tukhbatova, R. I., Wang, C.-C., Valtueña, A. A., Lankapalli, A. K., Kondrashin, V. V., et al. (2018). Analysis of 3800-year-old Yersinia pestis genomes suggests Bronze Age origin for bubonic plague. *Nat. Commun.* 9:2234.

Swain, F. M. (1969). Paleomicrobiology. *Annu. Rev. Microbiol.* 23, 455–472.

Taylor, G. M., Goyal, M., Legge, A. J., Shaw, R. J., and Young, D. (1999). Genotypic analysis of mycobacterium tuberculosis from medieval human remains. *Microbiology* 145, 899–904. doi: 10.1099/13500872-145-4-899

Tito, R. Y., Macmil, S., Wiley, G., Najar, F., Cleeland, L., and Qu, C. (2008). Phylotyping and functional analysis of two ancient human microbiomes. *PLoS One* 3:e3703. doi: 10.1371/journal.pone.0003703

Tobler, R., Rohrlach, A., Soubrier, J., Bover, P., Llamas, B., Tuke, J., et al. (2017). Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia. *Nature* 544, 180–184.

Tran, T.-N.-N., Aboudharam, G., Raoult, D., and Drancourt, M. (2011). Beyond ancient microbial DNA: nonnucleotidic biomolecules for paleomicrobiology. *BioTechniques* 50, 370–380. doi: 10.2144/000113689

Tsangaras, K., and Greenwood, A. D. (2012). Museums and disease: using tissue archive and museum samples to study pathogens. *Ann. Anat.* 194, 58–73. doi: 10.1016/j.aanat.2011.04.003

Turney, C. S. M., Fogwill, C. J., Golledge, N. R., McKay, N. P., van Sebille, E., Jones, R. T., et al. (2020). Early Last Interglacial ocean warming drovesubstantial ice mass loss from Antarctica. *Proc Natl. Acad. Sci. U.S.A.* 117, 3996–4006.

Vågene, ÅJ., Herbig, A., Campana, M. G., Warinner, C., Spyrou, M. A., and Valtueña, A. A. (2018). *Salmonella enterica* genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat. Ecol. Evol.* 2, 520–528. doi: 10.1038/s41559-017-0446-6

Wagner, D. M., Klunk, J., Harbeck, M., Devault, A., Waglechner, N., and Sahl, J. W. (2014). *Yersinia pestis* and the plague of justinian 541–543 AD: a genomic analysis. *Lancet Infect. Dis.* 14, 319–326. doi: 10.1016/S1473-3099(13)70323-2

Warinner, C., Rodrigues, J. F. M., Vyas, R., Trachsel, C., Shved, N., and Grossmann, J. (2014). Pathogens and host immunity in the ancient human oral cavity. *Nat. Genet.* 46, 336–344. doi: 10.1038/ng.2906

Wertheim, J. O. (2017). Viral evolution: mummy virus challenges presumed history of smallpox. *Curr. Biol.* 27, R119–R120. doi: 10.1016/j.cub.2016.12.008

Weyrich, L. S., Dobney, K., and Cooper, A. (2015). Ancient DNA analysis of dental calculus. *J. Hum. Evol.* 79, 119–124. doi: 10.1016/j.jhevol.2014.06.018

Weyrich, L. S., Duchene, S., Soubrier, J., Arriola, L., Llamas, B., Breen, J., et al. (2017). Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. *Nature* 544, 357–361.

Weyrich, L. S., Farrer, A. G., Eisenhofer, R., Arriola, L. A., Young, J., and Selway, C. A. (2019). Laboratory contamination over time during low-biomass sample analysis. *Mol. Ecol. Resour.* 19, 982–996. doi: 10.1111/1755-0998.13011

Wilbur, A. K., Bouwman, A. S., Stone, A. C., Roberts, C. A., Pfister, L. A., and Buikstra, J. (2009). Deficiencies and challenges in the study of ancient tuberculosis DNA. *J. Archaeol. Sci.* 36, 1990–1997. doi: 10.1016/j.jas.2009.05.020

Willerslev, E., and Cooper, A. (2005). Review paper. Ancient DNA. *Proc. R. Soc. B* 272, 3–16. doi: 10.1098/rspb.2004.2813

Wolfe, N. D., Dunavan, C. P., and Diamond, J. (2007). Origins of major human infectious diseases. *Nature* 447, 279–283. doi: 10.1038/nature05775

World Health Organization [WHO] (2016). *WHO | Global Leprosy Update, 2015: Time for Action, Accountability and Inclusion*. Available online at: http//www.who.int/lep/resources/who_wer9135/en/ (accessed October 18, 2016).

Yergeau, E., Hogues, H., Whyte, L. G., and Greer, C. W. (2010). The functional potential of high Arctic permafrost revealed by metagenomic sequencing, qPCR and microarray analyses. *ISME J.* 4, 1206–1214. doi: 10.1038/ismej.2010.41

Yousten, A. A., and Rippere, K. E. (1997). DNA similarity analysis of a putative ancient bacterial isolate obtained from amber. *FEMS Microbiol. Lett.* 152, 345–347. doi: 10.1111/j.1574-6968.1997.tb10450.x

Zhou, Z., Lundstrøm, I., Tran-Dien, A., Duchêne, S., Alikhan, N. F., and Sergeant, M. J. (2018). Pan-genome analysis of ancient and modern *Salmonella enterica* demonstrates genomic stability of the invasive para C lineage for millennia. *Curr. Biol.* 28, 2420.e–2428.e. doi: 10.1016/j.cub.2018.05.058

Ziesemer, K. A., Mann, A. E., Sankaranarayanan, K., Schroeder, H., Ozga, A., Brandt, B. W., et al. (2015). Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. *Sci. Rep.* 5:16498.

Zietz, B. P., and Dunkelberg, H. (2004). The history of the plague and the research on the causative agent *Yersinia pestis*. *Int. J. Hyg. Environ. Health* 207, 165–178. doi: 10.1078/1438-4639-00259

Zink, A. R., Molnár, E., Motamedi, N., and Palfi, G. (2007). Molecular history of tuberculosis from ancient mummies and skeletons. *Int. J. Osteoarchaeol.* 17, 380–391. doi: 10.1002/oa.909

Zink, A. R., Reischl, U., Wolf, H., and Nerlich, A. G. (2002). Molecular analysis of ancient microbial infections. *FEMS Microbiol. Lett.* 213, 141–147. doi: 10.1111/j.1574-6968.2002.tb11298.x

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 23 June 2020
doi: 10.3389/fevo.2020.00185

# CASCADE: A Custom-Made Archiving System for the Conservation of Ancient DNA Experimental Data

*Dirk Dolle*†, *Antoine Fages*†, *Xavier Mata, Stéphanie Schiavinato, Laure Tonasso-Calvière, Lorelei Chauvey, Stefanie Wagner, Clio Der Sarkissian, Aurore Fromentier, Andaine Seguin-Orlando and Ludovic Orlando**

*Laboratoire d'Anthropologie et d'Imagerie de Synthèse, CNRS UMR 5288, Faculté de Médecine Purpan, Toulouse, France*

OPEN ACCESS

**Edited by:**
*Nic Rawlence,*
*University of Otago, New Zealand*

**Reviewed by:**
*Axel Barlow,*
*Nottingham Trent University,*
*United Kingdom*
*André Elias Rodrigues Soares,*
*Uppsala University, Sweden*
*Christian N. K. Anderson,*
*Brigham Young University,*
*United States*

***Correspondence:**
*Ludovic Orlando*
*ludovic.orlando@univ-tlse3.fr;*
*orlando.ludovic@gmail.com*

† *These authors have contributed equally to this work*

**Specialty section:**
*This article was submitted to Paleoecology, a section of the journal Frontiers in Ecology and Evolution*

**Received:** *19 December 2019*
**Accepted:** *25 May 2020*
**Published:** *23 June 2020*

**Citation:**
*Dolle D, Fages A, Mata X, Schiavinato S, Tonasso-Calvière L, Chauvey L, Wagner S, Der Sarkissian C, Fromentier A, Seguin-Orlando A and Orlando L (2020) CASCADE: A Custom-Made Archiving System for the Conservation of Ancient DNA Experimental Data. Front. Ecol. Evol. 8:185. doi: 10.3389/fevo.2020.00185*

The field of ancient genomics has undergone a true revolution during the last decade. Input material, time requirements and processing costs have first limited the number of specimens amenable to genome sequencing. However, the discovery that archeological material such as petrosal bones can show increased ancient DNA preservation rates, combined with advances in sequencing technologies, molecular methods for the recovery of degraded DNA fragments and bioinformatics, has vastly expanded the range of samples compatible with genome-wide investigation. Experimental procedures for DNA extraction, genomic library preparation and target enrichment have become more streamlined, and now also include automation. These procedures have considerably reduced the amount of work necessary for data generation, effectively adapting the processing capacity of individual laboratories to the increasing numbers of analyzable samples. Handling vast amounts of samples, however, comes with logistical challenges. Laboratory capacities, equipment, and people need to be efficiently coordinated, and the progress of each sample through the different experimental stages needs to be fully traceable, especially as archeological remains of animals or plants are often provided and/or handled by many different collaborators. Here we present CASCADE, a laboratory information management system (LIMS) dealing with the specificities of ancient DNA sample processing and tracking, applicable by large and small laboratories alike, and scalable to large projects involving the analysis of thousands of samples and more. By giving an account of the specimen's progress at any given analytical step, CASCADE not only optimizes the collaborative experience, including real-time information sharing with third parties, but also improves the efficacy of data generation and traceability in-house.

**Keywords: ancient DNA, laboratory management, database, LIMS, traceability, conservation, collaborative sharing**

## INTRODUCTION

The first draft of the human genome was released in 2001, following 20 years of extensive collaborative efforts among large research centers across the world (Lander et al., 2001; Venter et al., 2001). The first prehistoric human genome was released almost a decade later (Rasmussen et al., 2010) and was immediately followed by the genome characterization of two extinct groups of hominins, Neanderthals (Green et al., 2010) and Denisovans (Reich et al., 2010). Now, another decade later, the number of human genomes characterized has become truly astronomical, and it is estimated that several millions of living individuals, and several thousand prehistoric ones, have had their genome sequenced (Marciniak and Perry, 2017; Brunson and Reich, 2019 Regalado, 2019). Extensive time-series of genomes are also becoming available for organisms other than humans, including their pathogens (e.g., *Yersinia pestis*, the agent of the plague, see Spyrou et al., 2019 for a review), animal domesticates (e.g., horses, Fages et al., 2019), and plants (e.g., maize, Kistler et al., 2018). The number of published metagenomes sampled from the environment (e.g., ancient lake sediments, Pedersen et al., 2015) and mammal-associated microbial communities (e.g., in ancient dental calculus, Mann et al., 2018) is also on the rise.

The reasons for such a success are manifold. First, the DNA data production capacity and costs of next-generation sequencing instruments have constantly improved (Metzker, 2010; Goodwin et al., 2016). Second, specific types of osseous material, such as petrosal bones (Pinhasi et al., 2015) and tooth cementum (Damgaard et al., 2015), have been found to show better DNA preservation rates than previously explored sources, and to be generally less prone to contamination by environmental microbes. These developments have lowered the sequencing efforts needed to retrieve significant coverage of the focal genome and have consequently reduced the time required for and costs incurred by ancient genome characterization. Third, an entire array of innovative molecular solutions has been developed to facilitate the recovery and manipulation of ancient DNA (aDNA) molecules, including DNA extraction (Dabney et al., 2013; Boessenkool et al., 2017; Glocke and Meyer, 2017; Korlević and Meyer, 2019), incorporation into DNA libraries (Gansauge and Meyer, 2013; Gansauge et al., 2017; Carøe et al., 2018; Rohland et al., 2018), handling of DNA damage (Gansauge and Meyer, 2014; Gansauge et al., 2017), and target enrichment of hundreds of thousands to over a million pre-selected loci across the genome (Haak et al., 2015; Harney et al., 2018; Mathieson et al., 2018; Olalde et al., 2018). As a result, aDNA projects have become increasingly large-scale, and it is now common that several hundred of specimens are processed within a single study. For example, the survey of genome-wide variation among humans in Iberia carried out by Olalde and colleagues included 271 individual specimens spanning the last ∼5,000 years (Olalde et al., 2019), while the study from Damgaard et al. (2018) released no fewer than 137 genome sequences of ancient humans from across the Eurasian steppes and spanning the last 4,000 years.

This increase in scale is impressive, especially when contrasted with the challenges ancient genomics still faces, first of which is sample availability. While new methods have expanded the range of suitable samples, finding relevant samples in the first place can be very difficult. Once a promising excavation site is found, the number of suitable samples at this site is *a priori* unknown and may only become apparent after months if not years. However, once unearthed, samples should be processed rapidly to prevent further degradation of endogenous DNA. Unfortunately, sample processing is often destructive and hence reduces the availability of samples for other fields like e.g., Archeology. It is therefore important to assess as swiftly as possible whether samples found at a given site are suitable for ancient genomics to limit material destruction as much as possible, either because initial screening is negative or enough material has been obtained. This also helps prioritizing areas and/or time periods that require further excavation and sampling. Timely processing however, can be challenging due to available laboratory capacity. Because of the risk of further degradation and contamination, the handling of aDNA requires dedicated, well-isolated clean rooms which have to undergo regular cleaning and sterilization cycles between uses in order to prevent additional contamination. As a result of these specific requirements suitable laboratory space is often limited. Strict experimental procedures, laboratory facility maintenance, and resource management are thus indispensable for efficient sample processing especially when trying to scale up analysis. It is furthermore imperative that laboratory personnel have access to the full status of any given sample at all times, especially as experiments may require the processing of specific samples and/or the replication of particular steps. In addition to informing decisions and providing context for future specimens, being able to track sample information throughout the whole aDNA data production process is paramount to a number of quality control meta-analyses, such as detecting batch effects, including contaminated experimental sessions. These processes are also crucial for assessing performance of steps and/or protocols, project reporting, or for obtaining preliminary background information for grant applications. Such achievements require tools that are up to the task of staying on top of constantly growing and changing data as well as all underlying procedural steps.

In this study, we present CASCADE, a Laboratory Information Management System (LIMS), tailor-made for the genetic processing of paleontological remains. The Custom-made Archiving System for the Conservation of Ancient DNA Experimental data implemented within CASCADE provides a user-friendly, web-based environment to track all experimental phases involved in the preparation, extraction, library construction, amplification and sequencing of aDNA. It delivers a full environment capable of both storing and querying the information from a web browser. It also supports barcode assisted identification of tube content at all documented experimental steps. It is available for free, together with a companion documentation that provides full installation instructions and user guidelines. We anticipate that CASCADE will facilitate the traceability, sharing and long-term conservation of the experimental metadata associated with aDNA analyses.

# MATERIALS AND METHODS

## CASCADE Setup

In order to make CASCADE as portable as possible, we decided to embed it inside a virtual machine (VM) available for all major operating systems (OS, e.g., Windows, Linux, Mac OSX). This was achieved using Oracle VM VirtualBox v5.2.[1] A VM is a fully self-contained simulated computer system that can be executed on any host system (i.e., a physical computer) that has the VM software installed and enough resources for the simulated system. The VM created for CASCADE uses only a single simulated CPU and 2GB of RAM, and hence should be able to run on most currently available computer systems. It was tested without returning problems on desktop and laptop computers with Intel CORE i5 and CORE i7 CPUs and 8GB of RAM using either Windows 10, MacOS El Capitan / Mojave, or Ubuntu 18.04 LTS as host systems. As guest OS (i.e., the operating system running inside the VM), we chose Ubuntu 18.04 LTS,[2] which is available free online.

The virtual disk image (VDI), i.e., the file that contains the simulated hard disk of the VM, requires around 10GB of disk space which is mostly due to the size of the guest OS and other required software. Software includes the NGINX[3] v.1.14.0 web-server executing the CASCADE source code, and the MySQL[4] v14.14 database management system that handles all the data stored. Other prerequisites are the Laravel[5] PHP framework v5.6, the Vue.js[6] JavaScript framework v2.6.10, and the Bootstrap 4[7] JavaScript & CSS library. CASCADE's back-end source code is fully written in PHP 7.2 using Laravel's classes while the front-end is written in JavaScript with the support of Vue.js. Styling is based on Bootstrap 4 with the addition of vector graphics from the Font Awesome[8] vector icon library. Vector graphics sourced from this library are under the Creative Commons Attribution 4.0 International license.[9]

CASCADE is accessed via web browser. It was developed specifically for Mozilla Firefox but has also been tested with recent versions of Google Chrome without encountering problems. Especially when the database reaches larger numbers of records (i.e., ten-thousands of entries), Chrome seems to perform better, although parts of the layout may be interpreted differently to Firefox. The source code as well as the fully installed and configured VM can be provided upon request (please direct requests to one of the following authors: LO, ludovic.orlando@univ-tlse3.fr; XM, xavier.mata@univ-tlse3.fr; CD, clio.dersarkissian@univ-tlse3.fr). We also provide an installation manual alongside the VM for those users willing to create and configure their own VM for running CASCADE. However, please note that the source code is available only upon request and requires an Atlassian Bitbucket[10] account which can be created for free. Using Bitbucket's collaborator feature allows us to safely exchange the code with trusted parties. In case such an account is not wanted, the source code can also be found in a folder inside the VM as indicated in the companion manual.

Our goal is to provide CASCADE free to all scientists interested. The procedures for obtaining CASCADE were, however, developed as precautionary measures to reduce the risk of hacking. To maximize safety, we also recommend installing CASCADE on computers connected to firewall-protected, laboratory-internal networks and allowing access from outside the network only through VPN.

It is the user's responsibility to ensure the security of the computer installation, network configuration and password management for any computer running CASCADE in order to maintain the safety of the data stored in CASCADE. Hence, it should be installed and run only on computers for which access control can be guaranteed at all times, either directly or through the network. Moreover, as CASCADE can be used to store personal data (e.g., from collaborators and study participants), it is subjected to legislation and rules about the protection of personally identifiable information in force in the users' country, state and/or institutions (e.g., GDPR in the European Union) and with which the user must comply (e.g., personal data anonymization, encryption, ethical clearance). Like all software under MIT license, CASCADE is provided "*as is*" without warranty of any kind. As a consequence, the authors of CASCADE cannot be held accountable for data/system loss as a result of security breaches.

## Database Schema Design

Each relational database project begins with the schema design during which all attributes required to describe every aspect of the system are defined. These are then further processed to remove redundancies and establish relationships between the different parts of the data. This procedure is called database normalization, which aims at satisfying so-called "*normal forms*" (Codd, 1970). A data model that reaches at least the third normal form is usually described as "*normalized*" (Date, 2003). Once in this configuration, most if not all anomalies that might arise from adding, deleting, or modifying data are removed. This feature guarantees the present and future integrity of the data and represents one of the major strengths of relational databases. Another one is that the relations established between the data sets allow flexibly combining them in different ways so as to query exactly the information required for certain analyses or tasks.

## Figures and Screenshots

Screenshots were taken using the screenshot tool of MacOS Mojave (v10.14.6) and subsequently processed in GIMP[11] v.2.10. Where necessary, text was replaced with a "– *REDACTED* –" label to not expose account and personal information of our collaborators. All other Figures were created using

[1] https://www.virtualbox.org/wiki/VirtualBox

[2] http://releases.ubuntu.com/18.04/

[3] https://www.nginx.com/

[4] https://www.mysql.com/

[5] https://laravel.com/

[6] https://vuejs.org/

[7] https://getbootstrap.com/

[8] https://fontawesome.com/

[9] https://creativecommons.org/licenses/by/4.0/

[10] https://bitbucket.org/product/

[11] https://www.gimp.org/

Inkscape[12] v0.92.4 and vector graphics from the Font Awesome vector icon library.

# RESULTS

## General Overview

The core functionality of any LIMS revolves around the handling and tracking of laboratory specimens. In the case of most aDNA projects, these requirements include sample registration, pulverization, and DNA extraction from the obtained powder, followed by library preparation, amplification, quantification, pooling and sequencing. Additional steps which might be used to increase quality, specificity, or yield of endogenous sequences consist of DNA damage repair by enzymatic treatment, and/or DNA capture methods for target enrichment of specific DNA sub-fractions. A commonly used repair method is the incubation of aDNA extracts with an enzymatic mixture consisting of Uracil DNA glycosylase and Endonuclease VIII (USER treatment, New England Biolabs), which can eliminate most or even all those cytosines that have been deaminated post-mortem (Rohland et al., 2015), and which represent the most common aDNA damage (Briggs et al., 2007).

All the steps described above are implemented in the CASCADE processing pipeline (**Figure 1**). At each stage of the pipeline, primary and secondary data relating to the experimental step are recorded. Primary data include attributes directly related to a given specimen, e.g., the condition of the sample, the amount of powder obtained, or the volume of enzyme used for treatment. Secondary data on the other hand refer to attributes that are only indirectly related to the specimen e.g., the coordinates of the excavation site, the return address of the sample provider, or the details of the protocol used for processing. All these data are handled by CASCADE leveraging a relational database recording 1,155 attributes (fields or columns) grouped into 97 different record types (stored as record sets or tables) which are linked together through a network of 374 private key – foreign key relations (**Tables 1–3**).

Upon registration of a given sample, the system generates a unique identifier and then incorporates additional information as the sample progresses through the pipeline. This feature enables unambiguous identification at every step until the sample reaches the final stage of data analysis. The partitioning of data and their interconnection by a web of relations allows users to flexibly combine individual data sets to query exactly the information required for a certain analysis or task. This flexibility is made available to users through a query system allowing them not only to create queries, but also to store and re-run them at a later stage (e.g., after new content has been added to the database). In effect, it grants users the power to add to the functionality of CASCADE as they see fit. In order to provide an example of what is possible, we have used this query system to generate a series of pre-built queries which we find useful in our own laboratory practice. These include, for example, summaries of

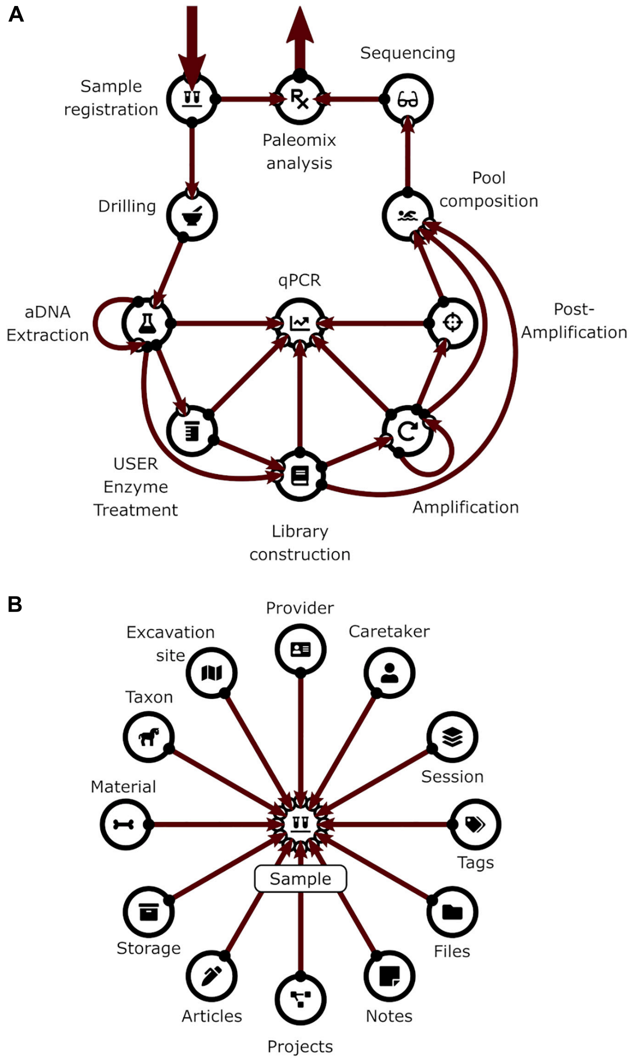


**FIGURE 1 | (A)** Pipeline backbone graph showing the flow of experimental data through CASCADE. Samples and their related data are registered first. Subsequently, they proceed through the pulverization or "*Drilling*" stage to the extraction step. Once extracted, they continue to library generation, either directly or after undergoing USER treatment. Should analyses reveal that the initial extraction round did not succeed in releasing satisfactory endogenous material, new extractions can also be generated from the remaining extraction pellets (loop from and to extraction) before being further processed. Generated libraries can be pooled immediately or after additional amplification. Again, should initial amplification prove not to be sufficient, re-amplification (loop from and to amplification), target enrichment, or other post-amplification protocols can be executed prior to pooling. All intermediates from extraction to post-amplification can be included in analyses via the feature "*qPCR*." Paths to the sequencing data generated for several samples of the same specimen (a "*Sample group*") are combined and fed in the Paleomix computational pipeline (Schubert et al., 2014) for retrieving basic estimates, including endogenous content. The results obtained represent the final pipeline output. **(B)** Display of the most important non-experimental data types related to sample data. Samples have the largest number of relations to non-experimental data. Some of these are shared with other experimental data types (e.g., "*Storage*," "*Notes*," "*Files*"). Icons in both panels are the same as those used for CASCADE's menus and taken from the Font Awesome vector icon library.

[12] https://www.inkscape.org

**TABLE 1 |** Data structure overview (part 1): presets.

| Data | Presets | | | For experiment types |
|---|---|---|---|---|
| | Tables | Fields | Relations | |
| *Contacts* | 12 | 105 | 38 | S,D,E,U,L,Q,A,P,O,R |
| *Excavation sites* | 4 | 36 | 11 | S |
| *Taxa* | 2 | 17 | 5 | S |
| *Materials* | 3 | 22 | 8 | S |
| *Articles* | 2 | 20 | 6 | S |
| *Projects* | 3 | 26 | 11 | S |
| *Protocols* | 2 | 17 | 6 | D,E,U,L,Q,A,P,O |
| *Oligos & Adapters* | 3 | 30 | 10 | L,A |
| *Sequencing & more* | 5 | 40 | 11 | A,R |
| *Storage locations* | 5 | 41 | 15 | S,D,E,U,L,Q,A,P,O,R |
| *Barcodes* | 1 | 8 | 2 | S,D,E,U,L,Q,A,P,O |
| *Files* | 5 | 44 | 14 | S,D,E,U,L,Q,A,P,O,R,N,I |
| *Notes* | 1 | 11 | 2 | S,D,E,U,L,Q,A,P,O,R,N,I |
| *Tags* | 2 | 17 | 4 | S,D,E,U,L,Q,A,P,O,R |
| SUM | 50 | 434 | 143 | |

*Tables dealing with data for "Contacts" include those for addresses, phone numbers, email addresses and personal data of collaborators and laboratory staff. These "Contact" records are then connected to experimental and other data e.g., representing sample providers, contact partners, experimenters, project members etc. Tables for "Taxa" and "Materials" deal with the origin and type of samples, "Oligos" form the basis for adaptors and primers for library building and amplification, respectively. "Storage locations" encompass the physical location and labels of fridges, freezers, boxes, and storage shelves for samples and laboratory specimens, while fields in the "Files" section point to their digital counterparts i.e., the data generated from them (e.g., pictures, 3D models, etc.). Each of the different data groups was designed to allow varying degrees in specificity that can be tailored to the needs of the individual laboratory. "Taxa" for instance can be as broad as a whole taxonomic kingdom or as specific as a certain phenotypic group within a breed. In a similar way, "Materials" can either be a full limb or a single bone or tooth fragment.*

**TABLE 2 |** Data structure overview (part 2): experiments and related data.

| Data | Experiments | | | Valid source types |
|---|---|---|---|---|
| | Tables | Fields | Relations | |
| *[S] Samples* | 2 | 47 | 17 | - |
| *[D] Drillings* | 1 | 23 | 6 | S |
| *[E] Extractions* | 1 | 21 | 6 | D,E |
| *[U] USER Treatments* | 1 | 20 | 6 | E |
| *[L] Libraries* | 1 | 21 | 7 | E,U |
| *[Q] qPCRs* | 2 | 27 | 9 | E,U,L,A,P |
| *[A] Amplifications* | 1 | 37 | 11 | L,A |
| *[P] Post-Amplifications* | 1 | 18 | 6 | A |
| *[O] Pools* | 2 | 28 | 8 | L,A,P |
| *[R] Sequencing runs* | 2 | 25 | 12 | P |
| *[X] Paleomix runs* | 4 | 46 | 13 | G |
| *[N] Sessions* | 14 | 272 | 90 | S,D,E,U,L,Q,A,P,O,R |
| *[G] Sample groups* | 2 | 16 | 6 | S |
| *[I] Imaging* | 3 | 26 | 10 | S |
| SUM | 37 | 627 | 207 | |

*"Samples" deals with the actual physical specimens provided to the laboratory. "Drillings" covers all aspects of powder generation from samples, "Extractions" handles the extraction of aDNA from generated powder, and "USER Treatments" the optional step of DNA repair. "Libraries" handles the process of library generation from extracts (treated or untreated) and "Amplifications" the multiplication of DNA fragments (in the library). "Post-Amplifications" includes different types of protocols (e.g., target enrichment through selective capture of endogenous DNA) applied prior to sequencing. "Pools" records type, number and relative concentrations of different libraries (amplified or non-amplified, with or without subsequent protocols applied) whereas "Sequencing runs" mainly provides attachment points at component resolution for the read files generated. "Paleomix runs" stores the data parsed from the report files produced by the Paleomix pipeline. "Sessions" logs which specimens were processed in the laboratory at the same time. "Sample groups" allows storing the relation between different samples (e.g., from the same individual) and provides the attachment point for Paleomix runs. Finally, "Imaging," deals with the photographic documentation of samples as well as 3D scans to document the samples' state prior to processing.*

**TABLE 3 |** Data structure overview (part 3): tables related to database functionality.

| Data | Functionality | | | Functionality |
|---|---|---|---|---|
| | Tables | Fields | Relations | |
| *Waiting list* | 1 | 11 | 3 | Monitors waiting specimen |
| *Defaults* | 1 | 9 | 2 | Allows definition of field presets |
| *Queries* | 1 | 17 | 2 | Handles all stored queries |
| *Administration* | 6 | 50 | 15 | General database administration |
| *Backup* | 1 | 7 | 2 | Logs manual backups |
| SUM | 10 | 94 | 24 | |

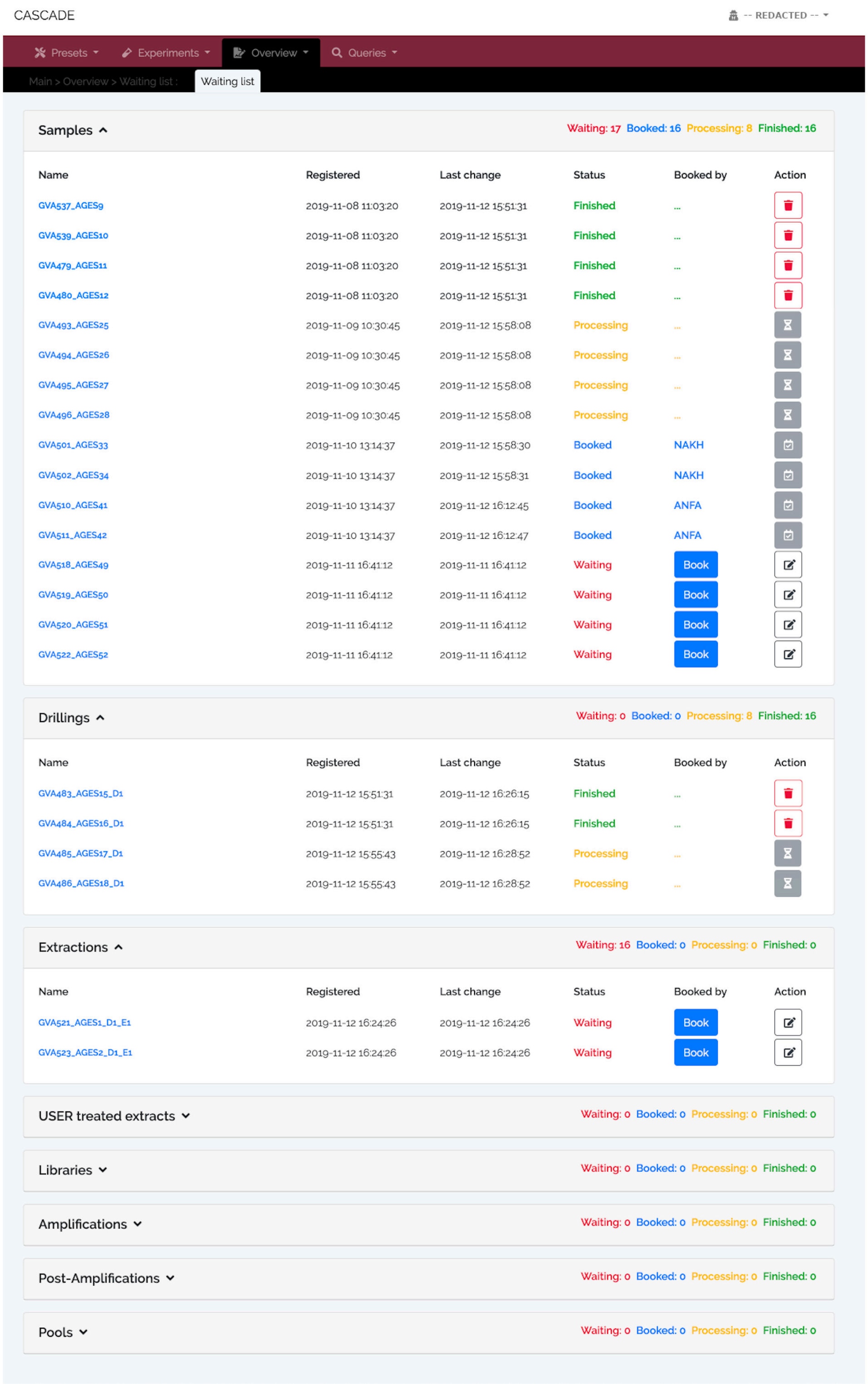
*"Waiting list" stores which specimens are being processed, booked, or waiting for the next step. "Defaults" allows users to set presets for each field displayed on forms that handle experimental data types in order to make data entry faster and more streamlined by reducing repetition. "Queries" store each query generated by CASCADE's query system, while "Administration" includes tables for registered database users, registration permissions, events that modified the database etc. Finally, "Backup" keeps track of every manual backup initiated by the administrator and provides download links to the backed-up data and the backup log.*

samples available and/or processed per individual archeological site, region and/or time period, and overviews of the status of any given sample in the experimental production chain.

## The Database User Interface

The database interface consists of two main parts, corresponding to the "*Presets*" and the "*Experiments*" sections. The "*Presets*" section describes all data that can exist independently of experiments but which form the majority of dependencies for experimental entries. These include, among other things, collaborator and user contact information, excavation sites, taxa, materials, protocols, and oligo-nucleotides used for preparing DNA library adapters, library indexing and amplification. The "*Experiments*" section mainly handles the actual experimental data, but also includes data types that are closely associated, such as imaging records (if available), sample groups, and analytical results of different types, depending on the experimental step considered.

In a typical use case scenario, any newly arrived sample is first checked for its "*Presets*" requirements (e.g., new collaborator, excavation site etc.) before the sample itself is added to CASCADE. Upon registration, the sample is automatically added to a "*Waiting list*" (**Figure 2**) which makes it immediately visible to all laboratory staff and enables them to book it for processing directly through this interface. Once processing starts, the sample status is continuously updated, which is



CASCADE

Samples ^ Waiting: 17 Booked: 16 Processing: 8 Finished: 16

| Name | Registered | Last change | Status | Booked by | Action |
|---|---|---|---|---|---|
| GVA537_AGES9 | 2019-11-08 11:03:20 | 2019-11-12 15:51:31 | Finished | ... | 🗑 |
| GVA539_AGES10 | 2019-11-08 11:03:20 | 2019-11-12 15:51:31 | Finished | ... | 🗑 |
| GVA479_AGES11 | 2019-11-08 11:03:20 | 2019-11-12 15:51:31 | Finished | ... | 🗑 |
| GVA480_AGES12 | 2019-11-08 11:03:20 | 2019-11-12 15:51:31 | Finished | ... | 🗑 |
| GVA493_AGES25 | 2019-11-09 10:30:45 | 2019-11-12 15:58:08 | Processing | ... | ⌛ |
| GVA494_AGES26 | 2019-11-09 10:30:45 | 2019-11-12 15:58:08 | Processing | ... | ⌛ |
| GVA495_AGES27 | 2019-11-09 10:30:45 | 2019-11-12 15:58:08 | Processing | ... | ⌛ |
| GVA496_AGES28 | 2019-11-09 10:30:45 | 2019-11-12 15:58:08 | Processing | ... | ⌛ |
| GVA501_AGES33 | 2019-11-10 13:14:37 | 2019-11-12 15:58:30 | Booked | NAKH | 📅 |
| GVA502_AGES34 | 2019-11-10 13:14:37 | 2019-11-12 15:58:31 | Booked | NAKH | 📅 |
| GVA510_AGES41 | 2019-11-10 13:14:37 | 2019-11-12 16:12:45 | Booked | ANFA | 📅 |
| GVA511_AGES42 | 2019-11-10 13:14:37 | 2019-11-12 16:12:47 | Booked | ANFA | 📅 |
| GVA518_AGES49 | 2019-11-11 16:41:12 | 2019-11-11 16:41:12 | Waiting | Book | ✎ |
| GVA519_AGES50 | 2019-11-11 16:41:12 | 2019-11-11 16:41:12 | Waiting | Book | ✎ |
| GVA520_AGES51 | 2019-11-11 16:41:12 | 2019-11-11 16:41:12 | Waiting | Book | ✎ |
| GVA522_AGES52 | 2019-11-11 16:41:12 | 2019-11-11 16:41:12 | Waiting | Book | ✎ |

Drillings ^ Waiting: 0 Booked: 0 Processing: 8 Finished: 16

| Name | Registered | Last change | Status | Booked by | Action |
|---|---|---|---|---|---|
| GVA483_AGES15_D1 | 2019-11-12 15:51:31 | 2019-11-12 16:26:15 | Finished | ... | 🗑 |
| GVA484_AGES16_D1 | 2019-11-12 15:51:31 | 2019-11-12 16:26:15 | Finished | ... | 🗑 |
| GVA485_AGES17_D1 | 2019-11-12 15:55:43 | 2019-11-12 16:28:52 | Processing | ... | ⌛ |
| GVA486_AGES18_D1 | 2019-11-12 15:55:43 | 2019-11-12 16:28:52 | Processing | ... | ⌛ |

Extractions ^ Waiting: 16 Booked: 0 Processing: 0 Finished: 0

| Name | Registered | Last change | Status | Booked by | Action |
|---|---|---|---|---|---|
| GVA521_AGES1_D1_E1 | 2019-11-12 16:24:26 | 2019-11-12 16:24:26 | Waiting | Book | ✎ |
| GVA523_AGES2_D1_E1 | 2019-11-12 16:24:26 | 2019-11-12 16:24:26 | Waiting | Book | ✎ |

USER treated extracts ˅ Waiting: 0 Booked: 0 Processing: 0 Finished: 0

Libraries ˅ Waiting: 0 Booked: 0 Processing: 0 Finished: 0

Amplifications ˅ Waiting: 0 Booked: 0 Processing: 0 Finished: 0

Post-Amplifications ˅ Waiting: 0 Booked: 0 Processing: 0 Finished: 0

Pools ˅ Waiting: 0 Booked: 0 Processing: 0 Finished: 0

**FIGURE 2 |** Waiting list showing specimens at different processing stages. Registered samples ("*Samples*," top), "*Drillings*" (middle), and "*Extractions*" (bottom). No records are available for the other experimental data types at this moment, and, hence, none are displayed. The "*Samples*" panel shows four samples each either "*Waiting*" to be processed, "*Booked*" for processing (by two different users "*NAKH*" and "*ANFA*"), currently undergoing "*Processing*," or already "*Finished*." "*Finished*" entries show the delete trash button on the right allowing the user to remove them from the list and avoid clutter. Deletion does not happen automatically; it thus provides users with feedback of what has been done. Waiting entries have an edit button instead, which leads to a menu allowing all users to attach notes to the entry, e.g., special experimental conditions must be applied, location of the material, etc.

communicated through the waiting list and also displayed on the sample's detail page. For each pipeline step a new entry is automatically created ensuring that the list is always up to date and informing everyone interested about the current status of the sample and the progress it has made. To illustrate this feature, we have pre-filled CASCADE with a toy example, which shows specimens at different stages of the processing pipeline. More details about the database interface can be found in the companion manual.

## Experimental Data Entry

In our own laboratory experience, new samples tend to arrive in batches rather than individually. Furthermore, we also almost exclusively process groups of samples in each pipeline step in order to maximize our clean lab capacity. We hence decided to organize experimental data entry for each step in "*Sessions*," with each session corresponding to a group of samples processed at the same time. Not only does this allow us to keep track of which specimens were subjected to the same conditions, and hence to spot batch effects (e.g., failure to amplify caused by using a new tube of enzyme or switching reagent supplier), it also helps distinguish whether contamination present in one particular specimen could have been introduced in the laboratory or prior to arrival. It further makes data submission more convenient as entries that share many of their parameters can be submitted together rather than one by one, hence reducing the workload and the likelihood of introducing mistakes. On top of that, we decided to allow users to pause data submission and continue it at a later point in time. As a result, laboratory staff can make efficient use of the inevitable waiting time e.g., during pulverization, centrifugation, or incubation periods, which is a real asset, especially in a laboratory setting. While waiting for a specific step to be completed, the data accumulated until the ongoing experimental step can be submitted and the process suspended once the experimental step has finished. This feature thus enables to leverage any available experimental downtime thereby improving overall laboratory efficacy. In order to achieve this, however, we had to implement an intermediate set of tables to hold the temporary data. We therefore decided to provide this mechanism only for experimental data types.

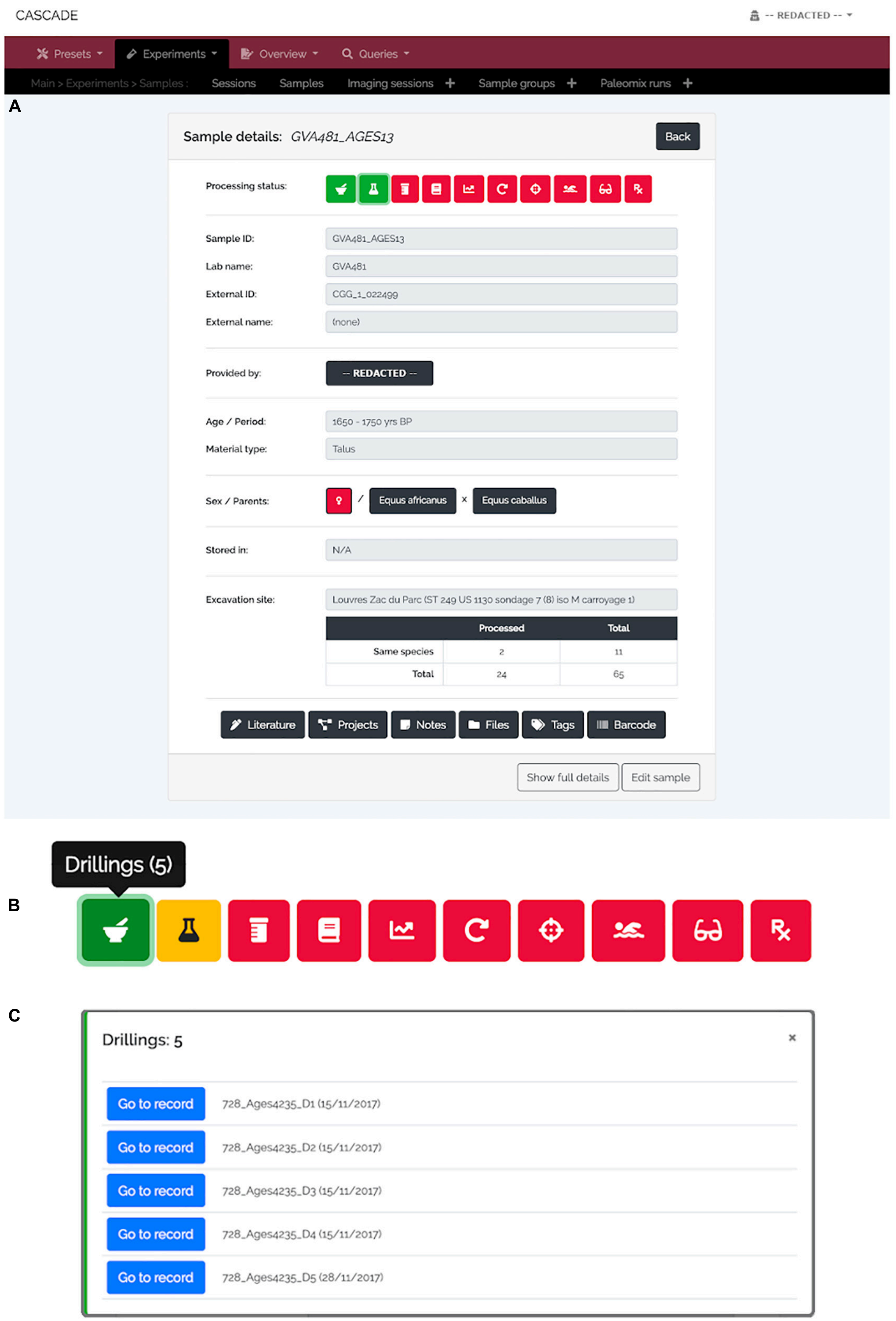
A specimen's progress through the different experimental steps (i.e., pulverization, DNA extraction, USER treatment, library construction, qPCR, amplification, capture/target enrichment, pooling, and sequencing) can be easily monitored via a "*Status bar*" feature available on the detail page of the corresponding sample (**Figure 3A**). A click on the icon of any type of experiment in this bar (**Figure 3B**) displays a list of all related entries of that type stored in the database (**Figure 3C**) and links directly to the corresponding detail pages should the processing of the focal entry be completed. While this feature allows tracking individual samples, tracking whole sets of samples is often equally desirable. For this reason, three higher-level detail menus, one each for collaborators, excavation sites, and taxa, were implemented (**Figures 4–6**). Each of these menus shows a list of all related samples (e.g., provided by a certain collaborator, excavated at a given site, or deriving from the same taxon) and hence helps sample management, provides archeological context, and supports decision making.

## The Query System

As mentioned above, one of the great strengths of relational databases is to offer the possibility of freely combining data from different related records. This feature facilitates the detection of patterns or correlations between parts of the data that otherwise might not have been visible. This kind of data manipulation is usually done using Structured Query Language (SQL) queries. However, due to the vast possibilities of SQL queries and the power that they have over the data and structure of a database, making them available to users is very risky. This risk is both due to the damage that inexperienced users can inadvertently cause but also due to intentional damage by hackers. Another problem is that even SQL queries providing simple output can be complex to write, especially for beginners.

To address these problems, we implemented a Graphical User Interface (GUI) for the creation and execution of SQL "*SELECT*" statements which protects against accidental damage by users while facilitating the construction of complex queries. To further assist users in creating more complex data requests, simpler statements (termed "*Moves*") can be created and stored and then later used in more complicated queries ("*Strategies*") like normal tables. "*Strategies*" further allow all possible set operations (i.e., union, intersection, exclusion, and subtraction) on tables and subqueries, and hence provide all data operations that most users will ever need. We also provide a number of predefined queries that we have found useful in our workflow and which appear in the "*Strategies*" section. The first retrieves the number of samples each of our providers has entrusted to us. This feature helps keep track of all our collaborators, regardless of the number of samples provided. Of equal importance is another query that retrieves the list of collaborators who have contributed samples to a specific project. While CASCADE provides the possibility of assigning people directly to projects, some people might be overlooked, a possibility prevented by this query. Two other search options provide the list of different taxa or samples in hand for a given excavation site. A sixth query returns the number of samples available per excavation site, which helps assess whether we have enough specimens from a given region. Two other queries return all samples based on their country of origin and the different types of material (i.e., tissues) available per taxon, which has proven useful when assessing which material offers the best preservation conditions. Finally, the last query helps with reporting for grant-funding bodies by retrieving the number of articles published for the different projects. These queries are illustrated in **Figure 7** (orange lines).

In addition to generating these types of targeted queries, the advanced search functionality built into CASCADE is equally useful for combining information from different tables into custom-built index views similar to those that already exist for the different menus. These "*virtual tables*" are first created as user-specific advanced searches and subsequently converted using the edit menu. Once converted, "*virtual tables*" cannot be used as source in advanced searches anymore, which makes their scope fixed, but are now available to everyone as queryable index views



Sample details: *GVA481_AGES13*

Processing status:

Sample ID: GVA481_AGES13

Lab name: GVA481

External ID: CGG_1_022499

External name: (none)

Provided by: -- REDACTED --

Age / Period: 1650 - 1750 yrs BP

Material type: Talus

Sex / Parents: ♀ / Equus africanus × Equus caballus

Stored in: N/A

Excavation site: Louvres Zac du Parc (ST 249 US 1130 sondage 7 (8) iso M carroyage 1)

| | Processed | Total |
|---|---|---|
| Same species | 2 | 11 |
| Total | 24 | 65 |

Drillings: 5

Go to record 728_Ages4235_D1 (15/11/2017)

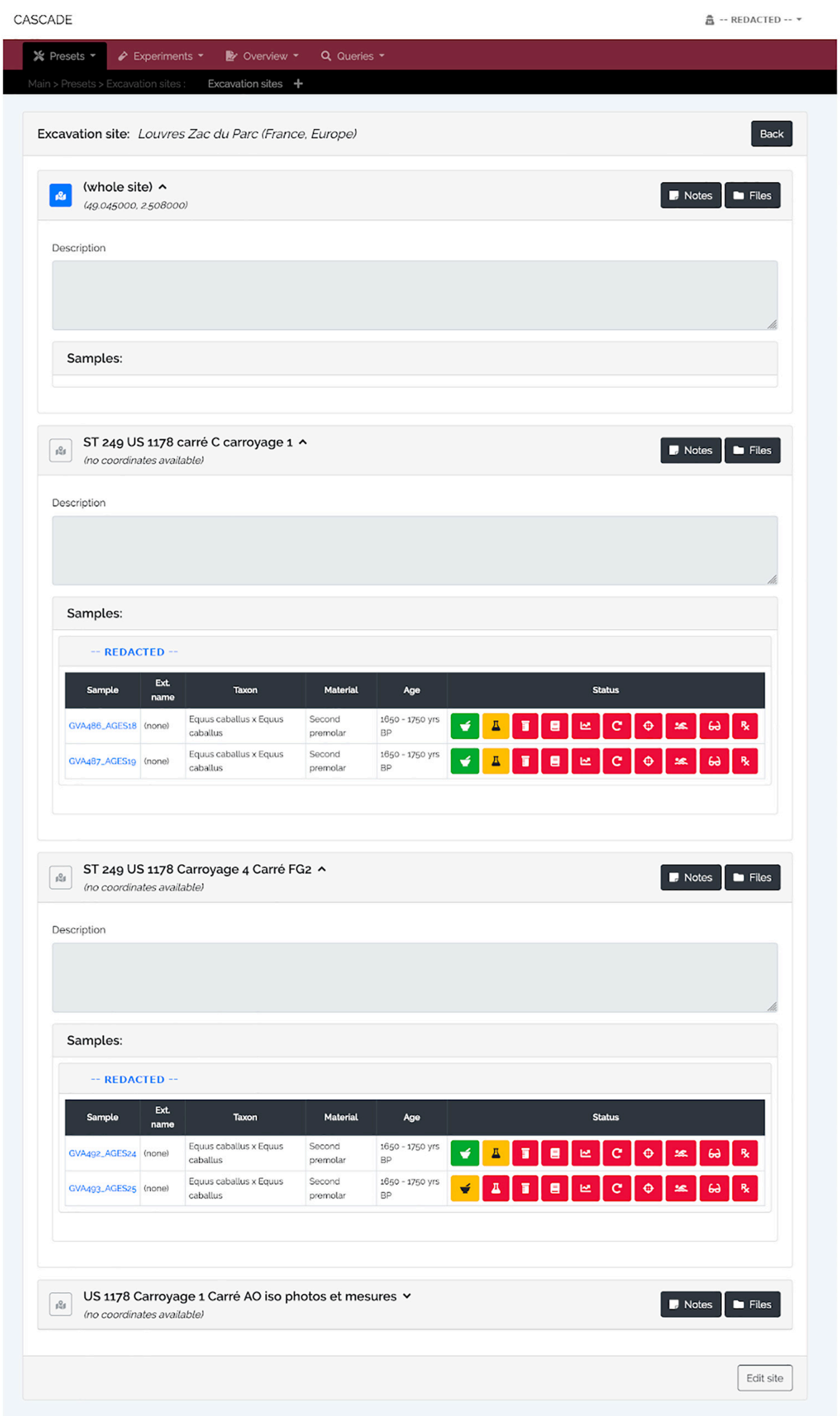
Go to record 728_Ages4235_D2 (15/11/2017)

Go to record 728_Ages4235_D3 (15/11/2017)

Go to record 728_Ages4235_D4 (15/11/2017)

Go to record 728_Ages4235_D5 (28/11/2017)

**FIGURE 3 |** Sample detail view and status bar. **(A)** For each sample, a "*Detail view*" is available that summarizes the most important details about a given sample. The full details are available via the "*Show full details*" button at the bottom left. At the top of the menu, the "*Status bar*" shows the current processing status of a sample (red = "*not done yet*," yellow = "*currently processed*," green = "*finished*"). **(B)** The icons in the status bar represent the different experimental types (from left to right: "*Drillings*," "*Extractions*," "*USER Treatments*," "*Libraries*," "*qPCRs*," "*Amplifications*," "*Post-Amplifications*," "*Pools*," "*Sequencing runs*," and "*Paleomix runs*") and open a pop-up menu **(C)** when clicked in order to list all available specimens per type. The "*Go to record*" button then allows jumping directly to the detail view of the record of interest.



Excavation site: *Louvres Zac du Parc (France, Europe)*

(whole site)
*(49.045000, 2.508000)*

Description

Samples:

ST 249 US 1178 carré C carroyage 1
*(no coordinates available)*

Description

Samples:

-- REDACTED --

| Sample | Ext. name | Taxon | Material | Age | Status |
|---|---|---|---|---|---|
| GVA486_AGES18 | (none) | Equus caballus x Equus caballus | Second premolar | 1650 - 1750 yrs BP | 🟩🟨🟥🟥🟥🟥🟥🟥🟥🟥 |
| GVA487_AGES19 | (none) | Equus caballus x Equus caballus | Second premolar | 1650 - 1750 yrs BP | 🟩🟨🟥🟥🟥🟥🟥🟥🟥🟥 |

ST 249 US 1178 Carroyage 4 Carré FG2
*(no coordinates available)*

Description

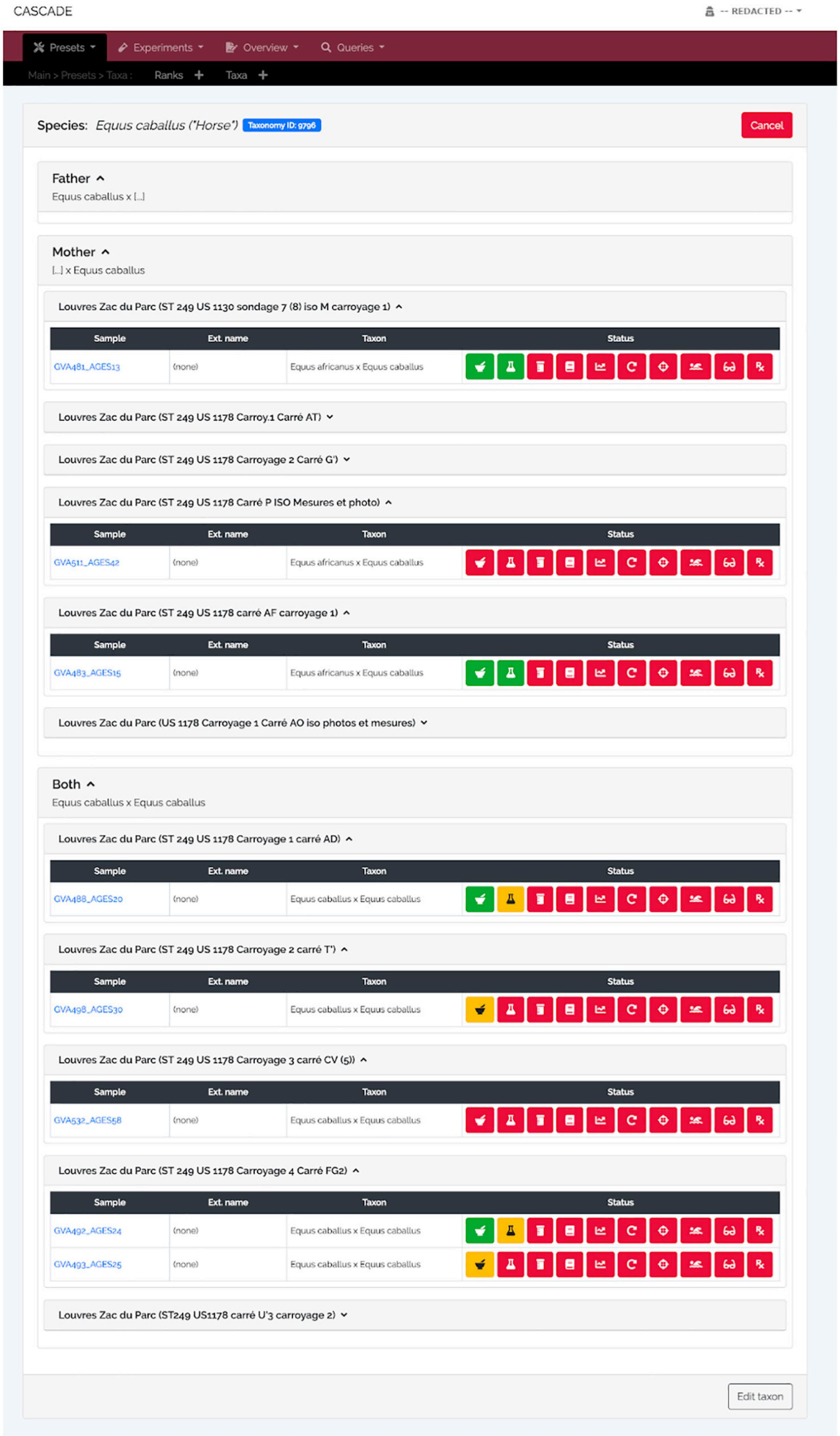
Samples:

-- REDACTED --

| Sample | Ext. name | Taxon | Material | Age | Status |
|---|---|---|---|---|---|
| GVA492_AGES24 | (none) | Equus caballus x Equus caballus | Second premolar | 1650 - 1750 yrs BP | 🟩🟨🟥🟥🟥🟥🟥🟥🟥🟥 |
| GVA493_AGES25 | (none) | Equus caballus x Equus caballus | Second premolar | 1650 - 1750 yrs BP | 🟨🟥🟥🟥🟥🟥🟥🟥🟥🟥 |

US 1178 Carroyage 1 Carré AO iso photos et mesures
*(no coordinates available)*

**FIGURE 4 |** Excavation site detail view. For each site and its sub-sites (e.g., sectors in a larger area, different tombs or burial mounds, etc.) all available samples are listed, grouped by the sample provider. A click on the blue sample name leads to the sample's detail view (see **Figure 3A**) while clicking on the provider name (here shown as "*REDACTED*" for privacy reasons) leads to that of the provider (see **Figure 6**). The displayed status bar works in the same way as described previously (see **Figure 3B**). As an additional feature, should GPS coordinates for a given site be available, clicking on the then blue map icon (gray and inactive if no coordinates are available) will automatically lead to a Google Maps view of that specific location.



Species: *Equus caballus ('Horse')* Taxonomy ID: 9796

Father

Equus caballus x [...]

Mother

[...] x Equus caballus

Louvres Zac du Parc (ST 249 US 1130 sondage 7 (8) iso M carroyage 1)

| Sample | Ext. name | Taxon | Status |
|---|---|---|---|
| GVA481_AGES13 | (none) | Equus africanus x Equus caballus | 🟩🟩🟥🟥🟥🟥🟥🟥🟥🟥 |

Louvres Zac du Parc (ST 249 US 1178 Carroy.1 Carré AT)

Louvres Zac du Parc (ST 249 US 1178 Carroyage 2 Carré G')

Louvres Zac du Parc (ST 249 US 1178 Carré P ISO Mesures et photo)

| Sample | Ext. name | Taxon | Status |
|---|---|---|---|
| GVA511_AGES42 | (none) | Equus africanus x Equus caballus | 🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥 |

Louvres Zac du Parc (ST 249 US 1178 carré AF carroyage 1)

| Sample | Ext. name | Taxon | Status |
|---|---|---|---|
| GVA483_AGES15 | (none) | Equus africanus x Equus caballus | 🟩🟩🟥🟥🟥🟥🟥🟥🟥🟥 |

Louvres Zac du Parc (US 1178 Carroyage 1 Carré AO iso photos et mesures)

Both

Equus caballus x Equus caballus

Louvres Zac du Parc (ST 249 US 1178 Carroyage 1 carré AD)

| Sample | Ext. name | Taxon | Status |
|---|---|---|---|
| GVA488_AGES20 | (none) | Equus caballus x Equus caballus | 🟩🟨🟥🟥🟥🟥🟥🟥🟥🟥 |

Louvres Zac du Parc (ST 249 US 1178 Carroyage 2 carré T')

| Sample | Ext. name | Taxon | Status |
|---|---|---|---|
| GVA498_AGES30 | (none) | Equus caballus x Equus caballus | 🟨🟥🟥🟥🟥🟥🟥🟥🟥🟥 |

Louvres Zac du Parc (ST 249 US 1178 Carroyage 3 carré CV (5))

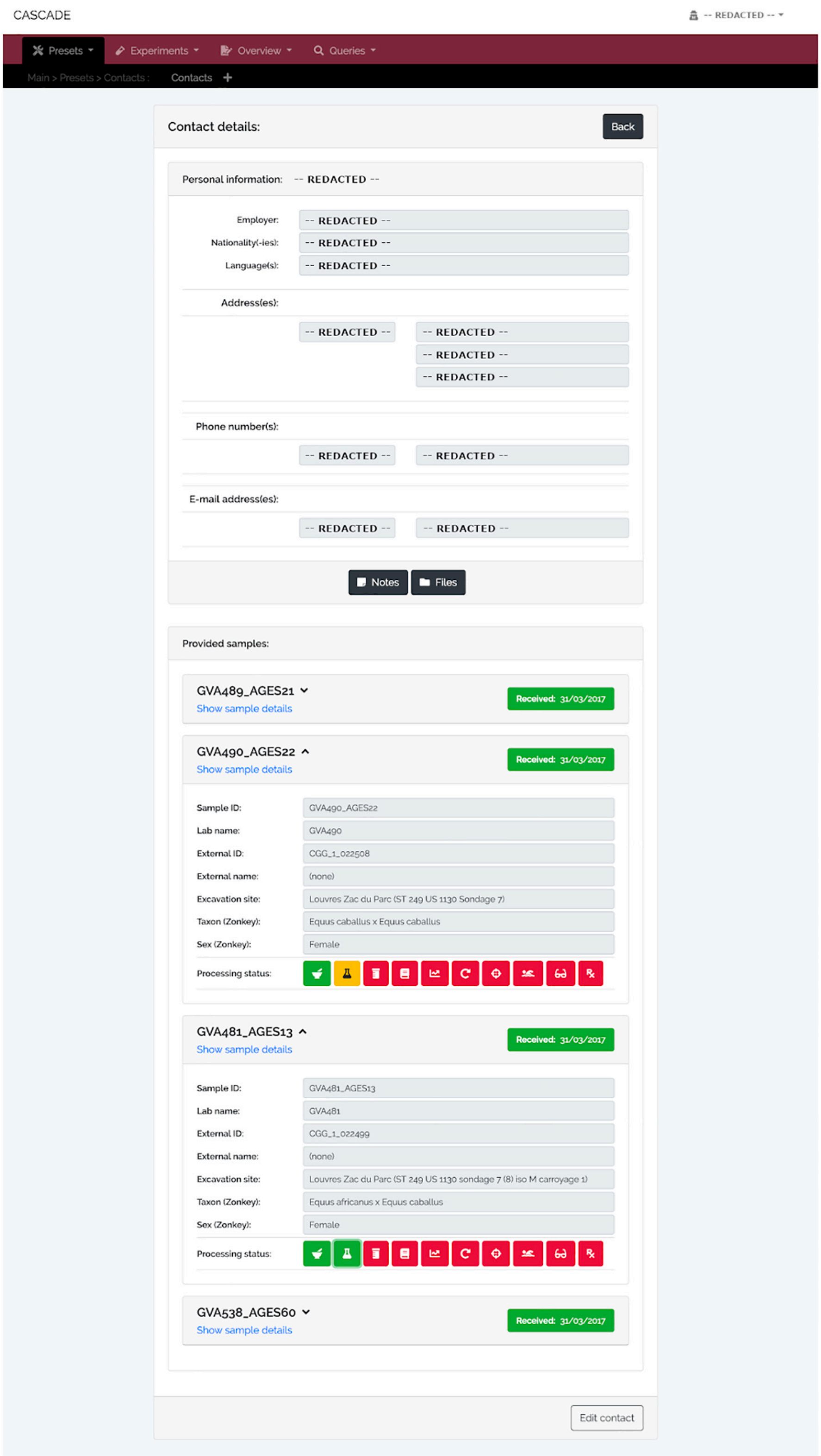
| Sample | Ext. name | Taxon | Status |
|---|---|---|---|
| GVA532_AGES58 | (none) | Equus caballus x Equus caballus | 🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥 |

Louvres Zac du Parc (ST 249 US 1178 Carroyage 4 Carré FG2)

| Sample | Ext. name | Taxon | Status |
|---|---|---|---|
| GVA492_AGES24 | (none) | Equus caballus x Equus caballus | 🟩🟨🟥🟥🟥🟥🟥🟥🟥🟥 |
| GVA493_AGES25 | (none) | Equus caballus x Equus caballus | 🟨🟥🟥🟥🟥🟥🟥🟥🟥🟥 |

Louvres Zac du Parc (ST249 US1178 carré U'3 carroyage 2)

**FIGURE 5 |** Taxon detail view. Similar to the detail view for excavation sites, the taxon detail view lists all samples stored in CASCADE that are associated with the selected taxon. However, only those samples are listed for which there is DNA-based taxon assignment. Samples for which the taxon is estimated on the basis of other criteria (e.g., morphology) will not be listed, as this would rapidly lead to excessively large, thus, impracticable numbers of samples to be displayed. Retrieved samples are grouped by the taxon of the father, mother, and both resulting in the first two groups listing exclusively hybrids (e.g., mules and hinnies). If an NCBI Taxonomy ID was submitted together with the taxon record, a blue button at the top links directly to the entry in the NCBI Taxonomy Browser.



Contact details:

Personal information: -- REDACTED --

| | |
|---|---|
| Employer: | -- REDACTED -- |
| Nationality(-ies): | -- REDACTED -- |
| Language(s): | -- REDACTED -- |

Address(es):

-- REDACTED -- -- REDACTED --
-- REDACTED --
-- REDACTED --

Phone number(s):

-- REDACTED -- -- REDACTED --

E-mail address(es):

-- REDACTED -- -- REDACTED --

Provided samples:

GVA489_AGES21
Show sample details
Received: 31/03/2017

GVA490_AGES22
Show sample details
Received: 31/03/2017

| | |
|---|---|
| Sample ID: | GVA490_AGES22 |
| Lab name: | GVA490 |
| External ID: | CGG_1_022508 |
| External name: | (none) |
| Excavation site: | Louvres Zac du Parc (ST 249 US 1130 Sondage 7) |
| Taxon (Zonkey): | Equus caballus x Equus caballus |
| Sex (Zonkey): | Female |
| Processing status: | |

GVA481_AGES13
Show sample details
Received: 31/03/2017

| | |
|---|---|
| Sample ID: | GVA481_AGES13 |
| Lab name: | GVA481 |
| External ID: | CGG_1_022499 |
| External name: | (none) |
| Excavation site: | Louvres Zac du Parc (ST 249 US 1130 sondage 7 (8) iso M carroyage 1) |
| Taxon (Zonkey): | Equus africanus x Equus caballus |
| Sex (Zonkey): | Female |
| Processing status: | |

GVA538_AGES60
Show sample details
Received: 31/03/2017

**FIGURE 6 |** Contact detail view. This menu displays the contact information for each record stored in CASCADE (here shown as "*REDACTED*" for privacy reasons). It also provides direct access to all samples ever contributed by the example collaborator, as well as their details and status. This property allows evaluating how far samples have progressed through the pipeline which helps this collaborator to assess on demand whether enough data have already been collected to proceed with a project's next step.

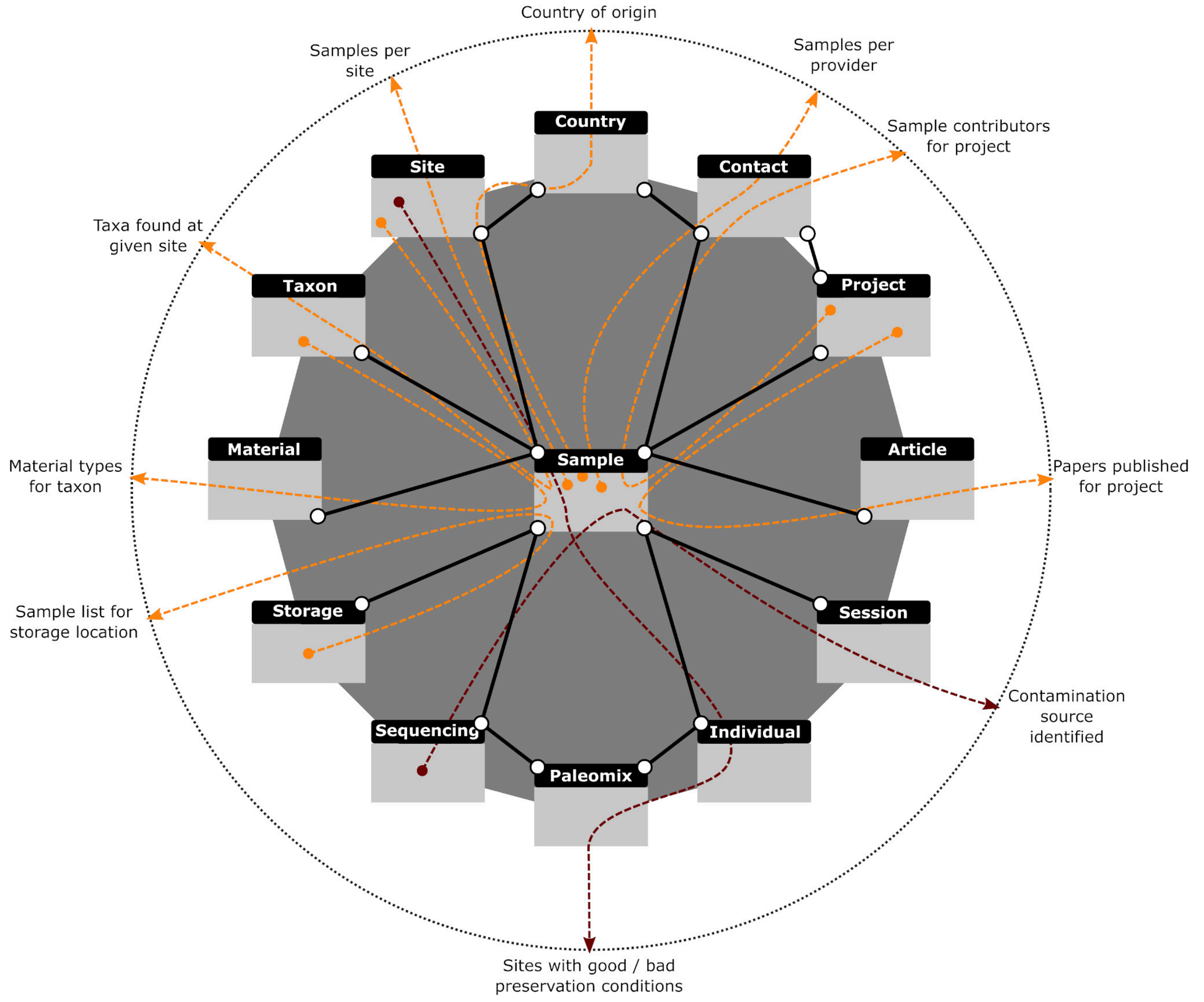


**FIGURE 7 |** Graphical representation of database queries. Colored arrows (red and orange) demonstrate how different paths through the same connected data sets provide answers to various questions (described at dotted circle). All queries in orange are described in the main text and provided with CASCADE. The red queries illustrate other, less generic queries used to answer very specific questions and hence were not pre-built. Gray rectangles with black headers represent the tables dealing with the data described in each header. Solid black lines represent the relations existing between those data sets. A search query created with our interface first combines the requested data sources (e.g., tables or other previously created queries) based on given criteria, filters the combined data according to user specifications, then selects whatever data fields were requested by the user and returns the result in table format.

which, in contrast to their built-in counterparts, provide the same filtering possibilities available in advanced searches. We have provided three such queries in the "*Tables*" section of the basic search feature.

Another possible use of the query feature is the generation of informative labels for bags, bottles, and tubes used at any step of the experimental pipeline handled by CASCADE. To achieve this result, customized queries retrieve exactly the data required for any individual laboratory's labeling needs and are exported as tab-delimited text files that can serve as input for most label-printing software and equipment available at the host laboratory. The data obtained can also be used for the production of one-dimensional and two-dimensional barcodes (e.g., QR codes). Due to their data density, the latter type is especially useful for labeling samples and tubes. This feature allows to store and retrieve all essential information about a specimen by simply scanning the attached barcode with barcode scanner or even a smartphone (in our laboratory we make use of "*Barcode Scanner*" for Android based on the ZXing open source barcode scanning library), without the need for a connection to CASCADE. Barcodes can also be used to directly access all sample information stored in the database. For this, users generate batches of IDs for their own personal use which can be assigned to any experimental data type and enable the recovery of the related data record. As these IDs can be created before any sample information is available, they can be printed in advance as barcodes and attached to individual tubes as experiments make progress. This is especially helpful in cases where no printers or computers are available wherever the different tubes are processed and/or stored (e.g., the ancient DNA clean rooms).

## DISCUSSION

The exponentially growing size of aDNA projects makes it increasingly difficult to keep all experimental metadata fully tractable for laboratory users and their collaborators. By integrating LIMS features tailored to the experimental procedures underlying aDNA analyses, CASCADE not only provides the first solution toward this objective, but also empowers experimental work and collaborative sharing through the possibility of automatic queries providing real-time information about ongoing progress and results. CASCADE can be accessed remotely from a web-browser by any user provided with a protected personal login account. It is made available for free, thereby helping to build the capacity of those smaller laboratories, which cannot afford the purchase of commercial LIMS solutions. It also contributes to the long-term preservation of important experimental information that may prove essential for the integration of available data to future projects, especially as the underlying methodology is constantly evolving.

In order to safeguard all experimental information handled by CASCADE, we have implemented two separate backup mechanisms. The first mechanism is configured during the installation and setup of the VM. It automatically generates full backups on a weekly basis in addition to daily incremental ones that allow a full reset of the database to its last functional state. The second mechanism can be manually triggered from inside CASCADE so as to force data download at critical stages. Like for the first mechanism, the output produced this way can be used to re-initialize the database. More importantly though, it can also be used to initialize a new copy of the database rather than just reset an existing one. In addition, it provides all data in a tab-delimited text format. This attribute guarantees that data stored in CASCADE will always be fully accessible in an easy-to-process format so that it can be transferred should newer and better lab management solutions become available.

## DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

## AUTHOR CONTRIBUTIONS

LO conceived the project and coordinated work. DD and AFa designed the database schema and coordinated the data restructuring process. DD created the virtual servers, figures, and programmed the database. AFa developed the pool feature concept, managed the input of all laboratory members on the database design, and curated the data. AFr developed the tag and sample group feature concepts and performed break testing of the database. AS-O developed the waiting list feature concept. XM tested installation procedures. AFa, SS, LT-C, LC, SW, CD, AFr, and AS-O tested the database. DD, AFa, and LO wrote the manuscript, with input from all co-authors. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported by the Initiative d'Excellence Chaires d'attractivité, Université de Toulouse (OURASI) and the Villum Fonden miGENEPI research project. LO has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 681605-PEGASUS). This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 795916-NEO.

## ACKNOWLEDGMENTS

We thank Dr. Naveed Khan and the staff of the AGES research laboratory (Archeology, Genomics, Evolution and Societies), especially Dr. Morgane Gibert, Dr. Catherine Thèves, and Dr. Tomasz Suchan, for their input on the database design, help with data restructuring, database testing and quality control. We also thank Corentin Deppe for his help in input data management.

## REFERENCES

Boessenkool, S., Hanghøj, K., Nistelberger, H. M., Der Sarkissian, C., Gondek, A. T., Orlando, L., et al. (2017). Combining bleach and mild predigestion improves ancient DNA recovery from bones. *Mol. Ecol. Res.* 17, 742–751. doi: 10.1111/1755-0998.12623

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., et al. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *PNAS* 104, 14616–14621. doi: 10.1073/pnas.0704665104

Brunson, K., and Reich, D. (2019). The promise of paleogenomics beyond our own species. *Trends Genet.* 35, 319–329. doi: 10.1016/j.tig.2019.02.006

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S. T., Sinding, M. H. S., and Samaniego, J. A. (2018). Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* 2, 410–419. doi: 10.1111/2041-210X.12871

Codd, E. F. (1970). A relational model of data for large shared data banks. *Commun. ACM* 13, 377–387. doi: 10.1145/362384.362685

Dabney, J., Knapp, M., Glocke, I., Gansauge, M. T., Weihmann, A., Nickel, B., et al. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15758–15763. doi: 10.1073/pnas.1314445110

Damgaard, P. B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., et al. (2018). 137 ancient human genomes from across the Eurasian steppes. *Nature* 557, 369–374. doi: 10.1038/s41586-018-0094-2

Damgaard, P. B., Margaryan, A., Schroeder, H., Orlando, L., Willerslev, E., and Allentoft, M. (2015). Improving access to endogenous DNA in ancient bones and teeth. *Sci. Rep.* 5:11184. doi: 10.1038/srep11184

Date, C. J. (2003). *An Introduction to Database Systems*, 8th Edn. London: Pearson.

Fages, A., Hanghøj, K., Khan, N., Gaunitz, C., Seguin-Orlando, A., Leonardi, M., et al. (2019). Tracking five millennia of horse management with extensive ancient genome time series. *Cell* 177, 1419–1435. doi: 10.1016/j.cell.2019.03.049

Gansauge, M. T., Gerber, T., Glocke, I., Korlevic, P., Lippik, L., Nagel, S., et al. (2017). Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res.* 45:e79. doi: 10.1093/nar/gkx033

Gansauge, M. T., and Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc.* 8, 737–748. doi: 10.1038/nprot.2013.038

Gansauge, M. T., and Meyer, M. (2014). Selective enrichment of damaged DNA molecules for ancient genome sequencing. *Genome Res.* 24, 1543–1549. doi: 10.1101/gr.174201.114

Glocke, I., and Meyer, M. (2017). Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. *Genome Res.* 27, 1230–1237. doi: 10.1101/gr.219675.116

Goodwin, S., McPherson, J., and McCombie, W. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi: 10.1038/nrg.2016.49

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., et al. (2010). A Draft Sequence of the Neandertal Genome. *Science* 5979, 710–722. doi: 10.1126/science.1188021

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522, 207–211. doi: 10.1038/nature14317

Harney, É, May, H., Shalem, D., Mallick, S., Rohland, N., Lazaridis, I., et al. (2018). Ancient DNA from Chalcolithic Israel reveals the role of population mixture in cultural transformation. *Nat. Commun.* 9:3336. doi: 10.1038/s41467-018-05649-9

Kistler, L., Maezumi, S. Y., Gregorio de Souza, J., Przelomska, N. A. S., Malaquias Costa, F., Smith, O., et al. (2018). Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. *Science* 362, 1309–1313. doi: 10.1126/science.aav0207

Korlević, P., and Meyer, M. (2019). Pretreatment: removing DNA contamination from ancient bones and teeth using sodium hypochlorite and phosphate. *Methods Mol. Biol.* 1963, 15–19. doi: 10.1007/978-1-4939-9176-1_2

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Mann, A. E., Sabin, S., Ziesemer, K., Vågene, ÅJ., Schroeder, H., Ozga, A. T., et al. (2018). Differential preservation of endogenous human and microbial DNA in dental calculus and dentin. *Sci. Rep.* 8:9822. doi: 10.1038/s41598-018-28091-9

Marciniak, S., and Perry, G. H. (2017). Harnessing ancient genomes to study the history of human adaptation. *Nat. Rev. Genet.* 18, 659–674. doi: 10.1038/nrg.2017.65

Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., Szécsényi-Nagy, A., Rohland, N., Mallick, S., et al. (2018). The genomic history of southeastern Europe. *Nature* 555, 197–203. doi: 10.1038/nature25778

Metzker, M. (2010). Sequencing technologies – the next generation. *Nat. Rev. Genet.* 11, 31–46. doi: 10.1038/nrg2626

Olalde, I., Brace, S., Allentoft, M. E., Armit, I., Kristiansen, K., Booth, T., et al. (2018). The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* 555, 190–196. doi: 10.1038/nature25738

Olalde, I., Mallick, S., Patterson, N., Rohland, N., Villalba-Mouco, V., Silva, M., et al. (2019). The genomic history of the Iberian Peninsula over the past 8000 years. *Science* 363, 1230–1234. doi: 10.1126/science.aav4040

Pedersen, M. W., Overballe-Petersen, S., Ermini, L., Sarkissian, C. D., Haile, J., Hellstrom, M., et al. (2015). Ancient and modern environmental DNA. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370:20130383. doi: 10.1098/rstb.2013.0383

Pinhasi, R., Fernandes, D., Sirak, K., Novak, M., Conell, S., Alpaslan-Roodenberg, S., et al. (2015). Optimal Ancient DNA yields from the inner ear part of the human petrous bone. *PLoS One* 10:e0129102. doi: 10.1371/journal.pone.0129102

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., et al. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* 463, 757–762. doi: 10.1038/nature08835

Regalado, A. (2019). *More Than 26 Million People Have Taken an at-Home Ancestry Test. MIT Technology Review*. Available online at: https://www.technologyreview.com/2019/02/11/103446/ (accessed December 2, 2019).

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., et al. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468, 1053–1060. doi: 10.1038/nature09710

Rohland, N., Glocke, I., Aximu-Petri, A., and Meyer, M. (2018). Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. *Nat. protoc.* 13, 2447–2461. doi: 10.1038/s41596-018-0050-5

Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., and Reich, D. (2015). Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370:20130624. doi: 10.1098/rstb.2013.0624

Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. protoc.* 9:1056. doi: 10.1038/nprot.2014.063

Spyrou, M. A., Bos, K. I., Herbig, A., and Krause, J. (2019). Ancient pathogen genomics as an emerging tool for infectious disease research. *Nat. Rev. Genet.* 20, 323–340. doi: 10.1038/s41576-019-0119-1

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science* 291, 1304–1351. doi: 10.1126/science.1058040

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 07 July 2020
doi: 10.3389/fevo.2020.00217

# Insights Into Aboriginal Australian Mortuary Practices: Perspectives From Ancient DNA

*Sally Wasef[1]*, Joanne L. Wright[1], Shaun Adams[1], Michael C. Westaway[1,2], Clarence Flinders[3], Eske Willerslev[4,5,6] and David Lambert[1]**

*[1] Australian Research Centre for Human Evolution, Environmental Futures Research Institute, Griffith University, Nathan, QLD, Australia, [2] Archaeology, School of Social Science, The University of Queensland, St Lucia, QLD, Australia, [3] Cape Melville, Flinders and Howick Islands Aboriginal Corporation, Cairns, QLD, Australia, [4] Department of Zoology, University of Cambridge, Cambridge, United Kingdom, [5] Wellcome Trust Sanger Institute, Cambridge, United Kingdom, [6] Lundbeck Foundation GeoGenetics Centre, University of Copenhagen, Copenhagen, Denmark*

OPEN ACCESS

***Edited by:***
*Nic Rawlence,*
*University of Otago, New Zealand*

***Reviewed by:***
*Raymond Tobler,*
*The University of Adelaide, Australia*
*Wolfgang Haak,*
*Max Planck Institute for the Science of Human History (MPI-SHH), Germany*

****Correspondence:***
*Sally Wasef*
*s.wasef@griffith.edu.au*
*David Lambert*
*d.lambert@griffith.edu.au*

***Specialty section:***
*This article was submitted to Paleoecology, a section of the journal Frontiers in Ecology and Evolution*

***Received:*** *30 January 2020*
***Accepted:*** *11 June 2020*
***Published:*** *07 July 2020*

***Citation:***
*Wasef S, Wright JL, Adams S, Westaway MC, Flinders C, Willerslev E and Lambert D (2020) Insights Into Aboriginal Australian Mortuary Practices: Perspectives From Ancient DNA. Front. Ecol. Evol. 8:217. doi: 10.3389/fevo.2020.00217*

Paleogenetics is a relatively new and promising field that has the potential to provide new information about past Indigenous social systems, including insights into the complexity of burial practices. We present results of the first ancient DNA (aDNA) investigation into traditional mortuary practices among Australian Aboriginal people with a focus on North-East Australia. We recovered mitochondrial and Y chromosome sequences from five ancestral Aboriginal Australian remains that were excavated from the Flinders Island group in Cape York, Queensland. Two of these individuals were sampled from disturbed beach burials, while the other three were from bundle burials located in rock shelters. Genomic analyses showed that individuals from all three rock shelter burials and one of the two beach burials had a close genealogical relationship to contemporary individuals from communities from Cape York. In contrast the remaining male individual, found buried on the beach, had a mitochondrial DNA sequence that suggested that he was not from this location but that he was closely related to people from central Queensland or New South Wales. In addition, this individual was associated with a distinctive burial practice to the other four people. It has been suggested that traditionally non-locals or lower status individuals were buried on beaches. Our findings suggest that theories put forward about beach burials being non-local, or less esteemed members of the community, can potentially be resolved through analyses of uniparental genomic data. Generally, these results support the suggestion often derived from ethnohistoric accounts that inequality in Indigenous Australian mortuary practices might be based on the status, sex, and/or age of individuals and may instead relate to place of geographic origin. There is, however, some departure from the traditional ethnohistoric account in that complex mortuary internments were also offered to female individuals of the community, with genomic analyses helping to confirm that the gender of one of the rockshelter internments was that of a young female.

**Keywords: Aboriginal Australians, bioarchaeology, genomic enrichment, mitochondrial DNA, paleogenetics, ancient DNA**

## INTRODUCTION

A better understanding of how people lived in the past can be revealed by an examination of their skeletal remains, which can assist in reconstructing their life history. The examination of their burial context reveals information on how they were treated after death. Traditionally this level of understanding has been reconstructed using methods from biological anthropology. These methods can include studies of craniometrics and biological traits that can provide insights into the genetic relatedness of an individual Pardoe (1993). Moreover, the examination of the health status of an individual which provides insights into how the individual may have been nursed or cared for by that society during life (Tilley, 2015). Taphonomic investigations of burial sites and the arrangement of the deceased in the grave provide important information on the status of an individual (Pearson, 1999). However, the methods based on archeological and anthropological assessments are sometimes insufficient to establish the sex and biological kinship relationships particularly when ancestral remains are heavily eroded or damaged and when comparative datasets simply do not exist for multivariate analyses of metric and non-metric data. The uses of ancient DNA (aDNA) can provide more precise information about the biological affinity among individuals in past populations to complement the other bioarcheological findings. The field of ancient DNA has expanded rapidly since its inception in 1984 (Kutanan et al., 2017). It has now impacted a large number of different disciplines including biological anthropology. Ancient DNA studies have the potential to provide new information about cultural traditions and specifically burial practices.

Biparentally inherited genomic data has been successively used in population studies for repatriation purposes (Heupink et al., 2016; Malaspinas et al., 2016). However, due to the highly fragmented nature of aDNA, this whole-genome approach could be inefficient when studying the burial practices of poorly preserved remains (Collard et al., 2019). It is typically more feasible to recover high copy number uniparental mitochondrial DNA (mtDNA) from human remains (Wright et al., 2018).

The Flinders Island group is located in the tropical north-east of Queensland, off the eastern coast of Cape York Peninsula. The group consists of seven continental Islands, sitting within the Princess Charlotte Bay, west from Cape Melville (**Figure 1**). Over the past two centuries, Indigenous people from these Islands have been regularly involved in maritime industries. Cultural changes that have included increased European contact have also resulted in the removal and theft of many Indigenous cave bundle burials (Horsfall, 1991). Not only were ancient remains affected by these activities, but so too were the Indigenous people of Flinders themselves. Since the Second World War many Indigenous people were forcefully removed from Flinders Island (Wurrima) to the mainland as a result of government regulations at that time (Rigsby and Chase, 1998).

Princess Charlotte Bay was first recorded by British navigators in 1815 (Jack, 1921). Princess Charlotte Bay was situated within the early shipping routes of the east coast of Australia, allowing early contact between Europeans and Traditional Owners in the 19th Century (King, 1827; Coppinger, 1883; Roth, 1898). Although there has been extensive archeological research conducted in the Flinders Group, no research to date has extended to mortuary practices. Little attention has been paid to many of the burial sites on the islands, or indeed throughout many parts of Cape York. Archeological examinations carried out by Beaton (1985) revealed that there has been an Aboriginal occupation of the Flinders group for at least 2,300 years. Beaton (1985) also suggested that the initial occupation of Flinders Island occurred ∼2,500 years ago and was probably closely related to the introduction of Papuan/Melanesian outrigger canoes. This interpretation is now under question with the discovery of archeological evidence at Endean Shelter illustrating that Aboriginal occupation of the islands extended back to 6,280 calBP (Collard et al., 2019). Ethnographic records provide detailed information on mortuary protocols after initial contact, but little is known about how these customs may have differed in pre-European contact.

Knowledge of Aboriginal mortuary practices in north-east Australia is limited to ethnographic accounts by anthropologists and observers in the 19th and early 20th Centuries (Roth, 1898, 1907). According to Roth, people of Cape York interpreted death as a result of spiritual intervention or human agency, rather than natural phenomena. It was believed that spirits of the dead could harm the living (Roth, 1907). The deaths of prominent and/or powerful people were often avenged by their remains being carried from camp to camp and defleshed before finally being buried or interred in trees or caves. While old, less esteemed, or infirmed people were given simpler burials with minimal ceremony and often buried within close proximity to the site of death (Roth, 1907). Also, at Torilla, south of Princess Charlotte Bay, Roth observed that women were usually buried immediately after death, bundled in bark and carried from camp to camp (Roth, 1907). In this study, we propose a mortuary narrative constructed from genomic data recovered from five ancient individuals who were interred in the Flinders Island Group.

We suggest that our understanding of the mortuary practices of Aboriginal Australians can be improved using ancient DNA methods. We investigate the possible link between mortuary practices and kinship among five individuals excavated within the Flinders Island Group in 2015–2016 (Adams et al., unpublished). Three burials were located within rock-shelters and two on beaches. The two contrasting sites pose an interesting question about the kinship among individuals, and at different sites. To gain a direct insight into the kinship relationships we used complete mitochondrial genome sequences and Y chromosome data, whenever it was available. Although we obtained uniparental genomic data by applying whole-genome target enrichment coupled with Next Generation Sequencing (NGS) methods, the autosomal data is not the focus of this publication.

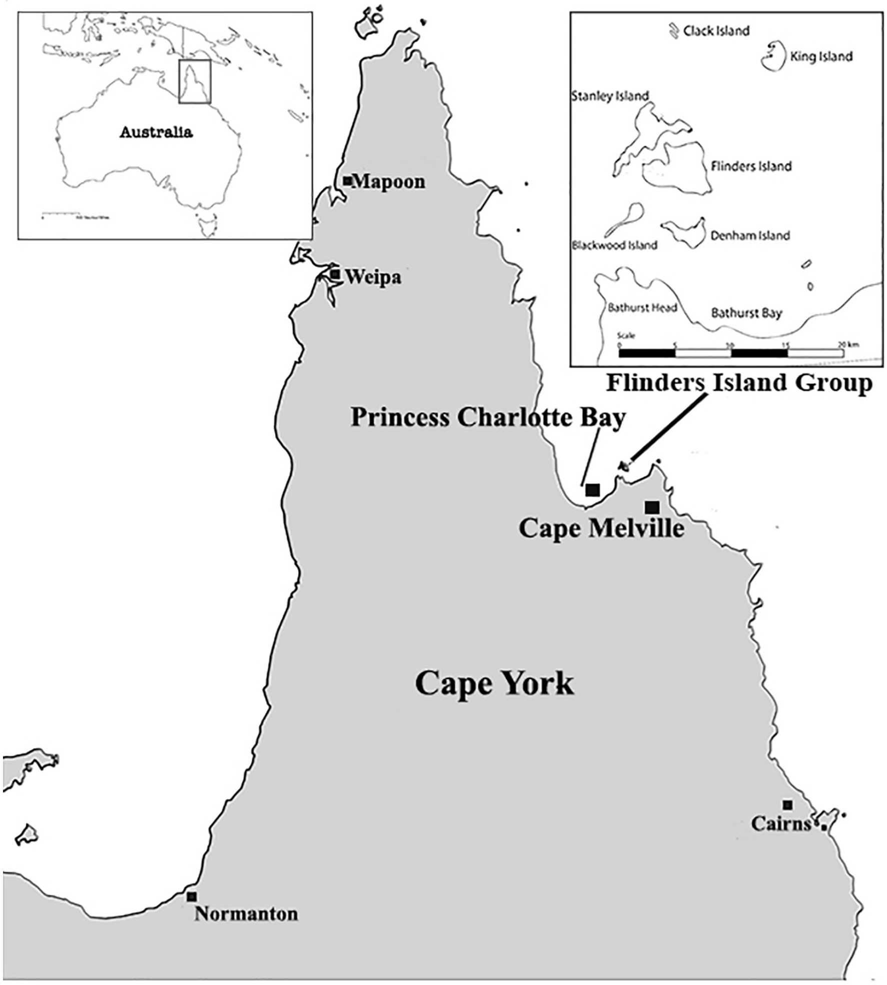


**FIGURE 1 |** Map of Cape York, showing the Flinders Island group location in the tropical north-east of Queensland off the eastern coast of Cape York.

# MATERIALS AND METHODS

## Archeological Samples

Archeological fieldwork included rescue excavations of two disturbed beach burials and recording of three bundle burials from two interment caves in the Flinders Islands (Adams et al., unpublished). The research focused on investigating burial sites and recovering samples for paleogenetic and isotope research. Orientation, shape, and size of each burial were recorded. Anatomical measurements were used for preliminary determination of sex, age, ancestry, and pathologies (Adams et al., unpublished). All observation and recording of available skeletal elements were completed in the field following excavation (Adams et al., unpublished). Immediately upon completion of recording and sampling Traditional Owners reinterred all excavated elements at a safe location. With consent from Traditional Owners, tooth and bone samples were collected for radiometric dating, isotopic assessment and aDNA analyses.

### Details of the Individuals Studied

*Flinders Island individual 1 (FLI1)*

This adult male was discovered eroding from a beach burial in an area of Flinders Island known as Apa Spit (Hale and Tindale, 1933) or Wathirrmana (Sutton et al., 1993) (**Figure 2B**). The remains were initially excavated by Traditional Owners and Queensland Police, who determine that it was a traditional burial, at a depth of ca. 1.2 m below the original ground level (Adams et al., unpublished). The individual has a north-east orientation with their face directed east. They had been interred on their back with their legs partially flexed and a large (ca. 40 cm diameter) limestone rock placed on his chest. The individual's hands were placed palm down on the thighs, and their feet were crossed. The rock, which was removed during the preliminary investigation, was the only identifiable grave good (Adams et al., unpublished).

*Stanley Island individual (STI1)*

In 2015, this female was discovered eroding from the beach foredune sands on Stanley Island by two fishermen, who removed the crania for a photo before re-burying the remains (Adams et al., unpublished) (**Figure 2A**). The burial was located in a large (ca. 1 km wide), flat, sandy cove that is surrounded by limestone boulders. The individual had been buried facing south-east. Wear of the frontal and supraorbital ridges of the cranium indicating a long-time exposure of the remains.

#### *Flinders Island individual 2 (FLI2)*

A set of remains found inside a rock-shelter located close to the beach on the east coast of Flinders Island (**Figure 2C**). The rock-shelter faces east and is ca. 10 m wide and ca. 2 m deep. Graham Walsh, in his early survey of the islands in the 1980s, recorded two sets of remains in the rock-shelter (Walsh, 1985) suggesting that they were bundle burials. Only one set of these was found belonging to an adult male (FLI2), to be still existing in the rock-shelter which was partially covered as a result of heavy weathering of the roof.

#### *Flinders Island individuals [FLI3 (B2) and FLI4 (B3)]*

The second rock-shelter is located on the north coast of Flinders Island. It faces north-west and is ca. 30 m long, ca. 6 m deep, and up to 1.5 m high. Again, the rock-shelter was surveyed by Walsh who recorded the presence of six bundle burials involving ornate bark coffins and matting. In 2016, only two of the burials remained Walsh's (Walsh, 1985) Burial 2 (B2) of a young male and Burial 3 (B3) of an early 20's female (**Figures 2D,E**). Although the cylinder-coffins and skeletal remains associated with four of the six burials were missing, the outer paperbark wrapped around the cylinders remained either *in situ* or nearby, except for Burial 6 which had no evidence of discarded paperbark. One cylinder-coffin remained with FLI4 (B3). It had been opened but retained fine twine that likely bound the post-cranial skeletal elements.

## Radiocarbon Dating

Human bone collagen from each of the five sets of remains was directly dated using AMS radiocarbon methods. This procedure was conducted at the Research School of Earth Sciences, Australian National University, Canberra. Dates were calibrated using OxCal 4.2 (Bronk Ramsey, 2013) software and the Southern hemisphere calibration curve SHCal13 (Hogg et al., 2013) (**Supplementary Table S3**).

## Ancient DNA

Ancient DNA work was carried out in a dedicated facility in the Australian Research Centre for Human Evolution at Griffith University. Only the roots of teeth were used for the ancient DNA extraction of the five individuals. Methods for handling ancient DNA were as outlined by Knapp et al. (2012).

Each sample was initially decontaminated with 10% sodium hypochlorite for 10 min, followed by 80% ethanol, and 5 min under UV light. Subsequently, the skeletal material was processed using a Dremel rotary tool with a high-speed diamond cutter head, or manually with a sterilized scalpel blade. DNA was extracted from ∼50 mg of bone or tooth powder following the modified protocol outlined in Wright et al. (2018), which allowed for better recovery of shorter DNA fragments (∼30 bp). Negative controls were included throughout all procedures, each of which showed no contamination.

### DNA Library Construction Methods

Double-stranded Illumina DNA libraries were built according to the modified method of Meyer and Kircher (2010) as detailed in Wright et al. (2018). Using the NEBNext DNA Library Prep Master Mix Set for 454 (New England Biolabs ref: E6070) 21.25 μl of DNA extract was subjected to three consecutive steps: NEBNext end repair, NEBNext blunt end adaptor ligation, followed by an Adapter Fill-In reaction. A MinElute (Qiagen) purification step with 10× binding buffer PB (Qiagen) was carried out between the first and second steps.

All pre-PCR libraries were amplified using KAPA HiFi Hotstart Uracil + (Kapa Biosystems), according to manufacturer's instructions using Illumina single indexing primers. PCR amplification products were cleaned using 1× Axygen beads according to the manufacturer's instructions. Amplified libraries, including negative controls, were quantified and visualized for length distribution using the 5,000 High-Sensitivity DNA tapes on the TapeStation 4000 (Agilent Technologies), following the manufacturer's instructions.

### Whole-Genome In-Solution Target Capture

Between 100 and 500 ng of library amplified DNA was subjected to in-solution target enrichment using whole human genome myBaits WGE (Arbor Biosciences) as detailed in Wasef et al. (2018) and Wright et al. (2018). Target capture enrichment steps were performed according to manufacturer's instructions with the following modifications: the hybridization step was performed for 36–42 h at 57°C. The beads, and the bead binding buffers were heated to 57°C for 30 min before being used. Further cleaning steps were also performed at the same hybridization temperature. Post-capture libraries were amplified on beads using HiFi HotStart Uracil + ReadyMix (Kapa Biosystems) for between 14 and 17 cycles, and then visualized using the 5,000 High-Sensitivity DNA tapes on the TapeStation 4000 (Agilent Technologies).

### Ancient Sequencing

After target enrichment, ancient samples were sequenced on HiSeq 4000 Sequencing System (Illumina) at The Danish National High-Throughput DNA Sequencing Centre in Copenhagen. Sequences were base called using CASAVA 1.8.2 (Illumina), demultiplexed and FASTQ files were generated by the sequencing facility.

## Modern Aboriginal Genomes

In addition to the previously published mitochondrial genomes of Indigenous Australians (van Holst Pellekaan et al., 2006; Hudjashov et al., 2007; Rasmussen et al., 2011; Nagle et al., 2017; Tobler et al., 2017), we also incorporated the haplogroup data previously published in Malaspinas et al. (2016) and Wright et al. (2018) (**Supplementary Table S2**).

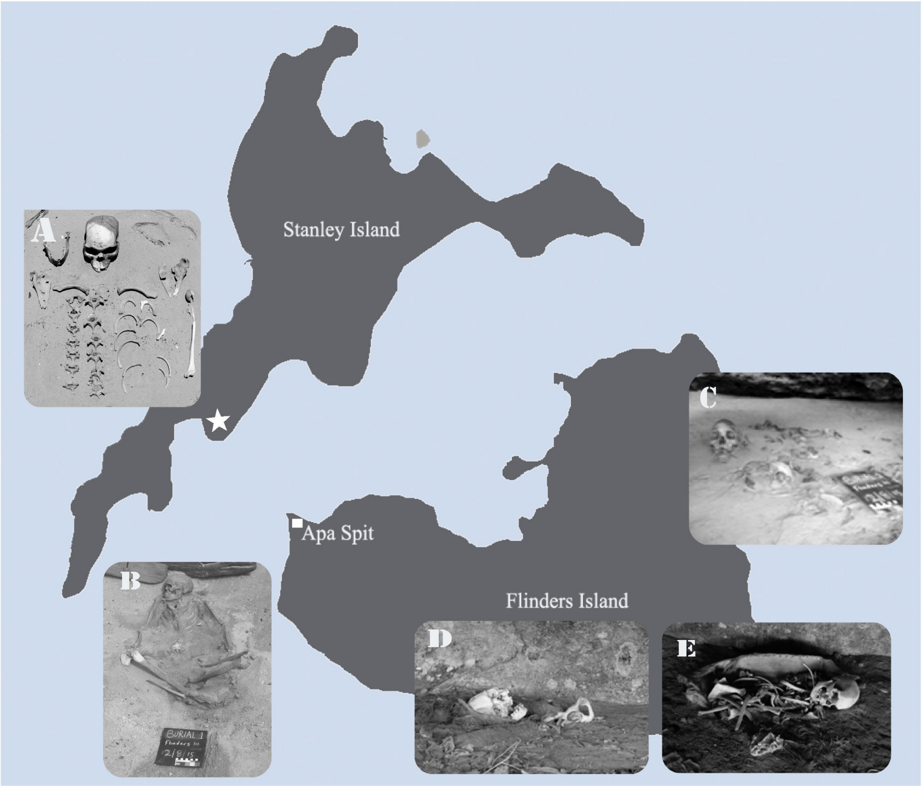


**FIGURE 2 |** Flinders Island Group burials. **(A)** Stanley Island female (STI1) remains, the burial location is indicated by the hexagon on the map. **(B)** Flinders Island adult male individual (FLI1) beach burial in the area of Apa Spit. **(C)** A set of remains found inside a rock-shelter (FLI2), located close to the beach on the east coast of Flinders Island. The second rock-shelter in the north coast of Flinders island where **(D)** FLI3 and **(E)** FLI4 both were buried.

### The Relationship Between This Study and the Griffith University Human Ethics Approval 2015 / 904

During the writing of this manuscript, and as a result of discussions with the Griffith University Human Ethics Committee (GUHEC), it was decided that the results presented below should not include any previously published genome sequences from contemporary Aboriginal Australians covered by the Griffith University Human Ethics approval 2015 / 904. This decision was made so that the current team had no unfair advantage over other researchers. Hence, we revised earlier analyses that included modern genomes and instead used only published haplotype data that can be found, for example in the **Supplementary Material** of Malaspinas et al. (2016), Wright et al. (2018).

As a result of these decisions, we used the mitochondrial genome data from 34 sets of ancient remains. The latter are not covered by the Griffith University Human Ethics approval. By using only published data, we showed that this modified approach did not significantly affect our conclusions.

## Genome Analyses

Adapter sequences were trimmed using fastx_clipper, part of Fastx_Toolkit (2009) 0.0.13[1], with reads shorter than 30 bases and low-quality bases removed using parameters -Q 33 –l 30. Reads were aligned to the human reference build GRCh37/hg19 for the nuclear genome or to the revised Cambridge Reference Sequence (rCRS) (accession number NC_012920.1). For mitochondrial data BWA 0.6.2-r126 software was used to align sequences (Li et al., 2010) with the following options: seed disabled (Schubert et al., 2012) and terminal low-quality trimming (using parameter -q15). Duplicate reads were removed using the MarkDuplicates tool from the Picard 1.68 tools package[2]. The mapped reads were sorted, indexed and merged using SAMtools 0.1.18 (Li and Durbin, 2009, 2011). The consensus mitogenome was generated using the SAMtools bcftools view –cg -command and converted to FASTA via SAMtools/bcftools/vcfutils (Li and Durbin, 2009). Qualimap was used to estimate the levels of coverage (Okonechnikov et al., 2015) and the number of mapped reads to the human reference genome (GRCh37/hg19).

Ancient DNA sequences were authenticated using MapDamage software (Jonsson et al., 2013), which uses levels of cytosine to thymine misincorporations in the 5′ end of fragments, and guanine to adenine misincorporations in the 3′ end. Schmutzi software was used to estimate modern

[1] http://hannonlab.cshl.edu/fastx_toolkit/

[2] http://broadinstitute.github.io/picard/

human contamination levels in the ancient mitochondrial sequences using the contDeam command which estimates contamination levels using deamination patterns (Renaud et al., 2015). Endogenous consensus sequences were generated using default settings. Both the Schmutzi generated consensus sequences and the original ancient sequences were then manually checked using the SAMtools tview command (Li and Durbin, 2009). Missing sites were replaced with "N". ANGSD was also used to estimate modern contamination levels in male samples.

Mitochondrial haplotypes were identified using HaploGrep 2.0. A total of 229 ancient and modern mitochondrial genomes (as detailed in **Supplementary Table S2**) were realigned using the online version of MAFFT software (Katoh et al., 2017). The mitochondrial consensus sequences were used to construct a Maximum Likelihood phylogenetic tree using the online version of RAXML with 1,000 bootstrap replications (Kozlov et al., 2019). The resulting Likelihood tree provided information about the maternal ancestry of each of the Flinders Island Group individuals (**Figure 3**).

Summary statistics of haplogroup frequencies in Queensland were estimated using 144 mitogenomes summarized in **Supplementary Table S4**. Arlequin software V3.5.2.2 (Excoffier and Lischer, 2010) was used to estimate haplotype frequencies and genetic distances (Fst) as pair-wise values, and to perform analysis of molecular variance by means of AMOVA (**Supplementary Tables S4**, **S5**). Using the SPSS V26.0.0.1 software package, Principal Component Analysis (PCA) was performed on the haplogroup frequencies detected in the Queensland populations investigated and in previously studied populations (**Supplementary Figure S1A**). Fst distance matrices of mtDNA haplotypes were used to construct MDS plots (**Supplementary Figure S1B**). Median-joining networks of haplogroups without pre- and post-processing steps were performed using Network[3] (**Supplementary Figure S2**).

Sex determination of all ancient Aboriginal Australian individuals was inferred using the method outlined in Skoglund et al. (2013). Y chromosome haplogroup assignments were performed for male individuals using Yleaf software (Ralf et al., 2018).

# RESULTS

## Radiocarbon Dating

**Supplementary Tables S1**, **S3** include the AMS $^{14}$C dates conducted on the bone collagen of each of the five individuals. The remains were dated between 147 and 473 calBP (**Supplementary Table S3**). These results show that all individuals recorded on both Flinders and Stanley Islands died before European colonization, making them suitable for comparison with the ethnographic mortuary record.

## Ancient DNA

We successfully recovered complete ancient mitogenomes from five ancient Flinders Group individuals, in addition to 29 we published previously (Wright et al., 2018), ranging between 2.3 and 331.9× coverage (**Supplementary Table S2**). All ancient DNA recovered were authenticated, and modern-day contamination levels were estimated (**Supplementary Table S1**). The characteristic aDNA damage patterns were estimated for each sample using MapDamage software (Jonsson et al., 2013). All samples exhibited damage patterns characteristic of ancient DNA, with elevated levels of cytosine to thymine misincorporations in the 5′ end of fragments, and guanine to adenine misincorporations in the 3′ end (Dabney et al., 2013). The mean read length also indicated that the DNA recovered was likely authentic (**Supplementary Table S1**). Recovered sequences were also consistent with Aboriginal Australian mitochondrial haplotypes and did not match those carried by any of the ancient DNA laboratory members.

Sex determination of the five individuals was determined bioinformatically, using the method detailed in Skoglund et al. (2013) and confirmed that FLI1, FLI2, and FLI4 were males, while STI1 and FLI3 were females.

## Genome Analyses

After constructing a Maximum Likelihood phylogenetic tree using 229 ancient, historical and modern mitochondrial genomes (**Supplementary Tables S1**, **S2** and **Figure 3**), it became clear that the five ancient individuals fell within previously described mitochondrial haplotypes of contemporary and ancient Aboriginal Australians (van Holst Pellekaan et al., 2006; Hudjashov et al., 2007; Rasmussen et al., 2011; Malaspinas et al., 2016; Tobler et al., 2017; Wright et al., 2018). All five mitochondrial genomes were Aboriginal Australians in origin. In the network generated, FLI1 clustered with two individuals from Central and South Queensland, both carrying the S* haplogroup (**Supplementary Figure S2**).

One hundred and twelve unique haplotypes were present among the 124 mitochondrial genomes from Queensland, showing high haplotype diversity (Hd = 0.9979). When comparing the haplotypes within Queensland, the Flinders Island group ancient samples showed a high mtDNA diversity (Hd = 1.000). Analysis of molecular variance based on haplogroup frequencies demonstrating the variation among different QLD populations are summarized in **Supplementary Table S5**.

FLI2 and STI1 have the recently identified Aboriginal mitochondrial haplotype P5b1 (Wright et al., 2018), which is between 12,000 and 28,000 years in age and appears to be restricted to Australia. Ancient Aboriginal Australians FLI2, STI1, NORA1 from Normanton, and the previously published A422, a contemporary individual from Queensland, all carry the P5b1 mitochondrial haplotype (**Figure 3A**). These four mitogenomes are the only Aboriginal Australians included in this research that carry the P5b1 haplotype. Ancient individual FLI4, found in the same rock-shelter as FLI3, carries the P5a1a haplotype, which is also present in contemporary

[3] http://www.fluxus-engineering.com/

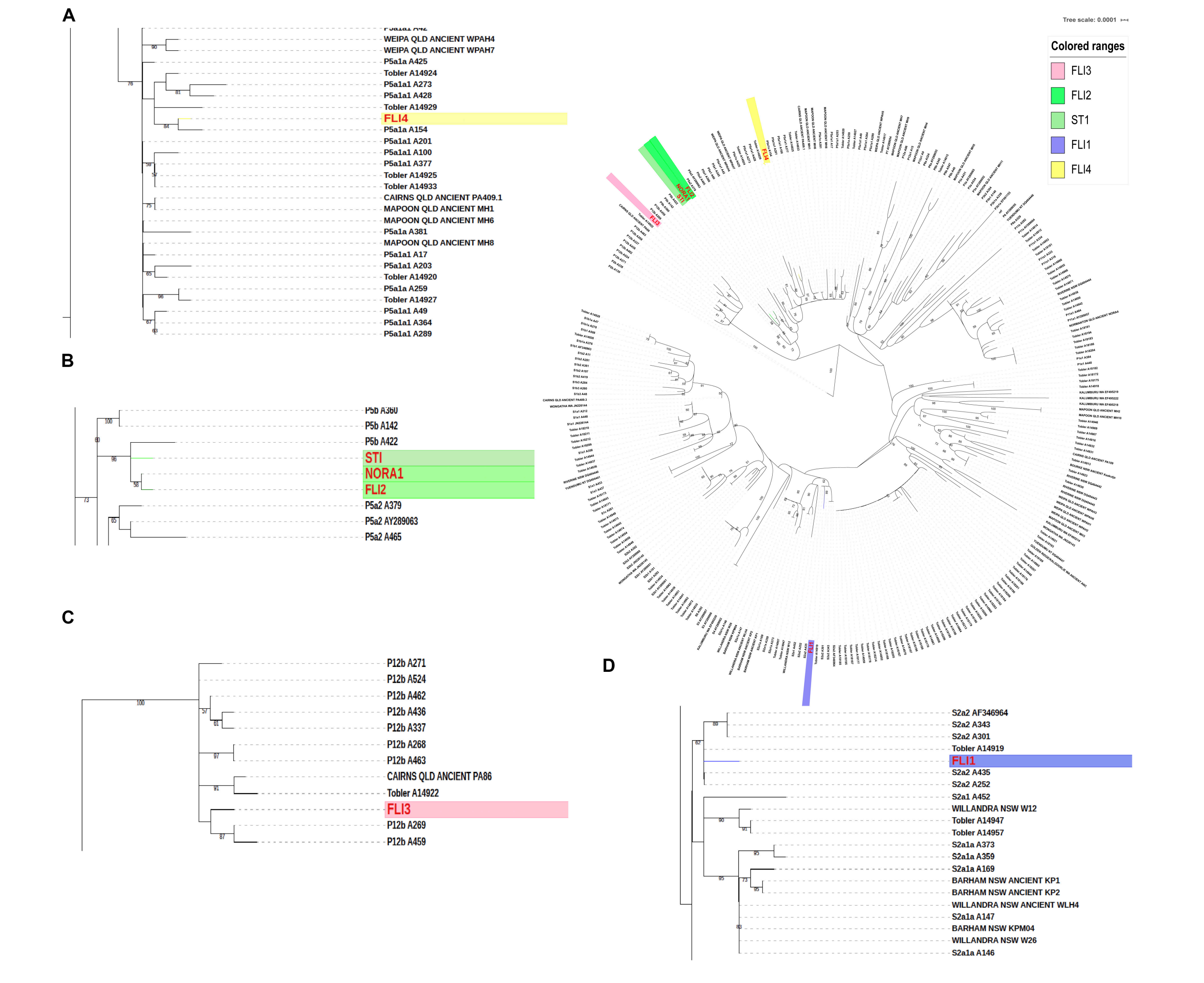


**FIGURE 3 |** Mitochondrial maximum likelihood phylogenetic tree. **(A)** FLI4 belongs to the P5a1a haplotype. **(B)** FLI2 and STI1 both belong to mitochondrial haplotype P5b1. **(C)** FLI3 belongs to the P12b haplotype. **(D)** The FLI1 showed a S2a haplotype.

Aboriginal Australians from Queensland (**Figure 3A**). FLI3 carries the mitochondrial haplotype P12b, also carried by 20 other individuals from Queensland. P12b representing the highest observed haplotype in Queensland with 13.4% (**Supplementary Table S4** and **Supplementary Figure S3**).

Unexpectedly, ancient individual FLI1 carries a S2a mitochondrial haplotype. This haplotype indicates a maternal ancestor for that individual who was not from the Flinders Island group or any close mainland community, but rather this haplotype is more closely related to haplotypes found in New South Wales, central Queensland and South Australia (**Supplementary Tables S4**, **S5**, **Supplementary Figure S3**, and **Figure 3C**).

## Y- Chromosome

Few Y-chromosome studies of Aboriginal Australians have been published to date, with the majority showing unique Aboriginal Australian Y-chromosome haplogroups, C* and K* predominantly, and M* in rare cases. As a result, there is a limited modern Y-chromosome database, which did not allow for phylogenetic analyses similar to the work done on the mitochondrial genomes (Bergstrom et al., 2016; Nagle et al., 2017). In previous studies, one constant, however, was the detection of considerable levels of Eurasian admixture in modern individuals, with a large number of research participants self-identifying as Aboriginal Australian carrying non-Indigenous Y- chromosome haplotypes. The level of Eurasian admixture varied from study to study, with ~32% being reported by Malaspinas et al. (2016), ~56% by Nagle et al. (2016), and ~70% by Taylor and Henry (2012).

FLI2 showed the S1c haplotype (characterized by Z41926, Z41927, Z41928, Z41929, and Z41930 SNPs), while FLI4 belongs to the S1a3a haplotype (which is a subclade of S1a~ previously known as K2b1a). Both of these haplotypes are unique to

Aboriginal Australians. The determination of FLI1's haplotype was not possible due to the low coverage of the Y chromosome.

## DISCUSSION

The ethnography of the Flinders Islands as discussed by Hale and Tindale (1933) suggests that the status of an individual in life was reflected in the complexity associated with their mortuary practices. Moreover, the Flinders Island group were often visited by people from the mainland and other islands. During their visits to the island they were often involved in ceremonies (Hale and Tindale, 1933; Rigsby and Chase, 1998). A map of Apa spit, where individual FLI1 was found eroding from the beach, drawn by Tindale during his visit to Flinders Island in the 1920s, showed visitation to the islands by at least four other tribal groups at that time. Visitors and non-locals were also offered different mortuary practices.

FLI1 was an older individual exhibiting extreme occlusal wear and periapical lesions (Adams et al., unpublished). Being an elderly individual at the time of death FLI1 may have represented one of the less esteemed members of the community that was not seen as a threat in the afterlife. He was buried a short time after death with no signs of extensive ceremony and complex interment. The rock placed on the torso of FLI1 represents a distinct funerary practice hitherto undescribed in available published literature for Australia. No other grave goods were observed in this burial. A modern Aboriginal interpretation of the purpose of the stone as a grave object, was provided by Traditional Owner Danny Gordon, who commented that the beach burial of FLI1 may have been an 'unliked man'(Danny Gordon pers. comm 2015). Alternatively, the FLI1 man may have died during his visit to the island and hence represents an example of a beach burial afforded to non-locals, as recorded by Hale and Tindale (1933). The mitochondrial genome of FLI1 provided more insights into his possible ancestry. This individual carries a mitochondrial haplotype (S2a) that is more dominant in New South Wales, especially among the Willandra Lakes communities (**Supplementary Tables S2**, **S4** and **Supplementary Figure S3**). We showed that the male beach burial (FLI1) carried a mitochondrial lineage that differed from other individuals, including contemporary communities from North Queensland. The haplotype difference of FLI1 from the other Flinders Island individuals may suggest that he was born away from the island but raised there, or was a visitor to the island at the time of his death.

The young woman (STI1) was also recovered from beach foredunes, so she might have been another example of a visitor burial. However, the recorded correlation between social status and the extent of ceremony and burial complexity is perhaps another explanation for the nature of her internment. Her placement on the beach indicating that this young woman did not meet the criteria for a more complex burial ritual (Roth, 1907; Adams et al., unpublished). She was buried articulated, and therefore not defleshed before burial, with no other burial goods found. However, it is important to note that this burial had been heavily disturbed, so it is possible that such goods may have existed prior to the disturbance of the burial (Adams et al., unpublished). Notably, we observed that STI1 carries a maternal haplogroup (P*) that dominates Aboriginal Australian communities in Queensland (**Supplementary Figure S3**), especially Cape York. particularly Cape York. Moreover, the STI1 woman shared an ancestral maternal lineage with the FLI2 individual who was buried in the rock-shelter. These results strongly suggest that the STI1 woman was likely a Flinders Island group individual, however, we could not rule out the possibility that she was a visitor from the nearby mainland.

Bark burial coffins are a feature of the Flinders Islands burial landscape, indicating a more complex form of funerary practice. All three sets of remains (FLI2, FLI3, and FLI4) recorded here represent individuals that were ritually prepared in accordance with burial protocols that have been ethnographically documented (Roth, 1898, 1907; Hale and Tindale, 1933). The genomic data of those three Flinders Island Aboriginal Australians, combined with modern genomes, form the basis of our understanding of the maternal haplotypes expected from that area of Queensland. Although the three belong to a common haplogroup (P*) for the region, they are from three different sub-haplotypes; P5b1, P12b, and P5a1a consecutively.

The absence of a close kinship relationship between FLI2 and FLI4 was supported by the little information gained through the study of the paternal lineage. Both FLI2 and FLI4 carried two different Indigenous Australians Y-chromosome haplotypes (S1c and S1a∼). Moreover, the C14 dates for those two individuals do not overlap after the Marine13 correction (**Supplementary Table S1**). This reveals that FLI3 (B2) and FLI4 (B3), despite being buried within the same cave, did not share a kinship relationship.

However, the Y chromosome results, when compared with the contemporary population from Queensland, did not provide any additional understanding of the burial practices performed on the island. This was a result of the significant levels of Eurasian admixture observed in the contemporary population.

## CONCLUSION

Although burials of Indigenous people often represent a poor source of DNA preservation, particularly in Australia, we have shown here that it is possible to recover full mitochondrial genomes, even in tropical contexts. The analysis of the ancient mitogenomes, in combination with a fine-scale genomic map of Aboriginal Australian mitochondrial haplotypes, can be used to better understand and test a range of hypotheses about mortuary practices and changes over time. Our genomic findings resolved to provide indicative answers to the questions about the beach burials, showing that FLI1 who was buried on the beach, most likely, for being a non-local, rather than being a less esteemed member of the Flinders Island community. While the STI1 female was more likely a less esteemed member of the community. These results do reflect some of the purported inequalities in Indigenous Australian mortuary practices that might have been based on age, sex or status of individuals, but they also reveal that ethnohistoric observations may reflect biases from early ethnographers. The presence of a young female FLI3 buried in the rock-shelter, indicates that both sexes were afforded complex mortuary rituals by their community.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in NCBI Genbank, NCBI Accession No.'s MK165665 to MK165690.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Griffith University Human Ethics. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## AUTHOR CONTRIBUTIONS

SW, MW, and DL conceived the project. SW, CF, and MW conceived the idea of the manuscript. DL, MW, and EW provided funding. MW, SA, and CF conducted the field work and collected ancient samples from the Flinders Island Group. SW performed the experiments and analyses of the data and drafted the manuscript. JW and SW checked the mitochondrial haplotypes. SW interpreted the data with help from DL, MW, CF, and JW. JW, DL, and SW revised the manuscript, with contributions from all the authors.

## FUNDING

This research was possible due to funding from the Australian Research Council LP140100387.

## ACKNOWLEDGMENTS

We thank all the Aboriginal Australian participants for supporting this research. We would like to acknowledge the QLD Police, DATSIP, and QPWS for assistance in the field. We also thank David McGahan for his assistant in the excavation of Stanley Island. We thank the National Throughput DNA Sequencing Centre at the University of Copenhagen for ancient DNA sequencing. We also thank Assoc. Prof. Craig Millar from The University of Auckland, New Zealand for valuable comments on the original manuscript. We would like to thank Prof. Andrew Smith, Pro Vice Chancellor (Griffith Sciences) for his guidance and support.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.00217/full#supplementary-material

## REFERENCES

Beaton, J. (1985). Evidence for a coastal occupation time-lag at Princess Charlotte Bay (North Queensland) and implications for coastal colonization and population growth theories for Aboriginal Australia. *Archaeol. Ocean.* 20, 1–20.

Bergstrom, A., Nagle, N., Chen, Y., McCarthy, S., Pollard, M. O., Ayub, Q., et al. (2016). Deep roots for aboriginal Australian Y chromosomes. *Curr. Biol.* 26, 809–813. doi: 10.1016/j.cub.2016.01.028

Bronk Ramsey, C. (2013). *OxCal 4.2 Web Interface Build*. Amsterdam: Elsevier.

Collard, M., Wasef, S., Adams, S., Wright, K., Mitchell, R. J., Wright, J. L., et al. (2019). Giving it a burl: towards the integration of genetics, isotope chemistry, and osteoarchaeology in Cape York, Tropical North Queensland, Australia. *World Archaeol.* 51, 602–619. doi: 10.1080/00438243.2019.1686418

Coppinger, R. W. (1883). *Cruise of the "Alert.": Four Years in Patagonian, Polynesian, and Mascarene Waters.(1878-82)*. Scotland: WS Sonnenschein.

Dabney, J., Knapp, M., Glocke, I., Gansauge, M. T., Weihmann, A., Nickel, B., et al. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15758–15763. doi: 10.1073/pnas.1314445110

Excoffier, L., and Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Fastx_Toolkit (2009). *fastx_toolkit (Version V0.0.13).*

Hale, H., and Tindale, N. (1933). Aborigines of princess Charlotte Bay. *Rec. S. Austral. Museum* 5, 63–116.

Heupink, T. H., Subramanian, S., Wright, J. L., Endicott, P., Westaway, M. C., Huynen, L., et al. (2016). Ancient mtDNA sequences from the First Australians revisited. *Proc. Natl. Acad. Sci. U.S.A.* 113, 6892–6897. doi: 10.1073/pnas.1521066113

Hogg, A. G., Hua, Q., Blackwell, P. G., Niu, M., Buck, C. E., Guilderson, T. P., et al. (2013). SHCal13 Southern Hemisphere calibration, 0–50,000 years cal BP. *Radiocarbon* 55, 1889–1903. doi: 10.2458/azu_js_rc.55.16783

Horsfall, N. (1991). Report on Visit to Flinders Island Group, Princess Charlotte Bay (Confidential).

Hudjashov, G., Kivisild, T., Underhill, P. A., Endicott, P., Sanchez, J. J., Lin, A. A., et al. (2007). Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8726–8730. doi: 10.1073/pnas.0702928104

Jack, R. L. (1921). *Northmost Australia: Three Centuries of Exploration, Discovery, and Adventure in and Around the Cape York Peninsula, Queensland, with a Study of the Narratives of All Explorers by Sea and Land in the Light of Modern Charting, Many Original Or Hitherto Unpublished Documents, Thirty-nine Illustrations, and Sixteen Specially Prepared Maps*. Michigan: Kent & Company, Limited.

Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P. L., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684. doi: 10.1093/bioinformatics/btt193

Katoh, K., Rozewicki, J., and Yamada, K. D. (2017). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* 20, 1160–1166. doi: 10.1093/bib/bbx108

King, P. P. (1827). Narrative of a Survey of the Intertmpical and Western Coasts of Australia. *Perf. Betw. Years* 18:18.

Knapp, M., Clarke, A. C., Horsburgh, K. A., and Matisoo-Smith, E. A. (2012). Setting the stage - building and working in an ancient DNA laboratory. *Ann. Anat.* 194, 3–6. doi: 10.1016/j.aanat.2011.03.008

Kozlov, A., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35, 4453–4455. doi: 10.1093/bioinformatics/btz305

Kutanan, W., Kampuansai, J., Srikummool, M., Kangwanpong, D., Ghirotto, S., Brunelli, A., et al. (2017). Complete mitochondrial genomes of Thai and Lao populations indicate an ancient origin of Austroasiatic groups and demic diffusion in the spread of Tai–Kadai languages. *Hum. Genet.* 136, 85–98. doi: 10.1007/s00439-016-1742-y

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475, 493–496. doi: 10.1038/nature10231

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., et al. (2010). The sequence and de novo assembly of the giant panda genome. *Nature* 463, 311–317. doi: 10.1038/nature08696

Malaspinas, A. S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., et al. (2016). A genomic history of aboriginal Australia. *Nature* 538, 207–214. doi: 10.1038/nature18299

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 2010:pdb.prot5448. doi: 10.1101/pdb.prot5448

Nagle, N., Ballantyne, K. N., van Oven, M., Tyler-Smith, C., Xue, Y., Taylor, D., et al. (2016). Antiquity and diversity of aboriginal Australian Y-chromosomes. *Am. J. Phys. Anthropol.* 159, 367–381. doi: 10.1002/ajpa.22886

Nagle, N., Van Oven, M., Wilcox, S., van Holst Pellekaan, S., Tyler-Smith, C., Xue, Y., et al. (2017). Aboriginal Australian mitochondrial genome variation– an increased understanding of population antiquity and diversity. *Sci. Rep.* 7:43041.

Okonechnikov, K., Conesa, A., and Garcia-Alcalde, F. (2015). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 32, 292–294. doi: 10.1093/bioinformatics/btv566

Pardoe, C. (1993). "The Pleistocene is still with us: analytical constraints and possibilities for the study of ancient human remains in archaeology," in *Sahul in Review: Pleistocene Archaeology in Australia, New Guinea and Island Melanesia*, eds M. A. Smith, M. Spriggs, and B. Fankhauser (Canberra: Australian National University).

Pearson, M. P. (1999). *The Archaeology of Death and Burial*. Stroud: Sutton Phoenix Mill.

Ralf, A., Montiel González, D., Zhong, K., and Kayser, M. (2018). Yleaf: software for human Y-chromosomal haplogroup inference from next-generation sequencing data. *Mol. Biol. Evol.* 35, 1291–1294. doi: 10.1093/molbev/msy032

Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K. E., Rasmussen, S., Albrechtsen, A., et al. (2011). An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* 334, 94–98. doi: 10.1126/science.1211177

Renaud, G., Slon, V., Duggan, A. T., and Kelso, J. (2015). Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* 16:224.

Rigsby, B., and Chase, A. (1998). "The sandbeach people and dugong hunters of eastern cape york peninsula: property rights in land and Sea Country," in *Customary Marine Tenure*, eds N. Peterson and B. Rigsby (Camperdown: Australia University of Sydney).

Roth, W. E. (1898). *On the Aboriginals occupying the Hinterland of Princess Charlotte bay, together with a preface containing suggestions for their better protection and improvement*. Brisbane: Mitchell Library, State Library of NSW.

Roth, W. E. (1907). *North Queensland Ethnography: Bulletin*. Sydney: Australian Museum.

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., Al-Rasheid, K. A., Willerslev, E., et al. (2012). Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* 13:178. doi: 10.1186/1471-2164-13-178

Skoglund, P., Storå, J., Götherström, A., and Jakobsson, M. (2013). Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J. Archaeol. Sci.* 40, 4477–4482. doi: 10.1016/j.jas.2013.07.004

Sutton, P., Rigsby, B., and Chase, A. (1993). *Traditional Groups of the Princess Charlotte Bay Region*. Cairns: Cape York Land Council.

Taylor, D. A., and Henry, J. M. (2012). Haplotype data for 16 Y-chromosome STR loci in Aboriginal and Caucasian populations in South Australia. *For. Sci. Int. Genet.* 6, e187–e188. doi: 10.1016/j.fsigen.2012.05.005

Tilley, L. (2015). *Theory and Practice in the Bioarchaeology of Care*. Cham: Springer.

Tobler, R., Rohrlach, A., Soubrier, J., Bover, P., Llamas, B., Tuke, J., et al. (2017). Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia. *Nature* 544:180. doi: 10.1038/nature21416

van Holst Pellekaan, S. M., Ingman, M., Roberts-Thomson, J., and Harding, R. M. (2006). Mitochondrial genomics identifies major haplogroups in Aboriginal Australians. *Am. J. Phys. Anthropol.* 131, 282–294. doi: 10.1002/ajpa.20426

Walsh, G. L. (1985). *Site Recording Report and Site Development Suggestions*. National Park: Flinders Group.

Wasef, S., Huynen, L., Millar, C. D., Subramanian, S., Ikram, S., Holland, B., et al. (2018). Fishing for mitochondrial DNA in the egyptian sacred Ibis mummies. *bioRxiv* [preprint]. doi: 10.1101/473454

Wright, J. L., Wasef, S., Heupink, T. H., Westaway, M. C., Rasmussen, S., Pardoe, C., et al. (2018). Ancient nuclear genomes enable repatriation of Indigenous human remains. *Sci. Adv.* 4:eaau5064.

**Conflict of Interest:** SA is employed by Everick Foundation Pty Ltd., Brisbane, QLD, Australia.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Ecology and Evolution

ORIGINAL RESEARCH
published: 10 July 2020
doi: 10.3389/fevo.2020.00218

# The Challenges of Reconstructing Tropical Biodiversity With Sedimentary Ancient DNA: A 2200-Year-Long Metagenomic Record From Bwindi Impenetrable Forest, Uganda

***René Dommain[1,2]*, Morgan Andama[3], Molly M. McDonough[4,5], Natalia A. Prado[4,6,7,8], Tobias Goldhammer[9], Richard Potts[2], Jesús E. Maldonado[4,7,8,10], John Bosco Nkurunungi[11] and Michael G. Campana[4,7,8]***

*[1] Institute of Geosciences, University of Potsdam, Potsdam, Germany, [2] Human Origins Program, National Museum of Natural History, Smithsonian Institution, Washington, DC, United States, [3] Department of Biology, Faculty of Science, Muni University, Arua, Uganda, [4] Center for Conservation Genomics, Smithsonian National Zoological Park and Conservation Biology Institute, Smithsonian Institution, Washington, DC, United States, [5] Department of Biological Sciences, Chicago State University, Chicago, IL, United States, [6] Center for Species Survival, Smithsonian National Zoological Park and Conservation Biology Institute, Smithsonian Institution, Front Royal, VA, United States, [7] Department of Environmental Science and Public Policy, George Mason University, Fairfax, VA, United States, [8] School of Systems Biology, George Mason University, Fairfax, VA, United States, [9] Department of Chemical Analytics and Biogeochemistry, Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany, [10] Department of Biology, George Mason University, Fairfax, VA, United States, [11] Department of Biology, Faculty of Science, Mbarara University of Science and Technology, Mbarara, Uganda*

OPEN ACCESS

***Edited by:***
*Nathan Wales,*
*University of York, United Kingdom*

***Reviewed by:***
*Mikkel Winther Pedersen,*
*University of Cambridge,*
*United Kingdom*
*Charline Giguet-Covex,*
*UMR5204 Environnements,*
*DYnamiques et TErritoires de la*
*Montagne (EDYTEM), France*

****Correspondence:***
*René Dommain*
*dommain@uni-potsdam.de*

***Specialty section:***
*This article was submitted to*
*Paleoecology,*
*a section of the journal*
*Frontiers in Ecology and Evolution*

***Received:*** *18 November 2019*
***Accepted:*** *12 June 2020*
***Published:*** *10 July 2020*

***Citation:***
*Dommain R, Andama M, McDonough MM, Prado NA, Goldhammer T, Potts R, Maldonado JE, Nkurunungi JB and Campana MG (2020) The Challenges of Reconstructing Tropical Biodiversity With Sedimentary Ancient DNA: A 2200-Year-Long Metagenomic Record From Bwindi Impenetrable Forest, Uganda. Front. Ecol. Evol. 8:218. doi: 10.3389/fevo.2020.00218*

Sedimentary ancient DNA has been proposed as a key methodology for reconstructing biodiversity over time. Yet, despite the concentration of Earth's biodiversity in the tropics, this method has rarely been applied in this region. Moreover, the taphonomy of sedimentary DNA, especially in tropical environments, is poorly understood. This study elucidates challenges and opportunities of sedimentary ancient DNA approaches for reconstructing tropical biodiversity. We present shotgun-sequenced metagenomic profiles and DNA degradation patterns from multiple sediment cores from Mubwindi Swamp, located in Bwindi Impenetrable Forest (Uganda), one of the most diverse forests in Africa. We describe the taxonomic composition of the sediments covering the past 2200 years and compare the sedimentary DNA data with a comprehensive set of environmental and sedimentological parameters to unravel the conditions of DNA degradation. Consistent with the preservation of authentic ancient DNA in tropical swamp sediments, DNA concentration and mean fragment length declined exponentially with age and depth, while terminal deamination increased with age. DNA preservation patterns cannot be explained by any environmental parameter alone, but age seems to be the primary driver of DNA degradation in the swamp. Besides degradation, the presence of living microbial communities in the sediment also affects DNA quantity. Critically, 92.3% of our metagenomic data of a total 81.8 million unique, merged reads cannot be taxonomically identified due to the absence of genomic references in public databases. Of the remaining 7.7%, most of the data (93.0%) derive from Bacteria and Archaea, whereas only 0–5.8% are from Metazoa and 0–6.9% from Viridiplantae, in part due to unbalanced taxa representation in the reference data. The plant DNA record at ordinal level agrees well with local pollen data but resolves less diversity. Our animal DNA record reveals the presence of 41 native taxa (16 orders) including Afrotheria, Carnivora, and Ruminantia at Bwindi during the past 2200 years. Overall, we observe no decline in taxonomic richness with increasing age suggesting that several-thousand-year-old information on past biodiversity can be retrieved from tropical sediments. However, comprehensive genomic surveys of tropical biota need prioritization for sedimentary DNA to be a viable methodology for future tropical biodiversity studies.

**Keywords: sedimentary ancient DNA, tropical biodiversity, DNA preservation, sediment, tropical swamp, shotgun sequencing, metagenomic analysis**

## INTRODUCTION

The tropics hold the greatest biodiversity on Earth (e.g., Gaston, 2000; Jenkins et al., 2013) containing about three quarters of all species (Barlow et al., 2018). It remains a major scientific challenge to unravel the processes that led to this exceptional diversity and to test existing models on the diversification of tropical biota (e.g., Stevens, 1989; Mittelbach et al., 2007; Pennington et al., 2015). Yet elucidating the evolution of tropical biodiversity and the assembly of tropical biotic communities over time is generally hampered by the lack of a taxonomically complete fossil record and the absence of exhaustive phylo- and biogeographic information. In particular tropical rainforests – the most diverse terrestrial ecosystems – are largely characterized by a poor fossil record (Wing et al., 2009; Jacobs et al., 2010), which severely limits the reconstruction of diversity patterns. For example, the mammalian fossil record for the past 20,000 years of Africa only contains few taxa from the large Congo Basin (Jousse, 2017) where several hundreds of mammal species occur today (Jenkins et al., 2013). Past diversity and community structure of small rainforest vertebrates or insects may be reconstructed from amber deposits (e.g., Wilson, 1985; Sherratt et al., 2015), which, however, are of very localized occurrence (Poinar, 1992). Floristic diversity can be reconstructed from pollen records, but generally suffer from limited taxonomic resolution in comparison to the existing tropical plant diversity (e.g., Morley, 2000). Hence, a comprehensive picture of the history of tropical biodiversity is still far from being established.

In addition to addressing this challenging scientific problem, the ongoing collapse in biodiversity in the tropics (Ceballos et al., 2017; Barlow et al., 2018) resulting from anthropogenically driven habitat degradation, deforestation, forest fragmentation, and defaunation necessitates well-documented species inventories for successful conservation management. Establishing baseline data on the "pre-disturbance" occurrence of species within a geographic area is relevant for protected area planning, ecosystem restoration and assessments on the global vulnerability of species by the International Union for the Conservation of Nature (IUCN). However, the species composition of most tropical regions prior to extensive modern human disturbance is often not known, except for few areas for which historic accounts or natural history museum specimens exist. Therefore, the development of independent means for reconstructing species assemblages would be of enormous help to conservationists and policy makers.

Besides conventional paleoecological tools and historical records, emerging ancient DNA approaches are likely to be very useful for elucidating the hyperdiversity conundrum of the tropics and in contributing to the urgent conservation needs of tropical countries (Andersen et al., 2012; Hofman et al., 2015; Thomsen and Willerslev, 2015; Díez-del-Molino et al., 2018; Epp, 2019). Ancient DNA (aDNA) can be directly extracted from fossil bones, teeth (e.g., Adler et al., 2011) and plant macro-remains (e.g., Estrada et al., 2018), thereby providing genetic information on individual organisms and species. Ancient DNA may also be present in sediments and, if sufficiently preserved can potentially open a window into the past faunal and floral composition of ecosystems (e.g., Willerslev et al., 2003; Slon et al., 2017). Sedimentary ancient DNA (sedaDNA) has been successfully used to detect the presence of extant, extinct and introduced species, including rare and large vertebrates, in various sedimentary environments (e.g., Haile et al., 2007, 2009; Boessenkool et al., 2012; Giguet-Covex et al., 2014; Graham et al., 2016; Slon et al., 2017).

Sedimentary DNA originates from traces of organisms left in the environment which are incorporated into the sediments of lakes, swamps or caves (e.g., Hofreiter et al., 2003; Willerslev et al., 2003; Haile et al., 2007; Andersen et al., 2012; Parducci et al., 2017; Slon et al., 2017). DNA-containing remains include hair, feces, urine, skin, eggs, feathers, tissue, and seeds. Sedimentary DNA may be deposited directly by the locally present biota such as aquatic taxa in lakes, cave-dwelling organisms or plants rooting in swamps. In addition, DNA may also be transported into a sedimentary basin from its catchment via rivers or runoff (Parducci et al., 2017). Depending on the stage of decay, nucleic acids may either be intra- or extracellular. The extracellular DNA is then microbially processed or adsorbed by cations, clay minerals, apatite, silica or other sedimentary matter (e.g., Lorenz and Wackernagel, 1987, 1994; Ogram et al., 1988; Armbrecht et al., 2019).

Thus far, sedaDNA has primarily been studied in cold climate regions, particularly northern Eurasia, North America and New Zealand where preservation conditions are ideal (e.g., Willerslev et al., 2014; Parducci et al., 2017; Slon et al., 2017; Kisand et al., 2018). For example, sediments from Arctic permafrost have yielded DNA up to 400,000 years old (Willerslev et al., 2003). The vast majority of sedaDNA studies applied metabarcoding, which is the amplification and sequencing of widely used mitochondrial or chloroplast markers such as mitochondrial 16S rRNA, cytochrome b, cytochrome oxidase I, and the chloroplast *trn*L P6 loop. However, this approach restricts the detection of taxa to selected groups and amplifies certain species more efficiently than others due to polymerase chain reaction (PCR) biases, thus providing a biased perspective on the past taxonomic composition of a site (e.g., Ziesemer et al., 2015; Parducci et al., 2018; Adams et al., 2019). Due to PCR biases and PCR template duplications, robust quantitative comparison of taxonomic profiles is difficult (Adams et al., 2019), even with gene copy normalization (Starke and Morais, 2019). Furthermore, metabarcoding approaches discard the majority of extracted DNA, problematic for conservation of non-replaceable aDNA samples. Moreover, the short length of degraded aDNA precludes the use of long barcoding markers (Willerslev et al., 2014), while the detection of rare taxa necessitates numerous runs of PCR. In comparison, shotgun sequencing reveals the entire metagenomic composition of a sediment sample and thus allows for the most comprehensive description of taxonomic diversity and community composition (e.g., Ziesemer et al., 2015; Slon et al., 2017; Pedersen et al., 2016). This method avoids PCR amplification, which reduces the risk of contamination with modern DNA (Armbrecht et al., 2019) but requires more comprehensive genomic databases to reliably identify obtained sequences (e.g., Rawlence et al., 2014; Cribdon et al., 2020). Nevertheless, reliable quantification of organism prevalence is difficult because metagenomic profiles are biased by the differing sizes of individual genomes – species with larger genomes generate more DNA sequences than ones with smaller genomes and thus appear to be more frequent (e.g., Segata et al., 2012). This bias can be corrected, but only if the genome sizes of all identified organisms are known.

SedaDNA studies of tropical regions are very rare and thus far principally absent from hyperdiverse tropical rainforests (Pedersen et al., 2015; Epp, 2019). For example, sub-Saharan Africa is understudied for aDNA as a whole (Campana et al., 2013) and the history of its rich vertebrate fauna has not been investigated from a sedaDNA perspective. We are aware of only six studies from tropical Africa that used sedaDNA either to explore floristic diversity or aquatic organisms such as diatoms and *Daphnia* (Mergeay et al., 2007; Epp et al., 2010, 2011; Stoof-Leichsenring et al., 2012; Boessenkool et al., 2014; Bremond et al., 2017). A possible reason for the scarcity of studies in the tropics is the assumption that high temperatures are not conducive for long-term DNA preservation (Kistler et al., 2017). Recently it has been shown, however, that DNA can persist in tropical lacustrine and marine sediments under high temperatures for hundreds to thousands of years (Bremond et al., 2017; Gomez Cabrera et al., 2019). However, the preservation of DNA in different sediment types in general, and in tropical systems in particular, is insufficiently known (e.g., Epp, 2019). Dry sediments in temperate regions seem to preserve DNA relatively well (Hofreiter et al., 2003), but vertical transport and leaching of DNA may lead to contamination within stratigraphic profiles (Haile et al., 2007). The role of DNA leaching in saturated sediments needs to be examined since the polarity of DNA molecules should promote its degradation by hydrolysis (Lindahl, 1993; Pääbo et al., 2004). Given these issues, a much better understanding of transport, deposition, preservation and degradation of sedimentary DNA in the tropics is needed (Domaizon et al., 2017; Giguet-Covex et al., 2019).

To address these open questions and to explore the value of sedaDNA as a tool for tropical conservation biology, we investigated the metagenomic composition and patterns of DNA degradation in several sediment cores collected from one of the most-diverse tropical rainforests of equatorial Africa – Bwindi Impenetrable Forest in Uganda (Kingdon, 1973; Butynski, 1984). Bwindi Forest is part of the Albertine Rift – a globally important conservation region which is characterized by high levels of endemism and the richest vertebrate diversity in Africa (Plumptre et al., 2007). It also belongs to the Eastern Afromontane Biodiversity hotspot (Brooks et al., 2004) highlighting its biogeographic uniqueness and vulnerability to anthropogenic impacts. Bwindi is the richest forest in Uganda with regard to the number of recorded mammal and plant species (Plumptre et al., 2007; Olupot and Plumptre, 2010). Importantly, Bwindi harbors one of the two globally surviving populations of mountain gorillas (*Gorilla beringei beringei*) with a population of about 400 individuals (Roy et al., 2014). This forest reserve is also the only place where mountain gorillas are sympatric with chimpanzees (*Pan troglodytes schweinfurthii*) (Stanford and Nkurunungi, 2003). Our study leveraged an undisturbed swamp located in the middle of Bwindi Forest that contains sediments dating back to the Pleistocene (Marchant et al., 1997). Bwindi thus provides an excellent opportunity to test whether tropical diverse rainforest biota can be detected with sedaDNA.

Here we present metagenomic assemblages across the tree of life (i.e., Bacteria, Archaea, and Eukaryota) retrieved by shotgun sequencing of sedaDNA covering the past 2200 years. The overall objectives of our study are (1) to investigate conditions of sedaDNA preservation and degradation in relation to chemical and physical sediment and water properties in a tropical environment and (2) to characterize taxonomic diversity and metagenomic composition of Bwindi over time. Specifically, we asked the following questions:

(1) Is endogenous ancient DNA preserved in tropical swamp sediments?
(2) How do temperature, acidity, nutrient content, elemental composition, and sediment lithology and age influence DNA preservation?
(3) How precisely can the taxonomic level of shotgun-sequenced sediments from a diverse tropical site be resolved at present?

To this end, we compare DNA concentration, molecular fragment length and cytosine deamination patterns as signals of DNA degradation with a multivariate dataset consisting of the age, type, chemistry and temperature of the sediments. Using our metagenomic approach, we compare past and present ecological communities preserved in the sediment and present taxonomic assignments using three metagenomic classifiers and four reference databases with a focus on plant and animal detections, assess the accuracy of identifications by comparing plant DNA signals with pollen data and animal DNA assignments with modern occurrence data and finally examine biases introduced by metagenomic reference databases and bioinformatic approaches in the detection of taxa.

## MATERIALS AND METHODS

### Study Area

Bwindi Impenetrable National Park is a 331 km$^2$ large forest (0°53′–1°08′ S; 29°35′–29°50′E) in southwest Uganda (**Figure 1**), which supports an exceptional diversity of at least 135 mammal species (10 primate species), 381 bird species, 34 reptile species, 29 amphibian species and 393 tree species (Plumptre et al., 2007). Fifty-six tetrapod species and 74 plant species of Bwindi are Albertine Rift endemics (Plumptre et al., 2007).

Bwindi is located in the Kigezi Highlands (**Figure 1**), which are part of the eastern shoulder of the Albertine Rift – the western arm of the East African Rift. The bedrock geology of Bwindi consists of Precambrian metamorphic rocks such as schist, quartzite, shale and granite and the soils are ferrallitic (Butynski, 1984). The elevational gradient of this relatively small reserve ranges from 1190 m in its northern most part to over 2600 m in the southeastern corner based on the digital elevation model of **Figure 1**. The land surface is strongly dissected, resulting in substantial relief differences and steep slopes.

At Ruhija (2350 m, **Figure 1A**) mean monthly temperatures are between 13.4 and 19.1°C (September 2001 to August 2002: monthly median 15.1°C; Nkurunungi et al., 2004) and mean annual precipitation is 1378 mm recorded over the period 1987–2006 (Kasangaki et al., 2012). Two rainy seasons occur between March to May and September to November associated with the biannual passage of the tropical rainbelt. Bwindi is covered with mid-altitude and montane rainforest and a small area in the southeast with bamboo (*Arundinaria alpina*) forest. The dominating tree species are *Chrysophyllum gorungosanum*, *Entandrophragma excelsum*, *Neoboutonia macrocalyx*, *Newtonia buchanani,* and *Parinari excelsa* (Butynski, 1984; Olupot and Plumptre, 2010).

In the southcentral part of Bwindi lies Mubwindi Swamp – a 1 km$^2$ flat, slightly inclined peatland fed by several streams from a 12 km$^2$ large, forested catchment (**Figures 1**, **2A**). We cored Mubwindi Swamp due to its long sediment record and because it serves as habitat to various terrestrial and (semi) aquatic species, which enhances the possibility of DNA deposition. The swamp is mostly covered with sedge communities dominated by *Cyperus latifolius* and *Cyperus denudatus*, which are often associated with *Thelypteris* cf. *confluens*, *Alchemilla johnstonii*, *Crassocephallum paludum*, *Lobellia mildbraedi,* and other species.

Several large mammals use Mubwindi Swamp as a foraging or breeding site like Sitatunga (*Tragelaphus spekei*) and bushbuck (*Tragelaphus scriptus*) (Mugerwa et al., 2013; pers. obs.). Mountain gorillas visit the swamp's edge to feed on the thistle *Carduus kikuyorua* (Ganas et al., 2009; Rothman et al., 2014; pers. obs.). The swamp is also home to breeding populations of several Albertine Rift endemics, including the globally endangered Grauer's swamp warbler (*Bradypterus graueri*; Kahindo et al., 2017), Delany's swamp mouse (*Delanymys brooksi*; Kasangaki et al., 2003), and several frog species such as Bururi long-fingered frog (*Cardioglossa cyaneospila*; Blackburn et al., 2016) and Ahl's reed frog (*Hyperolius castaneus*; Drewes and Vindum, 1994). Bwindi forest apparently lies within the hybrid zone of forest and savanna elephants (*Loxodonta cyclotis*, *L. africana*; Mondol et al., 2015) and currently supports a population of less than 40 elephants (Kasangaki et al., 2012), which regularly visit the forests around Mubwindi and the swamp itself for foraging and drinking, particularly during the dry season (Babaasa, 2000, pers. obs.). In the catchment of Mubwindi Swamp occur other rare vertebrates such as African golden cat (*Caracal aurata*), yellow-backed duiker (*Cephalophus silvicultor*) and one of Africa's rarest birds, the African green broadbill (*Pseudocalyptomena graueri*) (Olupot and Plumptre, 2010; Mugerwa et al., 2013; pers. obs.).

### Core Collection, Logging, and Processing

Between August 4 and 7, 2017 we collected seven sediment cores from Mubwindi Swamp (1°4′ S, 29°45′ E, ca. 2090 m a.s.l.) with a modified Livingstone-type corer equipped with a square rod and steel barrel, which was 5 cm in diameter (Wright et al., 1984). Prior to collecting cores, the piston and core barrel were washed with boiling water to reduce DNA contamination. In the swamp center, we collected two long cores from two sites (site 1 and 2, **Figure 1**) that both reached a basal resistant gravel layer (cores designated MUB17-1A: 620 cm, MUB17-2C: 656 cm). To this end, we successively recovered one-meter-long core sections from the same bore hole and extruded the cores in the field. Site 1 was located in a mixed sedge-fern community dominated by *Cyperus denudatus* and *Thelypteris* cf. *confluens* (**Figure 2B**) and site 2 in a tall sedge community covered by dense stands of *Cyperus latifolius* and within 5 m of a stream (**Figure 2C**). At these sites we collected additional surface cores within 1 meter from the main cores for assessing DNA degradation and modern metagenomic composition (site 1: core 1E: 87 cm, site 2: cores 2A: 71 cm, 2D: 93 cm). For site 2, we combined core sections 2D-1 and sections 2C-2 through 2C-6 into a composite core (here referred to as "master core") for studying long-term patterns (**Figure 3**), whereas for site 1 we only present data from surface samples. Furthermore, we collected two short cores (4A, 5A) at the eastern edge of the swamp where elephants had just been active prior to our visit. Core MUB17-4A (83 cm) was recovered from a site covered with elephant dung (here referred to as "elephant dung site"; **Figure 2D**) and core MUB17-5A (79 cm)

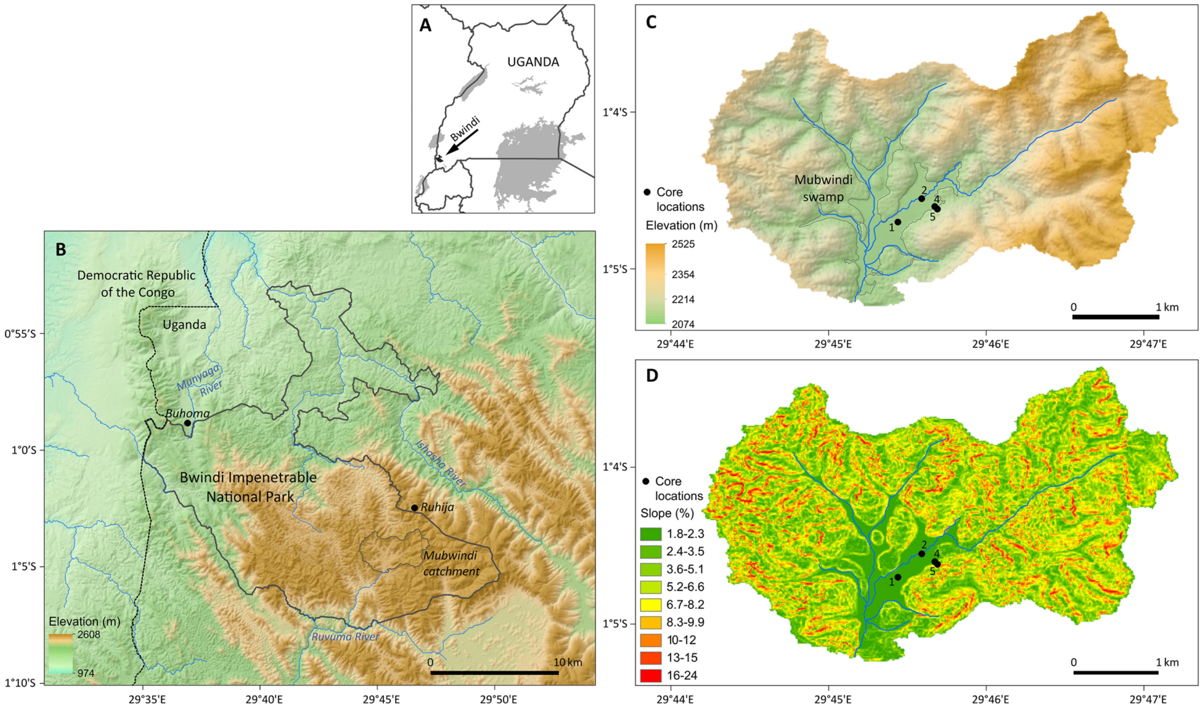


**FIGURE 1 |** Study area: **(A)** location of Bwindi Impenetrable National Park in Uganda, **(B)** digital elevation model of the Bwindi region with the national park boundary shown as thick gray line and the boundary of the catchment of Mubwindi Swamp as thin gray line, **(C)** elevation map and drainage network of the catchment of Mubwindi Swamp (bordered by gray line) with locations of coring sites 1, 2, 4 and 5, and **(D)** slope map of the Mubwindi catchment and swamp. Maps 2-4 and catchment boundaries are based on the TanDEM-X digital elevation model (© DLR).

retrieved from a flooded elephant wallow (here referred to as "elephant wallow site"; **Figure 2E**).

All core sections were wrapped in plastic wrap and stored in ABS (acrylonitrile butadiene styrene) tubes for transport and permanent storage. Cores were shipped to the LacCore facility (University of Minnesota, MN, United States) for non-invasive analyses, sub-sampling and permanent cold storage. At LacCore, cores were first scanned non-invasively for wet bulk-density using gamma-ray attenuation with a Geotek multi-sensor core logger at 0.5 cm intervals (two runs per core section). Limited high-energy irradiation using Gamma or X-rays is unlikely to significantly impact sedimentary DNA preservation (M. Muschick, pers. com.; Wanek et al., 2013; Wanek and Rühli, 2016). Each core was split longitudinally with a bleached diamond-bladed band saw into a working split section used for destructive sub-sampling and an archive split section used for non-invasive analyses and permanent storage. Each fresh archive section was photographed with a GeoScan IV digital linescan camera at 300 dpi.

Under sterile conditions, we collected samples for sedimentary ancient DNA analyses from the freshly split working sections of cores 1A, 1E, 2A, 2C, 2D, 4A, 5A (**Figure 3** and **Supporting Data**). Two-cm-thick samples were taken from the core's center with bleached and acetone cleaned tools after removing one centimeter of sediment from the core surface to avoid any contamination (**Figure 3**). All samples were immediately frozen and shipped to the Smithsonian Conservation Biology Institute (Washington, DC) for metagenomic analyses. From the archive core sections we obtained point magnetic susceptibility profiles (0.5 cm resolution) with a Geotek XYZ core logger and X-ray fluorescence (XRF) elemental profiles (0.5 cm sampling resolution; 15 s dwell time) with an ITRAX X-ray core scanner at Large Lakes Observatory (Duluth, MN, United States).

## Stratigraphic and Geochemical Core Analyses

The lithostratigraphy of the cores was determined by visual inspection of the core surfaces and with a dissecting scope. In addition, mineralogy was analyzed microscopically from smear slides and with a Hitachi TM 1000 tabletop scanning electron microscope. For measuring water and organic matter content we collected two-cm-thick samples contiguously from the master core and two-cm-thick samples parallel to every DNA sample location from the other analyzed cores. Water content was determined by drying sediment at 100°C for 12 h in an oven and organic matter (OM) content was determined by loss-on-ignition of the dried samples at 550°C for 4 h in a muffle furnace (Dean, 1974). In addition, 1-cm-thick samples were taken for elemental analyses (C, N, S, P) immediately adjacent to every sedDNA sample location. These samples were freeze-dried and ground to a powder. Total elemental contents of carbon (C), nitrogen (N), and sulfur (S) were determined by high-temperature combustion elemental analysis (infrared detection, Elementar Vario EL, IGB Berlin). Total phosphorus

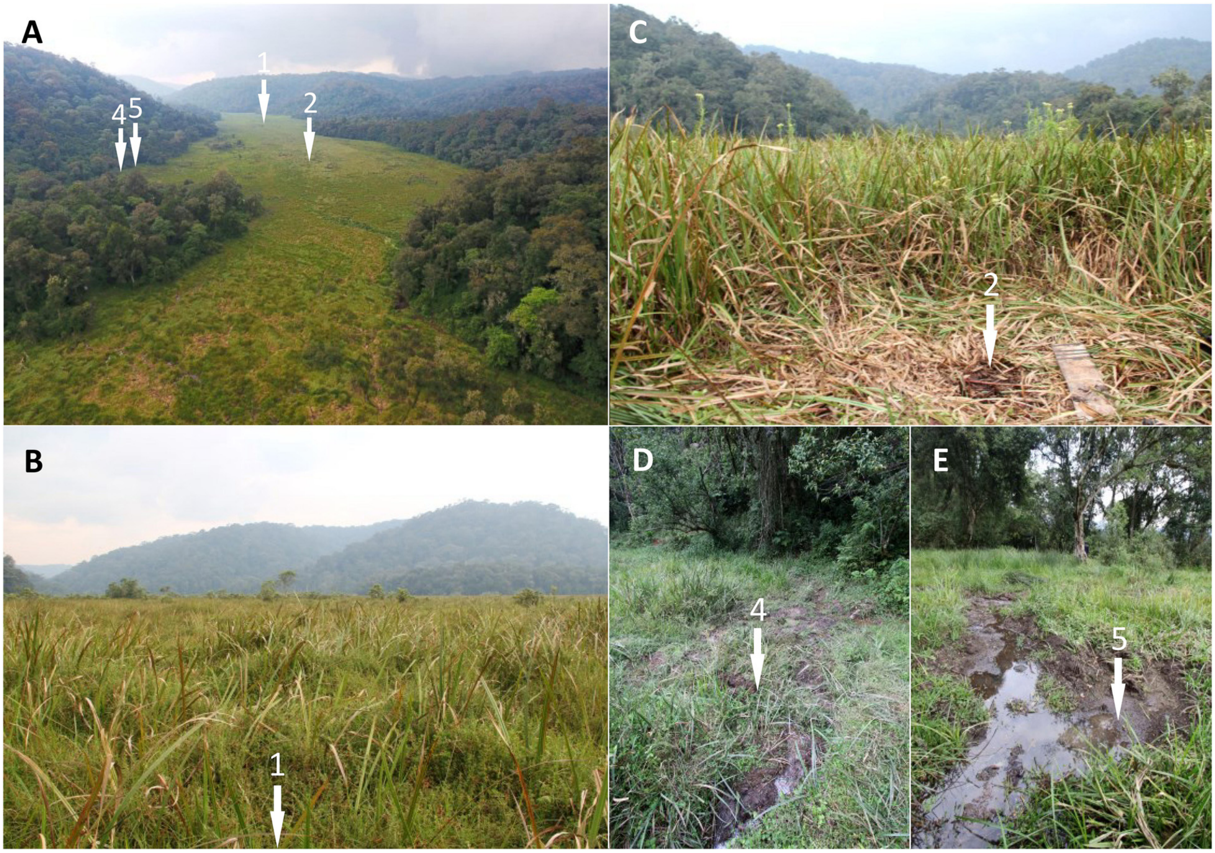


**FIGURE 2 |** Photographs of Mubwindi Swamp and coring sites. **(A)** Aerial photograph of Mubwindi from northeast to southwest with white arrows marking the coring sites. Note the elephant trails across the swamp in the foreground. **(B)** Coring site 1 at the swamp center, **(C)** coring site 2 (master core site) at the swamp center, **(D)** coring site 4 (elephant dung site) and **(E)** coring site 5 (elephant wallow). Note the recent elephant tracks in **(D,E)**.

content was determined by molybdenum blue spectrometry (Murphy and Riley, 1962) after thermal combustion (550°C) and hot potassium peroxodisulfate hydrolysis (Andersen, 1976; Ebina et al., 1983).

## Radiocarbon Dating

From cores 2C and 4A we collected plant macrofossils of non-aquatic plants and wood for AMS radiocarbon dating (**Table 1**). Site 5 (wallow) was not dated because of obvious bioturbation. Radiocarbon activity for 11 samples was determined by accelerator mass spectrometry at Lawrence Livermore National Laboratory (**Table 1**). We developed a Bayesian age-depth model for the master core with the BACON software package (Blaauw and Christen, 2011). All $^{14}$C dates were also calibrated into calendar years with the IntCal13 calibration curve (Reimer et al., 2013) in CALIB 7.1 (Stuiver and Reimer, 1993). Ages are presented as calendar years before present (i.e., present is 1950; "cal BP") and on the common era ("CE") notation.

## Field Measurements

To assess the influence of the local environment on DNA preservation, we conducted field surveys on water chemistry and physical and chemical soil properties of Mubwindi Swamp in June and September 2018 and July 2019. At all our four coring locations we measured soil pH and temperature changes with depth. To this end we successively retrieved 50-cm-long sediment core sections with a D-section core sampler and immediately upon core retrieval measured soil temperature in intervals of between 1 and 10 cm with an Extech digital thermometer, followed by pH with a Lutron soil pH meter at the same intervals. Close to our four coring locations, we also collected surface water samples from open pools or dug pits for the analysis of major dissolved elements, anions, and nutrients. Water samples were filtered (0.45 μm) and preserved by acidification (2M HCl) for subsequent analysis in the laboratory at IGB Berlin. Water temperature, pH and EC were measured at each sampling location with a handheld multiparameter probe (WTW Multi 3530). Dissolved aluminum (Al), calcium (Ca), iron (Fe), potassium (K), magnesium (Mg), manganese (Mn), and sodium (Na) were determined by inductively coupled-optical emission spectroscopy (ICP-OES, Thermo Scientific iCAP 6300). Dissolved chloride ($Cl^-$) and sulfate ($SO_4{}^{2-}$) were determined in samples without acidification by ion chromatography (conductivity detection after chemical suppression, Metrohm CompactIC). Dissolved ammonia

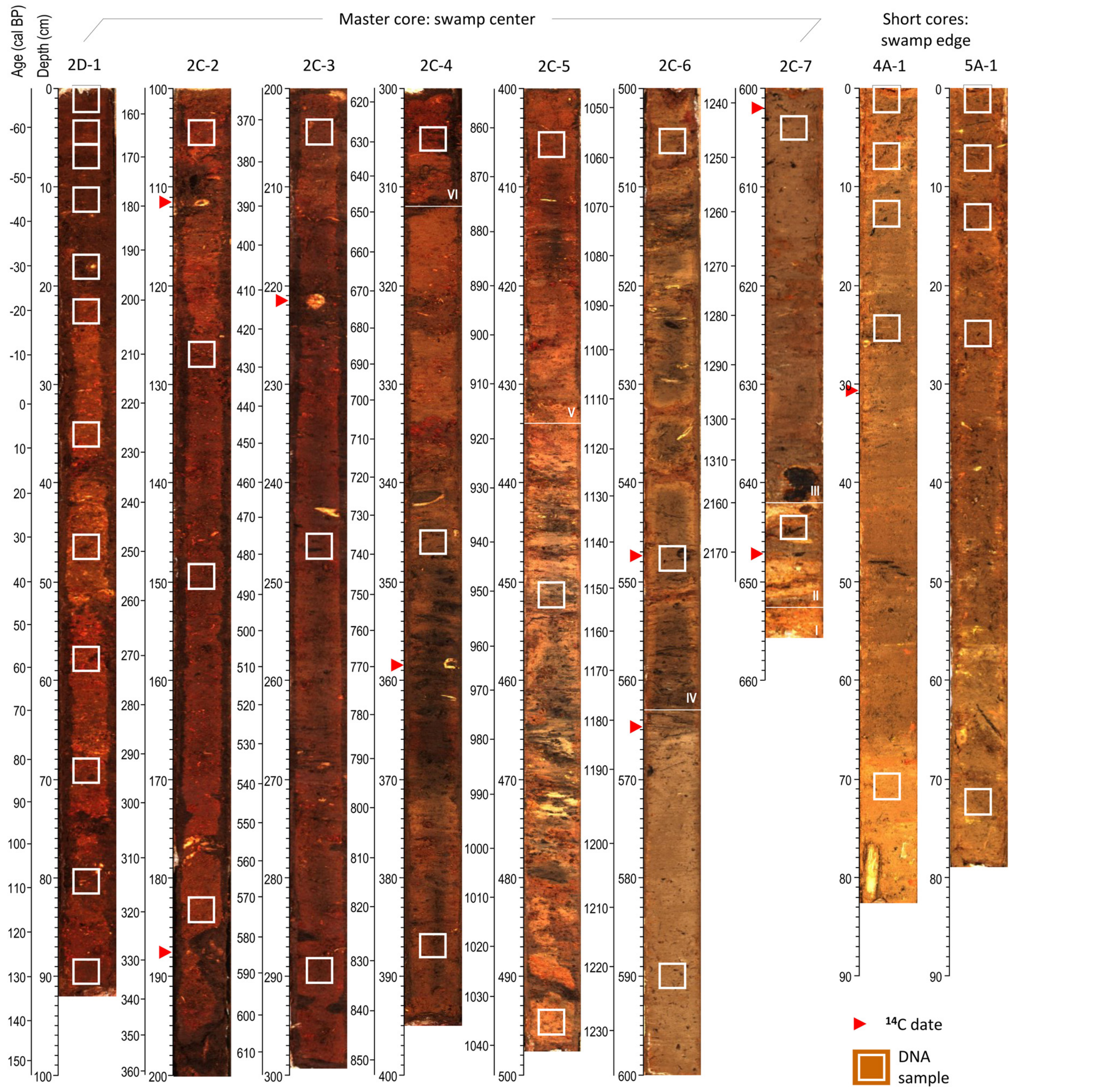


**FIGURE 3 |** Images of the analyzed core sections with age (calibrated median ages before present, only master core) and depth scales to the left of each image. White squares mark locations of sedDNA samples and red triangles locations of radiocarbon dates. The master core (MUB17-2C/D) consists of seven sections, with the top section shown on the left and the bottom section on the right. Core 4A-1 is the elephant dung site and 5A-1 the elephant wallow site. Roman numerals denote stratigraphic units of the master core, which are separated by white lines.

($NH_4^+$) and nitrate ($NO_3^-$) were determined by flow-segmented analysis (FSA, SEAL AutoAnalyzer 3). Dissolved organic carbon (DOC) was determined by thermocatalytic conversion infrared spectroscopy (Shimadzu TOC-L). Soluble reactive phosphorus (SRP) was determined by molybdenum blue spectroscopy (Murphy and Riley, 1962).

## Ancient DNA Extraction and Sequencing

We extracted DNA from 44 sediment samples (~0.25 g per extraction), 30 from our master core and the remainder from the short cores (**Figure 3** and **Supporting Data**), using DNeasy PowerSoil kits (Qiagen Inc., Germantown, MD, United States). To monitor DNA contamination during the coring and DNA extraction procedures, we also extracted DNA from two ~0.25 g samples of Lux bar soap (Unilever) used to grease the corer's piston and processed two additional sham extractions that contained only extraction reagents. DNA concentrations were measured using a Qubit® 2.0 fluorometer with the dsDNA HS assay kit (Thermo Fisher Scientific, Waltham, MA, United States).

**TABLE 1 |** Radiocarbon dates for the Mubwindi Swamp cores (MUB17-2C and MUB17-4A).

| Core section ID: MUB17- | Lab ID CAMS# | Mean depth below surface (cm) | Dated material | $^{14}$C Age (year BP) | $2\sigma$ bounds of calibrated age (yr cal BP) + (probability) | Median age estimate (yr cal BP) | Remarks |
|---|---|---|---|---|---|---|---|
| 2C-2 | 182574 | 111.5 | Wood | 145 ± 25 | 4–39 (0.178)<br>62–119 (0.212)<br>123–152 (0.121)<br>169–232 (0.320)<br>242–281 (0.168) | 150 | Master core |
| 2C-2 | 182575 | 187.2 | Wood | 280 ± 25 | 159–161 (0.003)<br>286–332 (0.444)<br>355–433 (0.554) | 375 | Master core |
| 2C-3 | 182576 | 221.8 | Wood | 300 ± 25 | 299–334 (0.270) 349–<br>439 (0.701)<br>442–455 (0.029) | 390 | Master core |
| 2C-3 | 182579 | 221.8 | Wood | 320 ± 25 | 306–341 (0.217)<br>347–462 (0.783) | 388 | Master core<br>Duplicate sample |
| 2C-4 | 182577 | 358.7 | Wood | 1030 ± 25 | 919–976 (1.0) | 946 | Master core |
| 2C-6 | 181997 | 547.9 | Seed | 1120 ± 35 | 939–942 (0.003)<br>954–1091 (0.919)<br>1107–1146 (0.051)<br>1159–1173 (0.027) | 1023 | Master core |
| 2C-6 | 178488 | 564.9 | Leaf fragment | 1245 ± 30 | 1077–1162 (0.282)<br>1171–1270 (0.718) | 1205 | Master core |
| 2C-6 | 178493 | 564.9 | Leaf fragment | 1280 ± 35 | 1091–1108 (0.017)<br>1129–1132 (0.003)<br>1147–1158 (0.013)<br>1173–1292 (0.967) | 1229 | Master core<br>Duplicate sample |
| 2C-7 | 178489 | 602.0 | Seed | 1230 ± 30 | 1068–1190 (0.659)<br>1199–1261 (0.341) | 1165 | Master core |
| 2C-7 | 178490 | 647.3 | Wood | 2160 ± 30 | 2057–2188 (0.557)<br>2190–2207 (0.023)<br>2228–2306 (0.420) | 2169 | Master core |
| 4A-1 | 182578 | 30.5 | Wood | 155 ± 25 | 0–36 (0.190)<br>68–118 (0.134)<br>131–154 (0.113)<br>166–231 (0.393)<br>244–284 (0.171) | 179 | Elephant dung site |

We built double-indexed, double-stranded DNA libraries using KAPA Library Preparation kits – Illumina (Roche Sequencing and Life Science, Kapa Biosystems, Wilmington, MA, United States) with iNext adapters (Glenn et al., 2019). To render the iNext stub compatible with the KAPA kit's A-tailing step, the stub sequence was modified by the addition of a thymine residue to the 3′-terminus (revised sequence: 5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGT-3′) and the stub complement sequence was extended by a guanine at the 5′-terminus (revised sequence: 5′-[phos]GACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG-3′). In addition to the previously described DNA samples and controls, we added four control reactions ("Library Controls") containing only water to monitor reagent contamination during library preparation. Libraries were amplified by 18 cycles of indexing PCR using KAPA HiFi Uracil + polymerase (Roche Sequencing and Life Science, Kapa Biosystems, Wilmington, MA, United States) according to the manufacturer's instructions. All purification steps during library preparation and amplification were performed using carboxyl paramagnetic beads (Rohland and Reich, 2012). Amplified libraries were visualized on 2% agarose gels stained with GelRed (Biotium Inc., Fremont, CA, United States) and quantified using a Qubit® 2.0 fluorometer with the dsDNA HS assay kit and quantitative PCR with the KAPA Library Quantification Kit (Roche Sequencing and Life Science, Kapa Biosystems, Wilmington, MA, United States). Libraries were pooled and submitted to Admera Health (South Plainfield, NJ, United States) where residual adapter-multimers were removed using AMPure XP beads (Beckman Coulter Life Sciences, Indianapolis, IN, United States) and final library pool quality was confirmed via visualization on a TapeStation (Agilent Technologies, Santa Clara, CA, United States). The quality-controlled pool was then 2 × 151 bp paired-end sequenced on a single lane of a Nextseq 500 (Illumina, Inc., San Diego, CA, United States).

## DNA Sequence Quality Control

Library sequence qualities were inspected using FastQC 0.11.5 (Andrews, 2016). We trimmed residual adapter contaminants and low-quality bases from the sequences using Trimmomatic 0.39 (parameters LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, MINLEN:30, HEADCROP:1, CROP:149, ILLUMINACLIP:NexteraPE-A-tail.fa:2:30:10 where Nextera-PE-A-tail.fa includes the A-tailing Nextera adapter sequences; Bolger et al., 2014). We merged paired reads using FLASH 1.2.11 (parameter -M 149; Magoč and Salzberg, 2011) and removed PCR duplicates from merged reads using CD-HIT-EST 4.6 (parameter -c 1; Li and Godzik, 2006).

## DNA Degradation and Cytosine Deamination

We estimated putative ancient DNA fragment length distributions from the deduplicated, merged reads using a previously developed custom Ruby script (Campana et al., 2014). While this estimate excludes a minority of library inserts longer than 288 bp, typical mean fragment lengths of ancient DNA are 150 bp or shorter (e.g., Green et al., 2009; Prüfer et al., 2010; Dabney et al., 2013a). We confirmed the accuracy of our fragment length estimates by inspection of library insert length distributions, both bioinformatically (**Supporting Data**) and by visualization on agarose gels.

We performed an *ad hoc* test to determine whether the Mubwindi Swamp sedimentary DNA demonstrated cytosine deamination patterns consistent with authentic ancient DNA preservation (Briggs et al., 2007). The sediment taxonomic profiles were dominated by Rhizobiales (see below). Many of these taxa, such as *Bradyrhizobium japonicum*, are plant symbionts providing functional benefits such as nitrogen fixation and nutrient provision (e.g., Kaneko et al., 2011; Erlacher et al., 2015). We observed intact plant tissues (roots, leaves and stems) primarily in the top layers of the sediment column. Therefore, we expect increased proportions of cytosine deamination in lower layers if "ancient" damaged Rhizobiales molecules are being preserved *in situ* (Briggs et al., 2007; Sawyer et al., 2012). Absence of an increasing cytosine deamination signal is consistent with these molecules originating from modern DNA, either from free-living taxa closely related to root-associated species or via leaching through the sediment column. To estimate the rates of terminal cytosine deamination, we aligned the deduplicated, merged reads against the *Bradyrhizobium japonicum* USDA6$^T$ reference genome (GenBank accession NC_017249.1; Kaneko et al., 2011) using BWA-MEM 0.7.17-r1188 (Li, 2013). The alignment was converted to bam format and sorted using SAMtools 1.9 (Li et al., 2009). Deamination patterns were analyzed using mapDamage2 2.0.8-dirty (Jónsson et al., 2013).

## Metagenomic Characterization

We aligned the deduplicated, merged reads against the National Center for Biotechnology Information (NCBI) non-redundant nucleotide (hereafter "nt," version dated 9 September 2019) and the Refseq Genomic (hereafter "Refseq," version dated 3 February 2020) databases using megaBLAST 2.6.0+ (Camacho et al., 2009) under default settings. MegaBLAST results were analyzed using the naïve lowest common ancestor (LCA) algorithm in MEGAN Community Edition 6.17.0 (nt analysis) or 6.18.5 (Refseq analysis) under default settings (Huson et al., 2016) except that "MinSupportPercent" was set to 0.005. We compared the two databases' read assignments using paired *t* tests in GraphPad QuickCalcs (GraphPad Software Inc.).[1] Relative metagenome compositions were compared using normalized counts to control for variation in sequencing depth, maintaining at least one read per identified taxon, and discarding unidentified sequences. Community compositions at the ordinal rank were compared by Principal Coordinates Analysis (PCoA) and neighbor-net analysis based on Bray-Curtis distance in MEGAN. We compared animal (Metazoa) and land plant (Embryophyta) taxa (ordinal level or higher) presence/absence in the sedimentary DNA record against regional species (Butynski, 1984; Kasangaki et al., 2003, 2008; Stanford and Nkurunungi, 2003; Olupot and Plumptre, 2010; Mugerwa et al., 2013; Decru et al., 2019) and pollen records (Marchant et al., 1997; Marchant and Taylor, 1998). Pollen data were obtained from the African pollen database hosted at ftp://ftp.ncdc.noaa.gov/pub/data/paleo/pollen/tiliafiles/apd/ (accessed 3 November 2019). We excluded samples with fewer than 50,000 unique reads from the animal and plant record comparisons as these produced too few eukaryotic sequences for accurate identification.

To better characterize the sediment microbial communities, we also analyzed the deduplicated, merged reads using MetaPhlan2 2.9.21 (Truong et al., 2015) and QIIME 2 2019.7 (Bolyen et al., 2019) following Ferrari et al. (2018). We performed MetaPhlAn2 analyses under default settings and generated heat maps clustering sediment samples and taxa using Euclidean distances. In the QIIME 2 analyses, we closed-reference clustered the sequences against a previously trained (using QIIME 2 2018.4) SILVA 16S database (build 132; Pruesse et al., 2007) at 99% identity using VSEARCH (Rognes et al., 2016). Clustered sequences were aligned with MAFFT (Katoh and Standley, 2013), and a phylogenetic tree was built using FastTree (Price et al., 2010). Both phylogenetic and non-phylogenetic diversity metrics were calculated with rarefaction to 50 and 500 sequences. Communities were compared using PCoA using Bray-Curtis and Jaccard distances in EMPeror (Vázquez-Baeza et al., 2013, 2017).

## Statistical Analyses

We examined the relation between DNA yield and environmental factors with exponential regression analysis and the linear dependence between multiple DNA quantity and degradation parameters by computing Pearson correlation coefficients in R 3.5.1 (R Core Team, 2018) using the PerformanceAnalytics 1.5.3 package (Peterson et al., 2019). Since we did not have an *a priori* model to explain DNA degradation in the Mubwindi Swamp sediments, we also assessed non-parametric relationships

[1] www.graphpad.com

between environmental parameters and DNA yield, terminal deamination, and sequence length using Spearman's ρ in R.

# RESULTS

## Core Stratigraphy, Chemistry, and Age

The master core (MUB17-2C/D) contains a stratigraphic sequence which grades from clastic to organic sediment from bottom to top and is divided into six stratigraphic units based on patterns of sediment composition, texture and geochemistry (**Figure 4**). The core bottom from 656 to 653 cm is an unsorted gravel layer (unit I), high in density (1.9 g $cm^{-3}$) and magnetic susceptibility, but low in water (<20%) and organic matter content of the dry mass fraction (<10%). High XRF counts were obtained for Ca, S, Al, and Fe (**Figure 5**). Unit II is a coarse sand layer from 653 to 642 cm containing wood and charcoal, but overall less than 10% organic matter. Its C (2.6%) and N (0.1%) content were the lowest of the entire core record, while XRF peaks occurred for Si, Al, Ti and Zn. The sediment type changes abruptly at 642 cm to a homogenous silty clay that continues until 563 cm (unit III). This sediment had, on average, a water content of ca. 50%, a dry mass organic matter content of ca. 20% and a C content of <10%, whereas average density is 1.3 g $cm^{-3}$. The major elements recorded by XRF are K, Ti, Fe and Al. Unit IV from 563 to 434 cm is a clay deposit with numerous horizontal layers of charcoal, buried leaf fragments and some smaller wood pieces. Its sediment had a density of ∼1 g $cm^{-3}$, an average sediment water content of 78%, a dry mass organic content of about 37%, a C content < 20% and a N content < 1%. The XRF elemental composition was similar to that of the previous unit, but showed higher counts for S, Co, and Ni. The sediment also contains pyrite and other Fe minerals. Unit V, which extends from 434 to 312 cm, is an organic-rich clay with nearly 60% average dry mass organic matter content and mean sediment water content of 86%. This sediment composition was generally reflected in slightly negative values for magnetic susceptibility, except for a high magnetic susceptibility interval between 372 and 340 cm corresponding to major peaks in XRF counts for Fe and Co and the occurrence of Fe oxides. In addition, relatively high XRF counts of S and a maximum S content of nearly 4% characterized this unit, consistent with the frequent occurrence

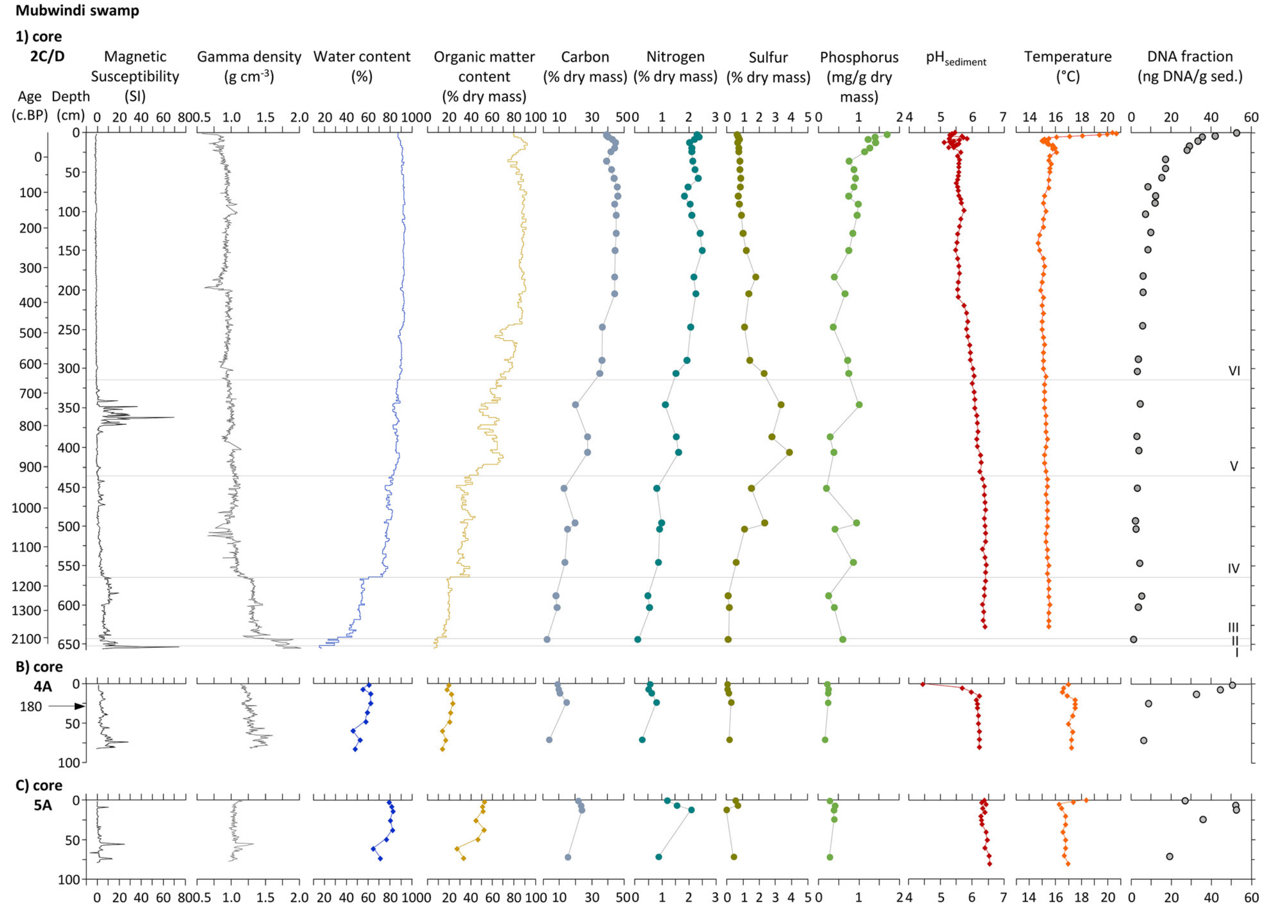


FIGURE 4 | Physico-chemical sediment stratigraphy of cores MUB17-2C/D (master), -4E (dung), and -5A (wallow) together with the DNA fraction of the sediment and a calibrated age scale (years cal. BP) for core 2C/D. pH and temperature were measured in the field close to the original coring locations. Roman numerals denote stratigraphic units of the master core, which are separated by horizontal lines.

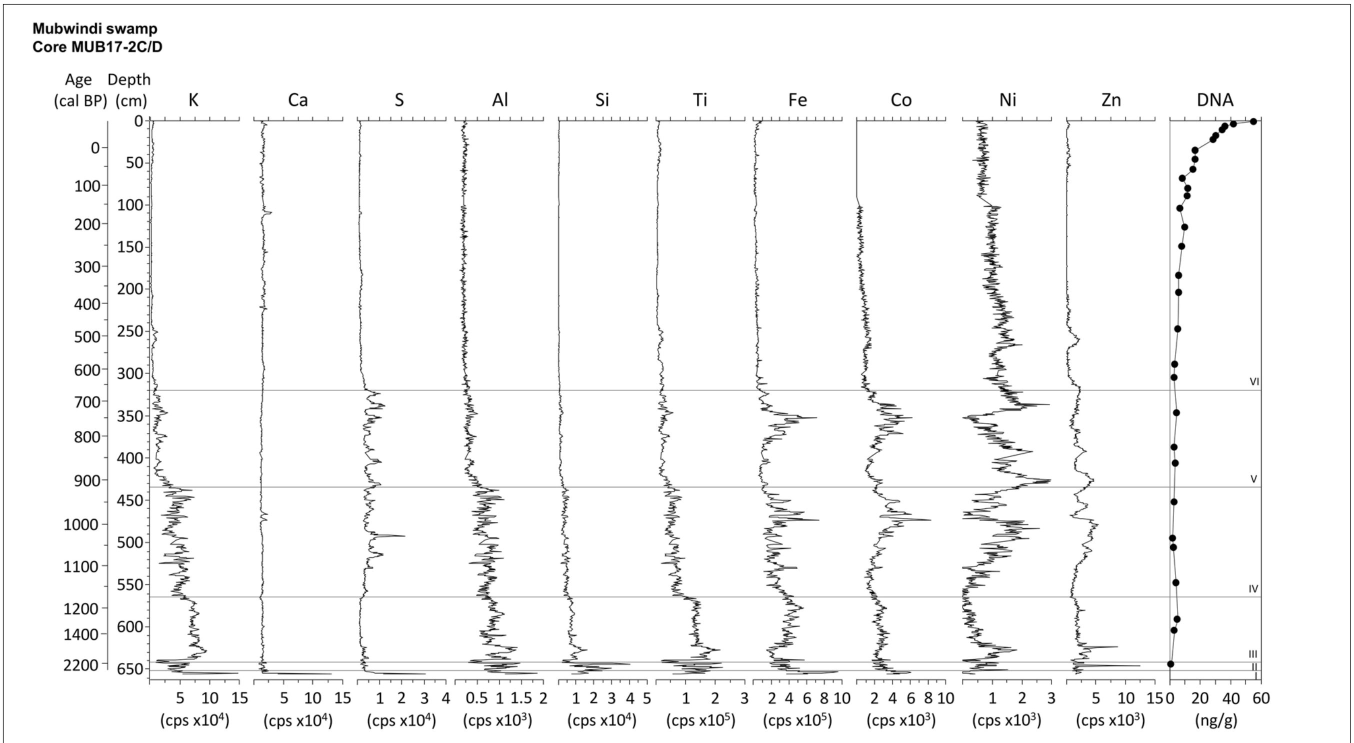


**FIGURE 5 |** XRF elemental composition of core MUB17-2C/D (master) in comparison to DNA yield. XRF units are counts per second (cps). Roman numerals denote stratigraphic units, which are separated by horizontal lines.

of pyrite. The final unit VI from 312 to 0 cm is water-saturated, sedge peat (water content > 90%) with an average organic matter content of the dry mass of 84%, a wet bulk density of 0.9 g $cm^{-3}$ and consistently negative magnetic susceptibility values. This stratigraphic unit has the highest C (average 42%) and N (average 2.1%) content of all units. The P content rises toward the top of the unit and is between 1.1 and 1.7 mg $g^{-1}$ dry mass in the upper 25 cm. The peat consists mostly of sedge rootlets and contains few small wood pieces (**Figure 3**). Nearly all XRF elements show significantly lower counts in the XRF profiles (except Ni) and subdued variability in this unit. The stratigraphic units III–V (clay sediments) are interpreted to represent a shallow lake environment, which filled with sediment to form a peatland (unit VI). In contrast to the master core, the short cores from the swamp's edge (4A, 5A, **Figure 3**) consist entirely of silty clay and the dry mass contained only between 15 and 50% organic matter. The elephant dung site (4A) had a water content of ca. 50–60% and about 20% dry mass organic matter. The elephant wallow core (5A) exhibited signs of bioturbation and consists of about 70–80% water and the dry mass had about 50% organic matter in the top 50 cm and about 30% below that depth. The most abundant element counts of the XRF analysis for these cores were Fe, Ti, and K.

## Age of Sediments

The radiocarbon chronology of site 2 (master core) showed that this central part of Mubwindi Swamp was an accumulating basin since ∼2200 cal BP (**Table 1**). However, the difference of nearly 1000 years between the two deepest radiocarbon ages of the master core, less than 50 cm apart, and the abrupt change in sediment from coarse sand to clay at 642 cm (transition from unit II to III) suggests a possible erosional surface and an associated depositional hiatus. Marchant et al. (1997) also found episodic sedimentation and several hiatuses in their cores. We therefore implemented a hiatus at 642 cm in the BACON age model calculation (**Supporting Data**). The resulting age-depth model shows relatively constant and very fast deposition since ∼1320 cal BP (642–0 cm). The associated average long-term sedimentation rate is 4.9 mm/yr, which means that one centimeter of sediment was formed in only about two years. The master core therefore provides a temporal high-resolution record: excluding the oldest DNA sample, the average time between adjacent DNA samples is ∼50 years, decreasing to ∼20 years in the upper one meter (**Figure 3**). The elephant dung site at the edge of the swamp has a median age of ∼180 cal BP (∼1770 CE) at a depth of 30 cm, indicating a slower sedimentation rate of ca. 1 mm/yr at the swamp's edge (**Table 1**).

## Physico-Chemical Soil and Water Properties

Mubwindi Swamp is an acidic swamp in a relatively cool tropical climate. Daytime soil temperatures in the sediment profile of site 2 (master core) ranged from maximum 20.6°C (0 cm) to 14.6°C (150 cm) (**Figure 4**). The average soil temperature of the entire 630 cm profile was 15.5°C – very similar to mean monthly air temperatures. Within the upper 10 cm of the sediment column the soil temperature declined rapidly by over 5°C, then continued to drop slightly until 140 cm, but below this depth increased slightly to 15.4°C (**Figure 4**). At the elephant dung site (core 4A) soil temperatures showed subdued changes, varying between 17.5 and 16.5°C and lacking a maximum at the surface, which instead was recorded at 20–30 cm. In contrast, the elephant wallow site (core 5A) had a temperature maximum at the surface (18.4°C) whereas below 10 cm temperature remained relatively constant at ca. 17°C.

Soil acidity along a 630 cm deep sediment column of the master core site ranged from pH 5.1 to 6.4 (**Figure 4**). The peat in the upper 20 cm is acidic (average pH 5.4) and showed the greatest variation in pH values. Between 25 and 210 cm, constantly moderately acidic conditions prevailed (pH of ~5.6) and below that depth pH continued to rise; the sediment was only slightly acidic below 330 cm ($\geq$ 6.1). The surface of the elephant dung site (core 4A) had a pH of 4.4 and became only weakly acidic below 10 cm depth ($\geq$6.1). The elephant wallow site (core 5A) exhibited very similar acidity of between pH 6.3 and 6.7 (slightly acidic to neutral) (**Figure 4**).

The low concentration of Ca and the moderately acidic pH values of the surface water classify Mubwindi Swamp as a transitional poor to intermediate fen (i.e., groundwater-fed peatland; **Table 2**). The water was generally depleted in nutrients, however the master core site (2) was slightly enriched in dissolved reactive phosphorus (SRP), ammonium ($NH_4^+$), and potassium (K). Both swamp center sites (1 and 2) were very rich in dissolved organic carbon (DOC). At the swamp edge (sites 4, 5), the surface water had elevated sulfate concentrations.

## DNA Content, Preservation, and Authenticity

All sediment samples contained Qubit-quantifiable DNA (range: 0.9-86.1 ng DNA/g wet sediment; Mean ± SD: 22.2 ± 21.1 ng/g). Qubit-quantifiable DNA was not detected in negative controls verifying that this DNA derives from the sediments and not contaminants. Between 193 and 24,239,573 read pairs (Mean ± SD: 2,840,581 ± 5,273,296 read pairs) were sequenced for each sediment library, yielding between 126 and 15,458,842 unique merged reads per library (Mean ± SD: 1,859,731 ± 3,425,701 unique reads). Between 1 and 46,643 read pairs (Mean ± SD: 14,440 ± 15,484 read pairs) were sequenced for each negative control, yielding between 0 and 76 unique merged reads per library (Mean ± SD: 19 ± 25.8 unique reads). These low numbers of unique reads in the blanks indicated that our results were not strongly biased by reagent contamination.

We observed a clear trend of decreasing quantities of DNA with greater depths (older age) of the cores (**Figures 4**, **6**, **7**), consistent with the preservation of ancient DNA *in situ*. The surface samples ($n$ = 7) of all our cores contained between 28.6 and 86.1 ng DNA/g sediment (Mean ± SD: 53 ± 17 ng/g, **Figure 6**). In the master core the amount of sedimentary DNA exponentially declined with depth ($r^2 = 0.755$, $p < 0.001$) from a maximum amount of 54.7 ng/g at the surface (0–2 cm) to 0.9 ng/g at 645 cm in the basal sand (**Figure 7**). The largest decline in DNA mass occurred in the upper 100 cm with reductions of more than 10 ng/g between adjacent samples from 2 to 3 cm and 20 to 30 cm. Below 100 cm, the amount of DNA was consistently less than 10 ng/g sediment and, below 250 cm (sediment older than 500 years), it was generally less than 5 ng/g ($\leq$10% of surface sample). Below this depth the mass of DNA remained relatively constant at an average fraction of 3 ng/g despite substantial changes in sediment type, water- and organic matter content and geochemistry in this part of the core (**Figure 4**).

The trend of rapidly declining DNA fraction in the upper meter was persistent across Mubwindi Swamp despite the obvious differences in sediment type and composition between swamp center and edge, i.e., peat vs. silty clay (**Figure 6**). In core 4A (elephant dung site) DNA declined exponentially ($r^2 = 0.683$, $p = 0.5322$) from 52.9 at 0–2 cm to 7.0 ng/g at 70 cm depth (**Figures 4**, **6**). In contrast, the surface sample of core 5A (wallow site) contained only half as much DNA as the samples from 5 to 7 and 10 to 12 cm which have 55 ng DNA/g sediment. Only below this depth did the DNA concentration decline, but the amount was still larger than in samples of similar depths in the other cores (**Figure 6**). Additional core samples are required to determine whether this variation is a result of local environmental conditions or simply variability in DNA taphonomy, which has been shown to differ even within a single bone (Green et al., 2010).

Fragment lengths and cytosine deamination patterns also supported the persistence of authentic ancient DNA in the sediment. Mean fragment lengths ranged from 127 to 175 bp (Mean ± SD: 146 ± 13 bp) and molecular fragment lengths were highest in the upper 30 cm of the cores (**Figure 6**). In the master core, fragment lengths declined with depth (exponential decline $r^2 = 0.543$, $p < 0.001$), but not as strongly as the DNA fraction. Conversely, both 5′ C→T and 3′ G→A increased exponentially with sediment depth and age in this core (5′ C→T: $r^2 = 0.519$, $p < 0.001$; 3′ G→A: $r^2 = 0.601$, $p < 0.001$), strongly suggesting that a portion of the sediment sequences derived from preserved aDNA. The quantity of DNA and fragment length both

**TABLE 2 |** Surface water chemistry parameters for the four coring locations.

| Site # | Temp. °C | pH | EC $\mu$S cm$^{-1}$ | SRP mg l$^{-1}$ | $NH_4^+$ mg l$^{-1}$ | DOC mg l$^{-1}$ | $Cl^-$ mg l$^{-1}$ | $NO_3^-$ mg l$^{-1}$ | $SO_4^{2-}$ mg l$^{-1}$ | Al mg l$^{-1}$ | Ca mg l$^{-1}$ | Fe mg l$^{-1}$ | K mg l$^{-1}$ | Mg mg l$^{-1}$ | Mn mg l$^{-1}$ | Na mg l$^{-1}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 Center | 14.4 | 6.3 | 39 | 0.06 | 0.06 | 33.2 | 2.1 | 0.01 | 0.3 | 0.06 | 2.8 | 2.6 | 2.0 | 1.9 | 0.2 | 1.4 |
| 2 Master core | 13.8 | 5.4 | 94 | 0.40 | 0.18 | 56.3 | 6.0 | <0.01 | 1.5 | 0.17 | 6.0 | 6.2 | 3.6 | 3.2 | 0.8 | 1.5 |
| 4 Dung | 21.2 | 5.9 | n.d. | <0.003 | 0.06 | 6.6 | 1.9 | <0.01 | 13.7 | <0.01 | 5.0 | 2.2 | 1.5 | 3.3 | 0.67 | 1 |
| 5 Wallow | 20.7 | 6.0 | n.d. | <0.003 | 0.09 | 3.3 | 4.7 | 0.08 | 8.9 | 0.02 | 3.0 | 0.3 | 0.8 | 2.3 | 0.48 | 2 |

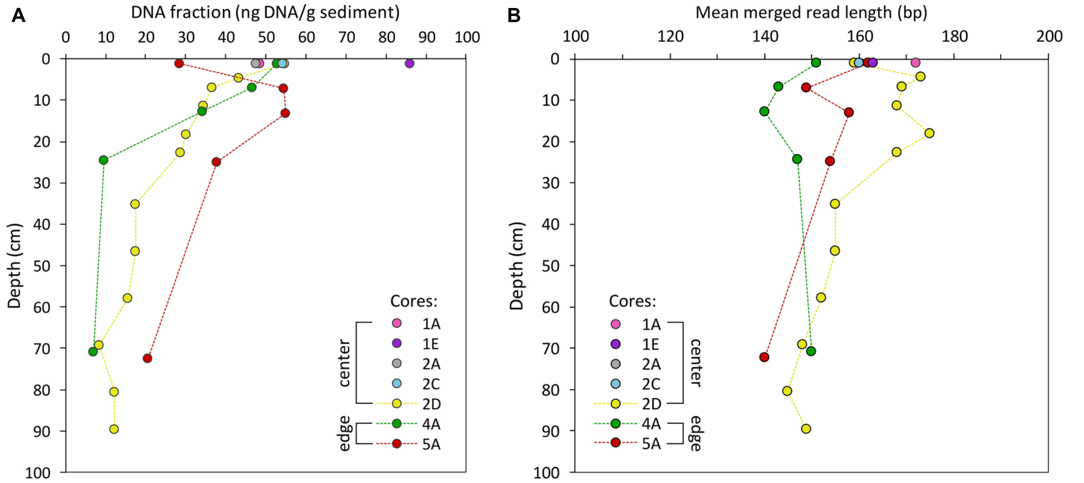


**FIGURE 6 |** DNA yield of surface samples (cores: MUB17-1A, -1E, -2A, -2C) and samples of the first meter of cores MUB17-2D, -4A, and -5A. **(A)** DNA fraction per gram sediment, **(B)** DNA molecule fragmentation presented as mean merged DNA molecule length (bp).

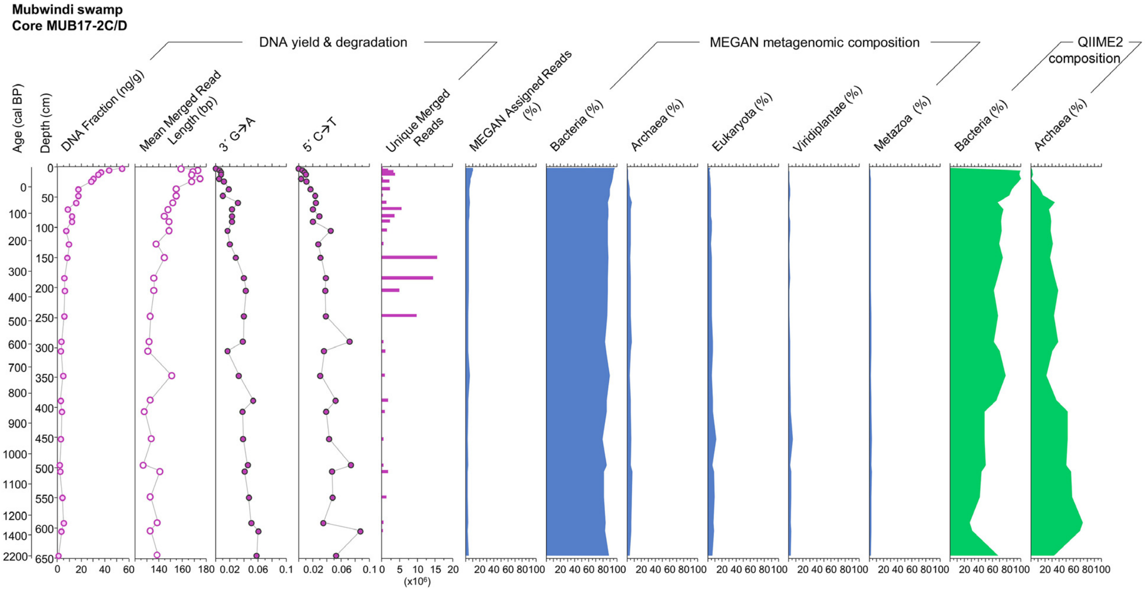


**FIGURE 7 |** DNA quantity, degradation patterns and metagenomic compositions derived from the MEGAN and QIIME2 pipelines for core MUB17-2 (master core) on a depth and calibrated age scale. DNA degradation is indicated by DNA fragmentation (mean merged read length) and 5′ C→T and 3′ G→A substitutions. The percentage of MEGAN-assigned reads is calculated as the number of MEGAN-assigned reads divided by the number of unique merged reads. The low percentage emphasizes the large number of unidentifiable reads due to missing coverage in genomic databases. Similarly, the percentage of each taxonomic unit presented for the MEGAN metagenomic composition corresponds to the number of taxon-assigned reads divided by the total number of MEGAN-assigned reads. Viridiplantae (green plants) and Metazoa (animals) are clades within the Eukaryota domain. Note the differences in the proportion of Bacteria and Archaea between MEGAN and QIIME2.

decreased down core and were significantly correlated ($r^2 = 0.701$, $p < 0.001$), whereas both parameters were significantly negatively correlated with both 5′ C→T and 3′ G→A as shown in **Figure 8**.

## MEGAN Metagenomic Characterization

The vast majority (nt database: 87.1–97.3%; Refseq database: 80.8–95.5%) of sediment sequences were not identifiable via megaBLAST and MEGAN analysis (nt: Mean ± SD: 94.1 ± 2.6%; Refseq: Mean ± SD: 90.9 ± 3.7%; **Figure 7**; **Supporting Data**). MEGAN-identified sequences were dominated by Bacteria (nt: range 79.0–96.9%, Mean ± SD: 88.2 ± 4.4%; Refseq: range 79.5–99.0%, Mean ± SD: 90.7 ± 4.4%), particularly Rhizobiales (nt: range 7.1–31.3%, Mean ± SD: 16.5 ± 4.8%: Refseq range 9.0–33.3%, Mean ± SD: 15.5 ± 4.8%) and Burkholderiales (nt: range 0.0–15.0%, Mean ± SD: 7.2 ± 3.0%; Refseq: range 1.3–14.1%, Mean ± SD: 6.2 ± 2.8%). The microbial profiling conducted in this project was undertaken to further our understanding of DNA preservation in this environment. So, while a full description of microbial profiles is beyond the scope of this paper, detailed microbial profiles and MEGAN results are available in the **Supporting Data**. The bacterial profiles were complex, representing 125 (nt database) to 156 (Refseq database) taxonomically described orders across all samples (**Supporting Data: Taxonomic Profiles**). Archaea, representing 15 (Refseq database) to 16 orders (nt database), comprised 0.0–7.5% of the nt assigned reads (0.0–6.9% of Refseq assigned reads) and increased with frequency with depth (nt: Mean ± SD: 4.2 ± 2.1%, $r^2$ = 0.69; Refseq: Mean ± SD: 3.4 ± 1.8%, $r^2$ = 0.28). The most frequent Archaea were Methanobacteriales (nt: range 0.0–4.3%, Mean ± SD: 1.1 ± 1.0%; Refseq: range 0.0–4.8%, Mean ± SD: 0.5 ± 1.0%), Methanomicrobiales (nt: range 0.0–7.4%, Mean ± SD: 1.0 ± 1.2%; Refseq: range 0.0–6.3%, Mean ± SD: 0.9 ± 1.0%), and Methanosarcinales (nt: range 0.0–5.7%, Mean ± SD: 0.9 ± 1.1%; Refseq: range 0.0–3.4%, Mean ± SD: 0.7 ± 0.8%). Eukaryotes were underrepresented, comprising only 0.0–13.8% of the identified reads in total (nt: range 0.0–11.5%, Mean ± SD: 4.5 ± 2.3%; Refseq: range 0.0–13.8%, Mean ± SD: 4.2 ± 2.8%) with 0.0–5.8% (nt: range 0.0–3.1%, Mean ± SD: 1.6 ± 0.8%; Refseq: range 0.0–5.8%, Mean ± SD: 2.5 ± 1.6%) deriving from Metazoa and 0.0–6.9% (nt: range 0.0–5.8%, Mean ± SD: 1.4 ± 1.1%; Refseq: range: 0.0–6.9%, Mean ± SD: 1.2 ± 1.3%) from Viridiplantae (**Figure 7**). Other than the relative decrease in Bacteria and corresponding increase in Archaea with depth, we observed little evidence of temporal structure at an ordinal scale (**Supporting Data: Taxonomic Profiles**). The relative increase in methanogenic Archaea between strata may not necessarily indicate changes in the past community structure, but rather current structure due to differing microbiota in deeper Mubwindi Swamp sediments (Vuillemin et al., 2017). PCoA and

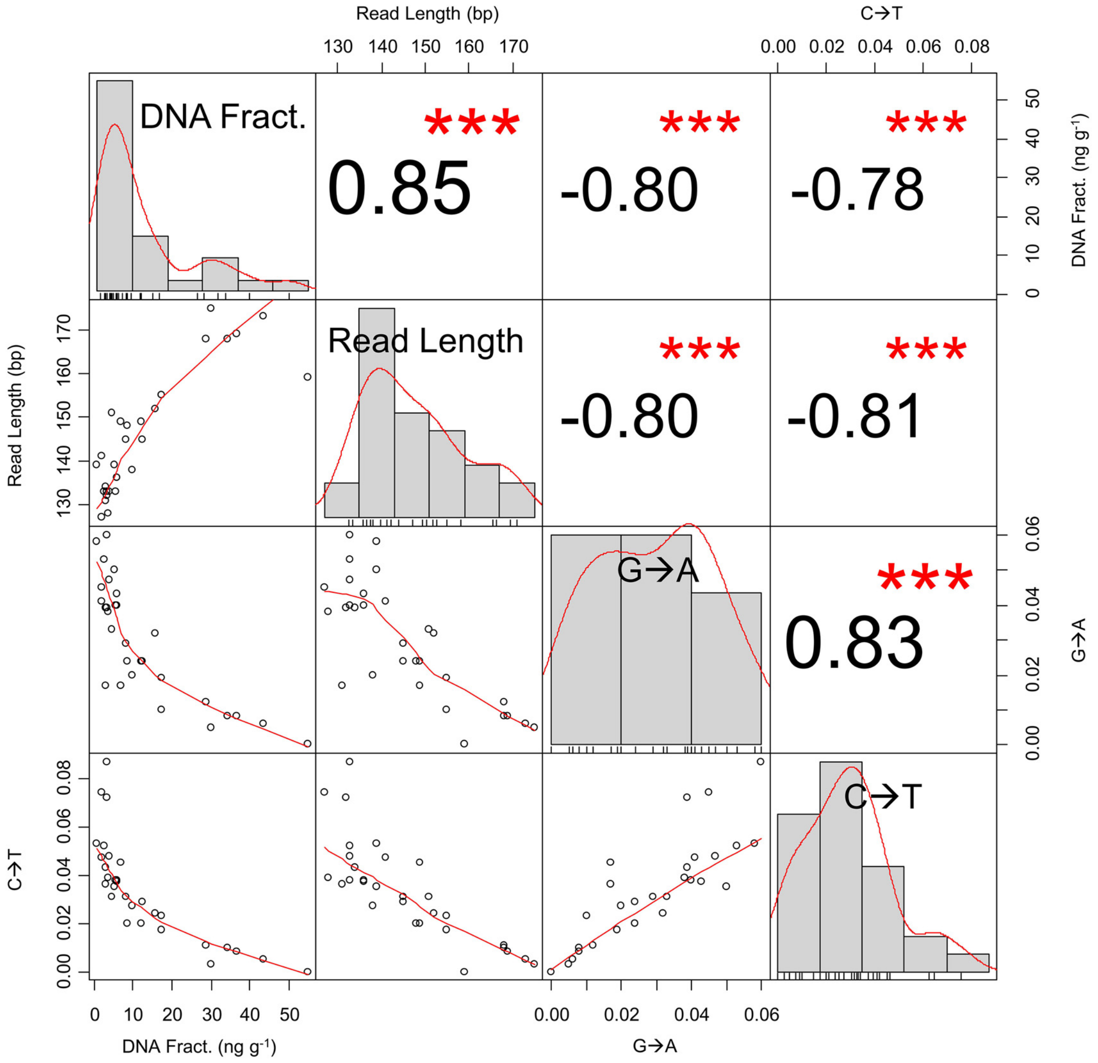


**FIGURE 8 |** Correlation matrix of DNA fraction, mean read length, and terminal cytosine deamination from core MUB17-2C/D. The histograms depict the distribution of sample values. All correlations (*r*) are highly significant with estimated *p*-values of 0 (denoted by "***").

neighbor-net analysis separated some of the surface samples from the short cores (samples 57, 59, 70, and 100) from the primary core samples. To a lesser extent, the short core samples 54, 56, 58, 60, 62, and 63 were also distinct. While these results were consistent with local community structure variation due to microenvironmental differences, the distinct samples were also the shallowest sequenced (120–37,261 unique sequences per sample), indicating that this pattern was probably a sampling artifact (**Supporting Data**).

The Refseq-aligned dataset assigned taxa to more sequences than the nt-aligned dataset (nt: Mean ± SD: 90,911 ± 150,070 sequences; Refseq: Mean ± SD: 142,931 ± 238,131 sequences; two-tailed $p = 0.0003$). The Refseq-aligned dataset assigned significantly more sequences to Eukaryota (nt: Mean ± SD: 4,042 ± 6,928 sequences; Refseq: Mean ± SD: 5,919 ± 10,897 sequences; two-tailed $p = 0.0036$), Bacteria (nt: Mean ± SD: 80,718 ± 132,524 sequences; Refseq: Mean ± SD: 129,720 ± 215,085 sequences; two-tailed $p = 0.0003$), and Archaea (nt: Mean ± SD: 3,598 ± 6,906 sequences; Refseq: Mean ± SD: 4,578 ± 8,665 sequences; two-tailed $p = 0.0013$). While the Refseq-aligned dataset assigned more reads to both Metazoa (nt: Mean ± SD: 1,569 ± 2,843 sequences; Refseq: Mean ± SD: 3,810 ± 7,390 sequences; two-tailed $p = 0.0022$) and Viridiplantae (nt: Mean ± SD: 1,038 ± 1,669 sequences; Refseq: Mean ± SD: 1,187 ± 1,911 sequences; two-tailed $p = 0.0041$) than the nt-aligned dataset, the effect was larger for Metazoa (mean 2.43 × more assigned sequences) than Viridiplantae (mean 1.14 ×). In fact, the increase in Metazoa was greater than the increase in total Eukaryota (mean 1.46 ×), likely reflecting the disproportionate number of animal genomes in the Refseq database (e.g., Brandies et al., 2019).

The taxonomic compositions of our MEGAN results were not significantly impacted by reagent contamination. The low contamination level (151 total unique reads in the negative controls) were consistent with rare cross-contaminations between sediment samples and negative controls and sporadic reagent contamination. For instance, the *Schistosoma*, *Mustela* and Percomorphaceae reads (see below) were probably the result of rare cross-contaminations from the sediment samples (e.g., via index switching: Kircher et al., 2012) as none are found frequently in reagents or were processed previously in our ancient DNA laboratory. Since our analyses were limited to high taxonomic ranks unlikely to be significantly biased by these rare contaminations, we list identified contaminants, rather than bias our dataset by excluding these taxa. Using the nt database, the majority of MEGAN-identified sequences ($n$ = 32 of 37 assigned sequences; 86%) in the negative controls derived from common laboratory contaminants including Hominoids ($n$ = 14), of which 12 were identified as human (*Homo sapiens*), Canids ($n$ = 1), and Bacteria ($n$ = 17). Identified bacterial species and strains in the blanks included *Acinetobacter johnsonii* XBB1 ($n$ = 1), *Alcaligenes faecalis* ($n$ = 1), *Aquaspirillium* sp. LM1 ($n$ = 1), *Brevundimonas* sp. ($n$ = 1), *Citrobacter freundii* complex ($n$ = 1), *Dietza* sp. oral taxon 368 ($n$ = 1), *Fusobacterium* sp. ($n$ = 1), *Pseudomonas putida* ($n$ = 1), *Salincola tamaricis* ($n$ = 1), *Sphingomonas hengshuiensis* ($n$ = 1), *Staphylococcus saprophyticus* subsp. *saprophyticus* ($n$ = 1), and *Stenotrophomonas* sp. ($n$ = 1). One contaminant sequence was identified as *Schistosoma japonicum*, and another three were identified as Percomorphaceae [including *Dicentrarchus labrax* ($n$ = 1), *Mastacembelus armatus* ($n$ = 1), and *Scophthalmus maximus* ($n$ = 1)]. Using the Refseq database, MEGAN assigned taxa to 62 contaminant sequences, of which 33 derived from Bacteria. These included *Acinetobacter* sp. ($n$ = 1), *Alcaligenes faecalis* subsp. *faecalis* NBRC 13111 ($n$ = 1), *Bradyrhizobium* sp. LSPM299 ($n$ = 1), *Brevundimonas bullata* ($n$ = 1), *Dietzia* sp. ($n$ = 1), *Egibacter rhizosphaerae* ($n$ = 1), *Fusobacterium* sp. ($n$ = 1), *Henriciella litoralis* ($n$ = 1), *Methylobacterium pseudosasicola* ($n$ = 1), *Microbacterium* sp. CGR2 ($n$ = 1), *Pedosphaera pravula* Ellin514 ($n$ = 1), *Pseudomonas* sp. ($n$ = 2), *Sphingobium phenoxybenzoativprans* ($n$ = 1), *Sphingomonas* sp. ($n$ = 1), *Staphylococcus* sp. ($n$ = 2), *Streptomyces griseus* subsp. *griseus* ($n$ = 1), *Terracidiphilus gabretensis* ($n$ = 1), and *Williamsia* sp. ($n$ = 1). Eukaryotic contaminants identified using the Refseq database included Percomorphaceae ($n$ = 7, including 1 assigned to *Lates calcarifer*), Primates ($n$ = 14, including 8 assigned to Homininae), Canidae ($n$ = 4, including 2 assigned to *Canis lupus*), and *Mustela putorius furo* ($n$ = 1).

## Plant and Animal Taxonomic Assignments

For land plants (Embryophyta) our MEGAN analysis against the nt database resulted in 18 taxonomic assignments, with 11 at ordinal rank. All 18 taxa were recorded in the master core, in which the richest samples occurred below 340 cm depth and single samples contained between 1 and 7 orders (**Figure 9**). The most commonly detected orders were in descending frequency Poales (100% presence), Fabales, Solanales, Rosales, and Brassicales. In comparison, the Refseq database yielded 21 plant taxonomic assignments with 14 orders. All these taxa were recorded in the master core, in which ordinal richness ranged from 0 to 10 orders per sample and was also highest below 340 cm depth (**Figure 9**). Fabales, Solanales, Poales, and Rosales were the most frequently detected plant orders with Refseq. Refseq identified all the taxa that were identified with the nt database, but in addition also Asparagales (1×), Gentianales (1×), and Caryophyllales (3×) (**Figure 9**).

Using the nt database, MEGAN assigned sedDNA sequences to 26 different animal taxa (Metazoa) of which 9 are at the level of order (**Figure 10A**). The most frequently recorded orders are Diptera (in all samples), Primates, Hymenoptera, and Cetartiodactyla. All of the detected taxa occur at present in Africa and members of all recorded orders in Bwindi Impenetrable Forest. In the master core between 1 and 5 orders/sample were detected and the number of assigned orders increased slightly with depth (age) (**Figure 10A**). In contrast, the Refseq database assigned sequences to 41 animal taxa representing 15 orders. Diptera, Hymenoptera, and Rodentia were detected in all samples, followed in frequency by Primates, Carnivora and Cetartiodactyla. In the master core, the number of detected orders per sample ranged from 3 to 13, with the minimum recorded

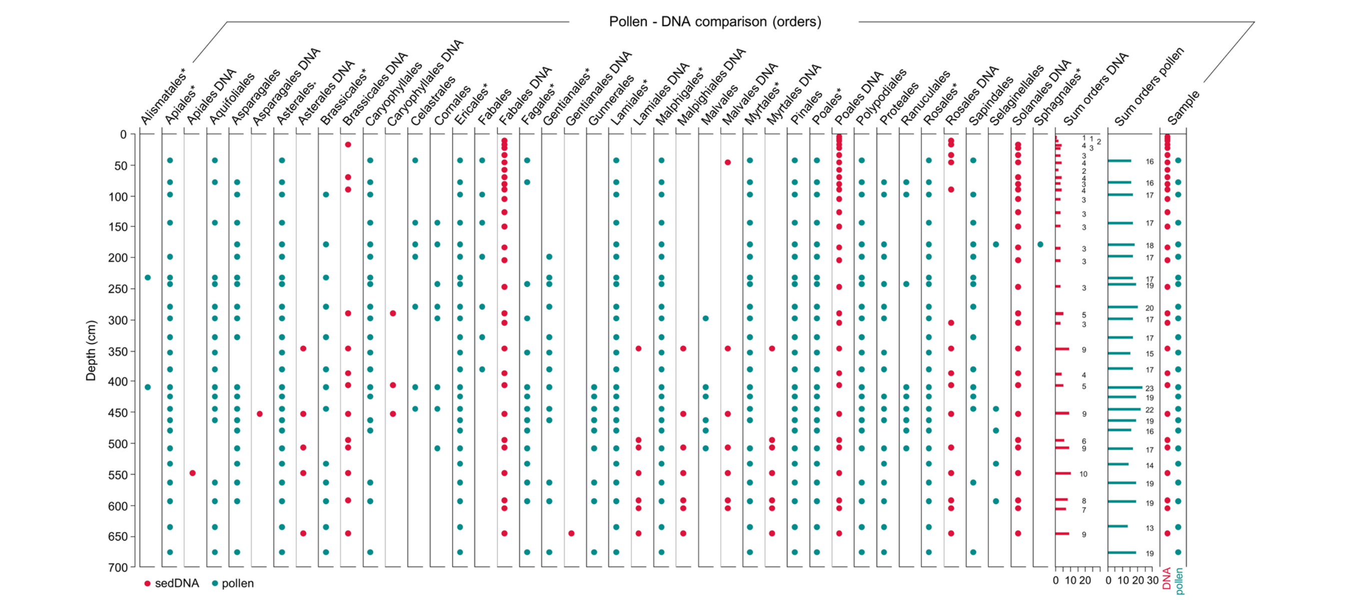


**FIGURE 9 |** Comparison of pollen and DNA record from Mubwindi Swamp. Plant DNA is from the master core MUB17-2D/C. DNA taxonomic assignments are based on both nt and Refseq output. The pollen data are from core MUB 3 of Marchant et al. (1997) and pollen types have been combined into their corresponding plant orders. Plant orders that include local pollen taxa are denoted by an asterisk. Each dot represents a record in the sediment core, red dots DNA and cyan dots pollen.

at 4 cm and the maxima between 300 and 500 cm depth (**Figure 10B**). Four assigned taxa (Protacanthopterygii, Metatheria (= marsupials), Cetacea, Octopoda) do not occur in (tropical) Africa and we consequently considered them misidentifications. Cetacea and Octopoda were only detected once but Metatheria was detected regularly in the master core (**Figure 10B**).

The taxonomic assignments for Metazoa from both reference databases had 22 taxa with the following seven orders in common: Cypriniformes, Rodentia, Primates, Cetartiodactyla, Hymenoptera, Lepidoptera, and Diptera. Excluding misidentifications, Refseq uniquely identified the orders Squamata, Passeriformes, Eulipotyphla, Carnivora, Chiroptera, Hemiptera, and Coleoptera whereas the nt database uniquely yielded Cyprindontiformes and Rhabditida. Together both databases detected a total of 45 assigned animal taxa of which 17 are at ordinal rank.

## Microbial Composition Using MetaPhlAn2 and QIIME 2

MetaPhlAn2 only identified a total of 11 bacterial ($n = 7$) and archaeal ($n = 4$) species/strains representing 5 bacterial and 1 archaeal orders. No taxa were identified for 30 of the sediment samples or for any of the negative controls (**Supporting Data**). We did not analyze these data further given that they were uninformative.

QIIME 2 identified between 0 and 1823 unique 16S sequences per sediment sample (Mean ± SD: 238 ± 412 sequences; **Supporting Data**). No 16S sequences were identified in the negative controls. Excluding samples with no retained reads, 27.3–100.0% were assigned to Bacteria (Mean ± SD: 74.6 ± 20.2%) and 0.0–72.7% were assigned to Archaea (Mean ± SD: 25.4 ± 20.2%). As in the MEGAN analyses, the percentage of reads assigned to Archaea increased with depth ($r^2 = 0.72$). Community compositions were exceptionally complex, with no taxa dominating the assignments (**Supporting Data**). In the dataset rarified to 50 sequences, PCoA separated the samples from the first meter of sediment from the remaining sediment samples (**Supporting Data**). PCoA did not reveal informative patterns in dataset rarified to 500 sequences due to the low sequencing depth eliminating most samples ($n = 8$ after rarefaction; **Supporting Data**).

# DISCUSSION

Sedimentary ancient DNA has been proposed as a powerful tool in conservation biology and paleoecology (e.g., Boessenkool et al., 2014; Rawlence et al., 2014; Epp, 2019). However, many aspects on the production, dispersal, deposition and accumulation of DNA in sediments of terrestrial ecosystems and on the taphonomy of sedaDNA remain to be studied (Epp, 2019). This need is particularly true for the tropics where most of the world's species occur. Our study highlights challenges and opportunities of sedaDNA approaches for reconstructing tropical biodiversity which we identify from the investigation of multiple profiles from a tropical swamp located within an exceptionally diverse African rainforest. The cores from Mubwindi Swamp allow us to examine sources of DNA, its taxonomic composition and the conditions of DNA preservation in different types of sediments. Our integrative approach allows for a comprehensive view of taphonomic

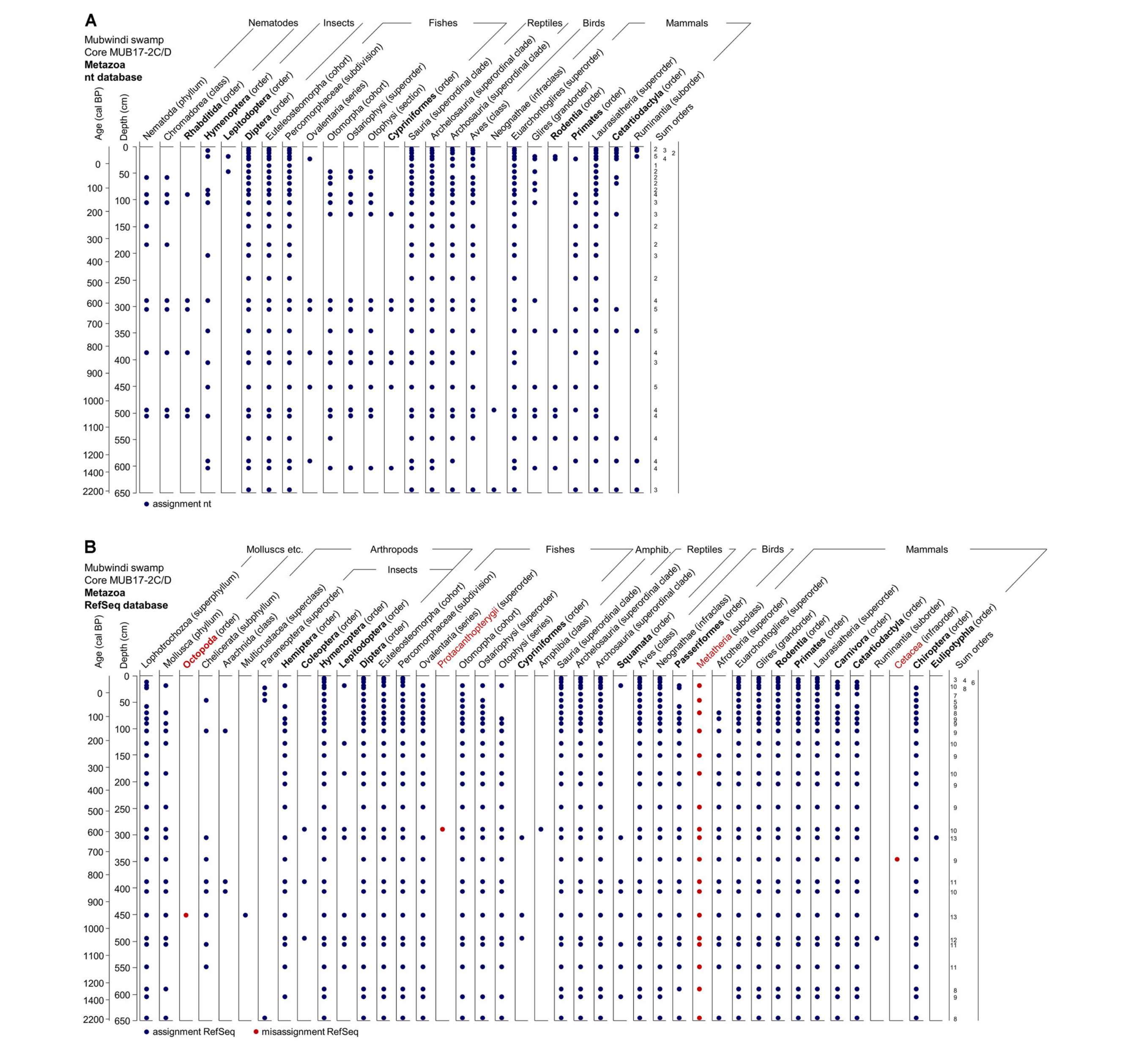


**FIGURE 10 |** Metazoa taxa observed in DNA preserved in the Mubwindi Swamp sediments. Metazoa DNA is from the master core MUB17-2D/C. DNA taxonomic assignments are based on both nt **(A)** and Refseq **(B)** output. Orders are in bold. Each dot represents a record in the sediment core. Taxa in red indicate likely taxonomic misassignments since these taxa are unknown in the area.

processes, and cross-validation of our results, in the Mubwindi Swamp (Armbrecht et al., 2019; Giguet-Covex et al., 2019).

## Sources of Metagenomic DNA

The deposition of DNA in a sedimentary basin and its representation of the locally present biota are influenced by the abundance of the locally present species, their biomass, their genome length, transport processes and by chance (but see Andersen et al., 2012; Yoccoz et al., 2012; Giguet-Covex et al., 2019). Four possible sources provide DNA to the sediments of Mubwindi Swamp including (1) DNA from micro- and macro-organisms living in the sediment, (2) DNA from organisms living on or utilizing the surface of the swamp, (3) DNA derived via streams and runoff from the catchment of the swamp (see Giguet-Covex et al., 2019), and (4) DNA derived by deposition from the air (e.g., pollen grains: Parducci et al., 2005). Given these sources, DNA extracted from the sediments will be a mixture of modern and ancient DNA.

Mubwindi Swamp was apparently an open water environment (shallow lake) from about about 1320 to ∼650 cal BP and then served as habitat for (semi)aquatic biota (e.g., Cypriniformes; **Figure 10**) and very likely as a water source to terrestrial fauna (e.g., Cetartiodactyla; **Figure 10**). At this stage the sediments received local authochtonous DNA and allochotonous DNA influx from the catchment that could be transferred to the central part of the basin. In the master core the taxonomic diversity is highest in the clayey sediment below ∼330 cm depth (older than ∼720 cal BP) corresponding to the shallow lake phase. Possibly the Mubwindi basin received more diverse DNA via erosion and runoff from its catchment during this phase than the peatland that formed later. Slightly higher sedimentation rates prior to 800 cal BP together with higher XRF counts of Al and Ti (**Figure 5**), which are indicators of erosion, support this interpretation. At about 650 cal BP the peatland had formed (**Figure 3**) and terrestrial habitat conditions became established that changed DNA transport pathways as a consequence.

Swamps and peatlands provide habitat to a diverse group of organisms that live in the organic deposits. Microbes will contribute modern DNA over the entire sediment column and thus disturb ancient DNA signals. They are expected to be most abundant in the aerobic zone of the peat surface of Mubwindi Swamp (water table: 0–14 cm measured during field work), where also high soil temperatures and in the center of the swamp also higher nutrient concentrations likely favor microbial activity (Dickinson, 1983; Dabney et al., 2013b; **Table 2**). Generally, the aerobic zone of peat deposits in swamps is habitat for various protozoa (e.g., amoeba) and soil fauna (e.g., mites, collembola, nematodes; Mason and Standen, 1983; Speight and Blackith, 1983) and also the rooting zone for most plants (Crawford, 1983). These organisms will add their DNA to the sediments as indicated by the detection of nematode DNA (**Figure 10A**). Deeper anaerobic conditions are likely to be associated with reduced microbial life but provide habitat for Archaea (**Figure 7**). Plants with anatomical transport mechanisms for oxygen will be able to survive in this waterlogged environment (Crawford, 1983). Radiocarbon dating of roots has shown that living herbaceous plants can insert their roots into waterlogged sediments of subtropical swamps to a depth of 60 cm (Glaser et al., 2012). Therefore, living plants may also be contributing DNA directly into the upper sediment column. We expect that the roots of the sedges *Cyperus latifolius* and *Cyperus denudatus*, which dominate most areas of Mubwindi Swamp and likely contribute a large proportion to the peat, represent a major source of sedDNA as indicated by the frequent detection of Poales (**Figure 9**).

Given these conditions most biotic activity should be concentrated in approximately the upper 25 cm of the sediment column. Indeed, from this zone most DNA was extracted from the cores and the DNA shows the lowest proportion of deamination and fragmentation (**Figure 7**). The extracted DNA pool from the upper sediment column thus contains a large contribution of modern DNA from living biota. Hence, the decline in the DNA yield with depth is not simply an indication of DNA degradation, but also of decreasing biotic activity in the sediment column. This high proportion of *in situ* derived DNA will also overwhelm signals of DNA derived from passing animals and influx from the swamp's catchment. Moreover, the biotically active zone will contribute a certain signal of time-averaging by all those species that can exist across a wider zone (>20 cm) of sediment. Time-averaging of sedaDNA may therefore be a specific problem in slowly accumulating systems. Permanent and deep lakes with anoxic bottom waters may be better sources for catchment studies of past biota. Yet modern microbial DNA will always be a significant part of any sedaDNA profile.

## Mubwindi Swamp sedaDNA Preservation

The collected sediment cores differ substantially in terms of their composition and the sediments show a large range in water content (15–94%), organic matter content (6–92%), nutrients (N, P, K), pH (4.4–6.7), and temperature (14.6–20.6°C; **Figure 4**). DNA is preserved in all sediment types and persists under various environmental conditions with few clear explanatory trends (**Figure 11** and **Supporting Information**). DNA is thus degrading at similar rates in different sedimentary environments with varying nutrient, pH and temperature conditions. General trends in our data are (1) the exponential decline in the amount of DNA with sediment depth and age, (2) the increase in DNA fragmentation with depth and age, and (3) the increase in 5′ C→T and 3′ G→A substitutions with depth and age (**Figure 7**). Nevertheless, the preserved DNA's read lengths, deamination rates, and quantities show significant monotonic relationships with almost all environmental factors within the master core, including those unlikely to have a direct causative relationship (e.g., sediment magnetic susceptibility: **Figure 11** and **Supporting Information**). This finding suggests that there is not a single environmental driver or simple set of factors that influence the degradation of DNA in swamp sediments (**Figure 11**; Dabney et al., 2013b) and complicates the generation of a comprehensive statistical model for DNA preservation in the Mubwindi Swamp due to environmental parameter non-independence. The roles of local sedimentation patterns and sediment types in DNA preservation demand further study.

At the master core site, the DNA in the basal sand layer was apparently deposited under drier conditions with neutral pH and lower nutrient levels than those above (**Figure 4**), which likely limited hydrolysis immediately after deposition (Lindahl, 1993; Pääbo et al., 2004). Although such conditions have been associated with comparably "good" sedaDNA preservation in caves (e.g., Hofreiter et al., 2003; Slon et al., 2017), at Bwindi the amount of remaining DNA and its fragmentation in the basal sediments are very similar to that of the overlaying wetter acidic organic peat deposits (**Figure 7**). Although higher taxonomic diversity was detected in the deeper minerogenic sediments than in the peat, we assume that catchment-scale transport processes rather than differential lithologic control on preservation account for these findings. Organic matter content also appears to have limited effects on the quantity of preserved DNA and its fragmentation (**Figure 4**). In the master core, DNA yield and fragmentation reach background levels below about 200 cm, where organic matter is present in a wide range of values (6–88%). Thus, the deposition in clastic sediment (sandy, silty, or clay deposits) or in organic peat deposits seems not to determine how well DNA is preserved in the long-term. This conclusion is supported by similar initial amounts of DNA yield in the surface of all cores and the similar decline in DNA yield in the first meter of the short cores (4A, 5A) (**Figure 6A**), which entirely consist of clay in contrast to the master core's upper peat layer (**Figure 3**).

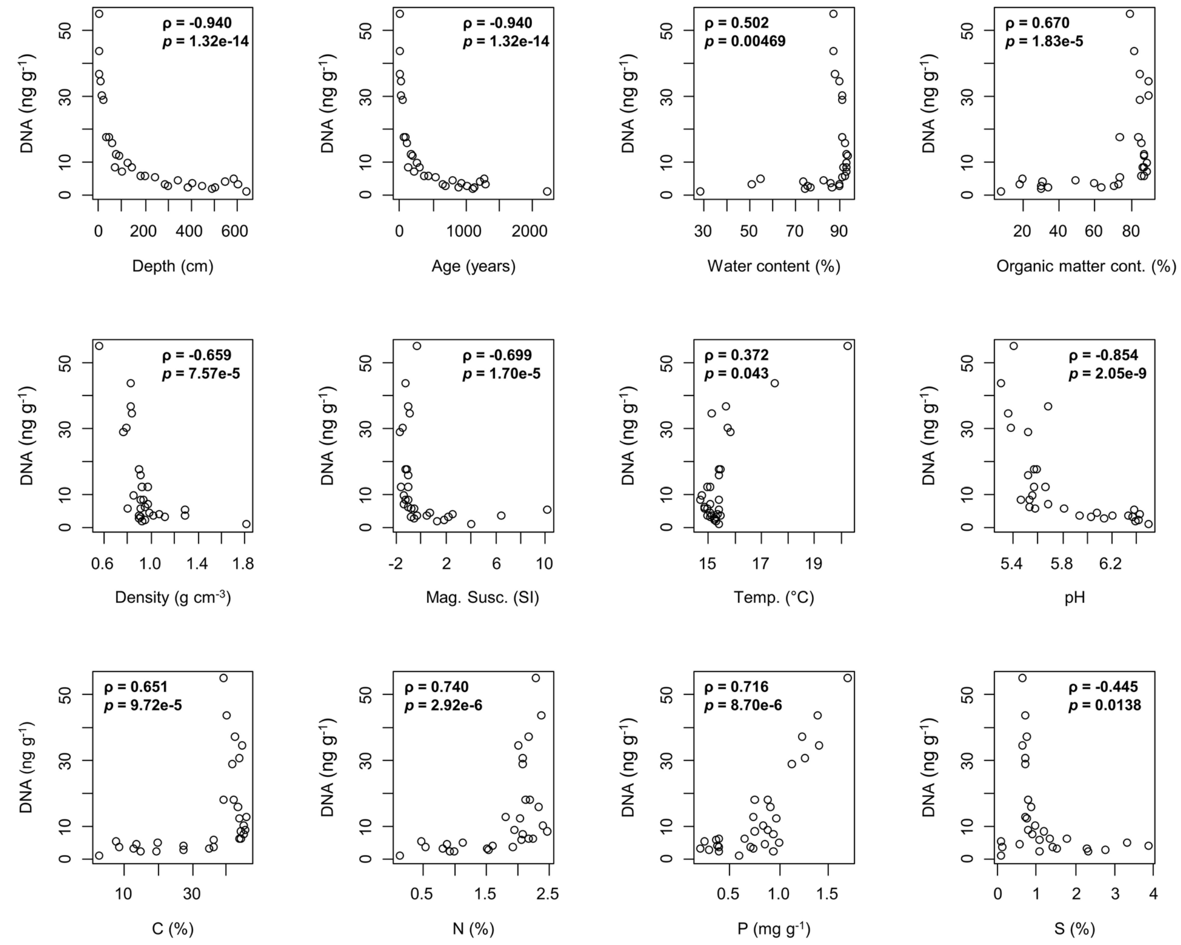


**FIGURE 11 |** Relationships between DNA fraction and environmental parameters for the master core MUB17-2C/D. The strength (Spearman's ρ) and statistical significance (uncorrected *p*-values) of non-parametric correlations are given for each comparison.

In addition, the variation in the elemental composition of the master core is also not reflected in variations of DNA yield or preservation stage. Despite large changes in the quantity of cations such as K, Al, or Fe we find no indication for preferential preservation through DNA adsorption in zone of higher concentrations of metals (**Figure 5**). The DNA patterns of the short cores with their high Fe XRF counts and clay content corroborate this inference.

As predicted by current models of DNA preservation (e.g., Dabney et al., 2013b; Kistler et al., 2017), the vast majority of DNA content loss occurs rapidly after deposition in the Mubwindi Swamp (**Figures 6**, **7**). Not only was there a sharp decline in DNA content below the top meter of bioactive sediment, we observed very few eukaryotic sequences, even in sites of recent elephant activity, suggesting that deposited DNA degrades rapidly in the acidic and warm (17–21°C) swamp environment (Giguet-Covex et al., 2019). Based on laboratory and field experiments, acidic, warm conditions are known to promote DNA hydrolysis and are non-conducive to long-term DNA preservation (Lindahl and Nyberg, 1972; Lindahl, 1993; Strickler et al., 2015; Kistler et al., 2017; Seymour et al., 2018). However, inference of the influence of pH and temperature on DNA degradation by direct comparison of DNA yield with these parameters is misleading. Most DNA is deposited at the sediment surface (where DNA yield, acidity, and temperature are highest), whereas DNA yield declines with depth as acidity and temperature decrease. This finding likely indicates the concentration of DNA from living biota in the sediment surface where an aerobic, warm, and more nutrient-rich environment facilitates abundant plant growth and microbial and soil faunal activity. DNA degradation under surface conditions only becomes apparent when deposition time is considered. In the swamp's center, DNA is deposited at the sediment surface at > 20°C and then exposed to the maximum burial temperatures for over 50 years (= 25 cm depth) until temperature stabilizes at about 15.5°C (**Figure 4**). This time of exposure to high temperatures (and low pH) will likely drive rapid DNA degradation, while at the same time the *in situ* microbial contribution to the DNA pool should also decrease as the sediment changes from aerobic to anaerobic as inferred from the presence of Archaea and the measured minimum position of the water table at -14 cm. Furthermore, the higher concentrations of sedimentary nitrogen and phosphorus near the surface likely contribute to DNA degradation by stimulating microbial decay (Dabney et al., 2013b; **Figure 4**) and are thus not reflecting good preservation conditions as suggested in **Figure 11**.

Although the reduced DNA content in earlier (deeper) strata likely correlates with decreased biotic activity at these depths, statistically, age and depth are the best predictors for the remaining fraction of DNA in the sediment (age/depth: $r^2 = 0.755$, $p < 0.001$). Furthermore, the decreasing fragment lengths and increased cytosine deamination of DNA sequences deriving from deeper, earlier sediments suggest that at least a portion of these molecules are preserved ancient DNA. Moreover, the pattern of increasing cytosine deamination with depth is not consistent with the DNA from deeper strata deriving solely from water-driven leaching of DNA from extant, living and recently deceased organisms through the sediment column despite up to > 90% sediment water content. In the case of the leaching scenario, we would expect all sediment DNA samples to exhibit roughly the same level of deamination since they would all be nearly of the same age. Furthermore, in the leaching scenario, we would expect a continuous gradient in DNA content down the core. Instead, we find that the DNA content stabilizes below the first meter of sediments. We conclude that, despite the acidic condition, genuine ancient DNA has survived in Mubwindi Swamp for over 2200 years and sedaDNA is therefore retrievable from tropical swamps.

Our data suggest a tropical sediment model in which most DNA degrades beyond recovery rapidly after deposition in the swamp (<200 years) (Bremond et al., 2017). The majority of DNA molecules in the most recent sediment layers likely derive from ongoing biotic activity rather than preserved ancient biomolecules. The small portion of DNA molecules that survive this initial stage of elimination become stabilized in the less acidic and cooler deeper sediment levels, which are more conducive to long-term DNA survival. These deeper sediments still preserve a sedaDNA record of diverse taxonomic assemblages. In contrast to other studies of African sedaDNA (Boessenkool et al., 2014; Bremond et al., 2017) we, in fact, observe no decline in the number of taxa with increasing sediment age. Instead the master core contains higher taxonomic diversity in sediment older than ca. 600 cal BP. The absence of a declining trend in taxonomic richness with age suggests that sedaDNA can provide information on past tropical biodiversity for several thousand years.

## Challenges for sedaDNA Studies

### DNA Recovery Biases

Recovered ancient DNA molecules will vary in length and quantity by DNA extraction and library preparation protocols (Rohland and Hofreiter, 2007; Meyer and Kircher, 2010; Dabney and Meyer, 2012; Dabney et al., 2013a; Gansauge and Meyer, 2013; Gansauge et al., 2017; Glocke and Meyer, 2017; Rohland et al., 2018). A wide-variety of aDNA extraction methods, including silica-based, alcoholic, and phenol-chloroform protocols, exist (Rohland and Hofreiter, 2007; Hagan et al., 2020), each optimized to various substrates. Similarly, a multitude of aDNA library preparation protocols have been developed, including double-stranded (e.g., Meyer and Kircher, 2010), single-stranded (Gansauge and Meyer, 2013; Gansauge et al., 2017) and single-tube approaches (Carøe et al., 2018). Even subtle experimental decisions can have significant downstream effects. For instance, within silica-based extractions, choices of extraction buffers (Rohland and Hofreiter, 2007), binding buffers and matrices (Dabney et al., 2013a; Rohland et al., 2018), and purification protocols (Rohland et al., 2018) have significant effects on DNA yields and length biases. The choice of library amplification polymerases has been shown to have significant impacts on molecular length and GC content (Dabney and Meyer, 2012).

Since sediments are heterogenous substrates, sedimentary DNA analyses may be prone to extraction and library preparation biases due to extraction and library protocols being insufficiently optimized across strata (e.g., sand vs. silty clay within the Mubwindi core). Moreover, community structure reconstructions will vary depending on chosen extraction method because these protocols vary in their recovery efficiency between intra- and extracellular DNA (e.g., Vuillemin et al., 2017). We used the PowerSoil extraction kit as it has been shown to be relatively efficient across a wide variety of substrates (including ancient samples: Hagan et al., 2020) and is therefore the recommended protocol of the Earth Microbiome Project (Marotz et al., 2017; Thompson et al., 2017) and the Human Microbiome Project (Aagaard et al., 2013). The microbial results are thus highly comparable to those of these large-scale projects. Nevertheless, the PowerSoil protocol has not been specifically optimized for sedaDNA extraction (e.g., Armbrecht et al., 2020) and is likely to have inefficiently extracted the shortest aDNA molecules (Rohland et al., 2018). This bias is compounded by our double-stranded library preparation protocol, which while more cost- and time-efficient, is less likely to recover the most damaged aDNA molecules than a single-stranded approach (Gansauge and Meyer, 2013; Gansauge et al., 2017; Glocke and Meyer, 2017).

### Mixture of Modern and Ancient DNA

The mixture of modern and ancient DNA in sediments complicates the reconstruction of past ecosystems. Extracting authentic eukaryotic ancient DNA from a pool of DNA dominated by modern and ancient microbial DNA is a major challenge due to their rarity (e.g., Slon et al., 2017). Every sedimentary deposit is also a habitat to microbes and so there is no depositional environment that solely contains ancient DNA. Authenticating past microbial life is even more challenging as ancient microbes will be very similar to extant species currently inhabiting the sediment (e.g., Campana et al., 2014). DNA results from environments impacted by significant leaching will be even harder to authenticate (Haile et al., 2007).

In addition, determining both the drivers of DNA decay and the half-life of sedaDNA is exceptionally challenging with a mixed assemblage of ancient and modern DNA. For example, a low pH is expected to lead to both faster DNA degradation and reduced biotic activity and therefore lower modern DNA contribution, with both effects contributing to lower DNA yield. Unlike with samples of known species (such as bone: Allentoft et al., 2012), any approach to calculate DNA degradation half-life must incorporate the DNA contributions from *both* living or recently deceased organisms in a sediment context, which could render results ambiguous.

## Limitations of Current Metagenomic Databases

Our ecological community reconstructions were limited by very low sequence identification rates (<13% in all sediment samples). Of a total 81.8 million unique, merged reads, only 6.3 million can be assigned to taxonomic entities (7.7%) with the Refseq database (4 million (4.9%) with the nt database). This result is unsurprising given that Ugandan taxa in general are nearly unrepresented in genomic reference databases. Nevertheless, low sequence identification rates from metagenomic sedaDNA studies are not limited to the tropics. Ahmed et al. (2018) and Parducci et al. (2019) were only able to assign ∼16 and 2.3% of reads, respectively, from Hässeldala Port (southern Sweden) sediments. Similarly, Slon et al. (2017) identified only between 4 and 21% of metagenomic reads from Eurasian Pleistocene cave sediments. Low identification rates are problematic for ancient DNA research as they increase experimental costs and decrease reliability in the generated community composition profiles.

Our data suggested systematic biases due to sequence database limitations in the reconstructed profiles. Eukaryotic sequences were very rare (nt: 4.4% of identified sequences, 0.2% of total merged reads; Refseq: 4.0% of identified sequences, 0.3% of total merged reads) in the Mubwindi Swamp sediments. The rarity of plant sequences is surprising given the dominance of plant biomass in the Mubwindi Swamp and the frequency of identified plant-associated microbial taxa. Similarly, Parducci et al. (2019) identified only 1,634 plant reads from ∼1 billion metagenomic sequences (representing < 0.1% of assigned reads) from Hässeldala Port sediments, indicating that poor plant identification is a common issue using current databases. Yet, at the taxonomic rank of order, the few plant DNA sequences evidently provide a reliable reconstruction of past floristic composition from the Mubwindi Swamp and its catchment. A comparison of the sedDNA plant data with published pollen records from Mubwindi Swamp (Marchant et al., 1997; Marchant and Taylor, 1998) showed good agreement between these independent datasets (**Figure 9**). Thirteen of the 14 orders detected by sedDNA were also recorded by pollen analysis (Marchant et al., 1997). The order Solanales not found as pollen, however, contains species that occur in Bwindi today (e.g., Stanford and Nkurunungi, 2003). One of the most frequently taxa detected by sedDNA is Poales, which includes the family Cyperaceae and which is the most common pollen type in the cores of Marchant et al. (1997). This observation is consistent with the dominance of sedges in Mubwindi Swamp and their addition of belowground biomass (roots, rhizomes) to the peat deposit.

Core MUB3 of Marchant et al. (1997) is closest to our master core allowing us to compare presence of taxa with depth (**Figure 9**). In general plant DNA detects about a quarter (∼15–60%) of the orders found as pollen in individual samples of similar depth. The majority of taxa identified by sedDNA matches pollen taxa that are of local origin (e.g., Poales, Rosales), that is pollen produced by plants that (potentially) grow in Mubwindi Swamp (**Figure 9**; Marchant et al., 1997). An exception is Fabales which were regularly detected by DNA, yet *Newtonia buchananii* a member of Fabales is a common tree in the swamp's direct catchment and a possible past DNA source (Marchant et al., 1997). These observations generally suggest that most of the sedDNA was derived from local sources.

The relatively good taxonomic correspondence and similar detection with depth between DNA and pollen indicates that shotgun-sequenced sedaDNA is reliably recording plant history in the tropics at the ordinal level. Given the large proportion of unknown plant DNA in our data we expect increases in both the detection of taxa (diversity) and the resolution of taxonomic level with future improvements of species coverage in genomic databases. For sedaDNA to outperform the taxonomic resolution of pollen analysis first necessitates genome-sequencing of the majority of plant species in Bwindi Impenetrable Forest and in tropical forests in general. However, we advocate to apply both methods in combination when reconstructing past floristic diversity.

This study, the first examining Metazoan assemblages with sedaDNA in tropical Africa, successfully revealed the past occurrence of 16 native orders of animals at Mubwindi Swamp. Species belonging to all these orders occur in Bwindi today and members of most of them, including Rodentia, Primates, Carnivora, Cetartiodactyla, Eulipotyphla, Chiroptera, Passeriformes, Squamata, Lepidoptera, Diptera, Hymenoptera, Hemiptera, and Coleoptera visit or inhabit Mubwindi Swamp at present (e.g., Butynski, 1984; Kasangaki et al., 2003, 2008; Decru et al., 2019; pers. observ.), supporting the correctness of these taxonomic detections in the sediment record. The co-occurrence of most animal taxa in our sedaDNA data-set, in particular the frequent occurrence of Afrotheria, Euarchontoglires (including Rodentia and Primates), Carnivora, and Cetartiodactyla (**Figure 10**) suggest that Bwindi Impenetrable Forest was inhabited by typical African faunal assemblages during the past 2200 years. Intriguing is the missing or low detection of various wetland taxa that are common in Mubwindi Swamp today such as Odonata, Trichoptera, Anura, Gastropoda or Eulipotyphla and which should therefore be important DNA sources. This result could suggest that these important groups are particularly underrepresented in genomic surveys. The detection of vertebrates is moreover biased by species which have been genome sequenced – mostly large, charismatic taxa and model species (Brandies et al., 2019). Smaller and elusive animal species that constitute most of the diversity, such as shrews and frogs will be less likely detected. A comprehensive reconstruction of animal diversity of Bwindi Impenetrable Forest would therefore be premature at this stage and systematic genomic surveys of Bwindi taxa are required in order to produce more reliable, precise ancient DNA taxonomic profiles that can serve as baseline data for conservation.

Assigning taxa below ordinal rank is still prone to large uncertainties given the inadequate coverage of Afrotropical taxa in reference databases. This challenge is emphasized by the assignment of reads to several exotic animal taxa that either have no Quaternary record in tropical Africa (Metatheria, Protacanthopterygii) or inhabit the ocean (Cetacea, Octopoda, partly Protacanthopterygii) (**Figure 10B**). The issue of dealing with higher taxonomic levels and more importantly the 92% of unidentifiable data restrict clear insights into community dynamics over time. Moreover, the choice of reference database has also significant effects on the resulting taxonomic composition of metagenomic data. Whereas the nt database derived record shows a scattered presence without clear temporal trend for most animal orders at Mubwindi, the Refseq derived data indicate a nearly continuous presence for over half of the detected orders, particularly for mammals (**Figure 10**). These contrasting results emphasize that multiple genomic reference databases should be explored in tandem when working with shotgun-sequenced sedaDNA data-sets.

We also found systematic biases due to differences between metagenomic analysis pipelines. QIIME 2 analyses found that Archaea comprised a much larger portion of the microbial community than in the MEGAN dataset (**Figure 7**). This result likely reflects differences in taxonomic database biases. Far more complete bacterial genomes have been sequenced than archaeal ones (25,496 bacterial genomes vs. 1,680 archaeal in the NCBI Genome database as of 28 October 2019), resulting in a bias toward the preferential identification of bacterial sequences. Although the same bias exists in the 16S dataset (592,561 bacterial and 25,026 archaeal sequences in the SILVA build 132 non-redundant 16S database), 16S has been better characterized for both domains, reducing the bias's impact. The MEGAN profile therefore probably undercounts the presence of Archaea relative to Bacteria, with the QIIME 2 profile being more representative of the true community. Similarly, MEGAN documented dominance of the microbial communities by Rhizobiales and Burkholderiales, which was not observed in the QIIME 2 dataset. In this case, it is difficult to determine which pipeline produced the more accurate community reconstructions. The choice of database for microbial detection is also critical: our MEGAN analyses found slight, but significant, taxonomic compositional differences between the nt- and Refseq-aligned datasets. As found previously (Ye et al., 2019), MetaPhlAn2 performed poorly as a metagenomic classifier. We recommend against its further use in sedaDNA research. The megaBLAST/MEGAN pipeline is common in sedaDNA analyses (e.g., Parducci et al., 2019), even being called the "gold standard" by Cribdon et al. (2020). QIIME (Caporaso et al., 2010) and QIIME 2 are common in sedaDNA analyses using metabarcoding (e.g., Ziesemer et al., 2015). We therefore employed these classifiers for comparability with extant datasets. Nevertheless, metagenomic standards and classification algorithms are advancing rapidly (Ye et al., 2019; Cribdon et al., 2020), which will necessitate development of more accurate ancient DNA classifiers such as PIA (Cribdon et al., 2020) and MALT (Vågene et al., 2018).

## CONCLUSION

We present one of the richest biogeological and DNA datasets for tropical sediments yet published. Using this dataset, we model ancient DNA preservation in the Mubwindi Swamp, showing that age and depth are the strongest determinants of DNA preservation and fragmentation, but that almost all environmental parameters have monotonic relationships with DNA degradation. Surprisingly, microenvironmental conditions and lithology show few systematic impacts on DNA preservation in this sedimentary basin. We show that metagenomic sedimentary DNA can provide valuable insights into past tropical biodiversity, but that further development of genomic databases is necessary to provide robust, detailed community reconstructions. The actual taxonomic composition and resolution of our recovered DNA would likely change if all sequences could be identified. This problem of a skewed reconstruction of taxonomic composition will be present in any future sedaDNA study for the tropics until the genomes of most tropical species have been sequenced.

Besides a fundamental improvement of genomic databases, the transport, deposition, and ultimate representation of sedaDNA of different ecosystems need to be systematically investigated. Calibration studies should inventory a site's extant diversity and compare its biota with the taxonomic composition of environmental DNA samples from various locations and transport pathways (streams etc.) within the site's catchment and basin. Until taxonomically representative databases are generated and further DNA taphonomic studies are completed, sedaDNA cannot be fully utilized for biodiversity studies in the tropics.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the Sequence Read Archive accession PRJNA573108.

## AUTHOR CONTRIBUTIONS

RD, MM, RP, JM, and MC secured the funding. RD, MM, JM, and MC designed the study. RD, MA, MM, JN, and MC performed the field sampling. RD, MA, and JN secured the necessary permits for the study. RD, MA, NP, TG, and JN performed the experiments. RD and MC analyzed and archived the data. RD and MC wrote the manuscript with contributions from all co-authors. All authors contributed to the article and approved the submitted version.

## FUNDING

This material was based upon work supported by the Global Genome Initiative under Grant No. (GGI-Exploratory Science-2016) awarded to RD. This work was financially also supported by Geo.X, the Research Network for Geosciences in Berlin and Potsdam. RD acknowledges a KoUP seed fund from the University of Potsdam for establishing collaboration with Mbarara University of Science and Technology. The Smithsonian Institution (Human Origins Program, National Museum of Natural History and the Smithsonian National Zoological Park and Conservation Biology Institute) supported this research. Elevation data in **Figure 1** are based on the TanDEM-X Science DEM granted to RD by the German Space Agency (DLR 2017). We acknowledge the support of the open-access publication fund of the University of Potsdam.

## ACKNOWLEDGMENTS

We acknowledge the LacCore facility for their support in equipment renting, core storage and processing. This research and core export were approved by the Uganda National Council of Science and Technology (Ref. No. NS13ES) and by the Uganda Wildlife Authority (Ref. No. UWA/COD96/05). We are grateful for the support from Lornah Nabukwasi, Jessica Quinta, Lydia Olaka, Casim Umba Tolo, Julius B. Lejju, Matt Tocheri, Bodo Bookhagen, Manfred Strecker, Paul Glaser, and Mark Shapley and from Jörg Gelbrecht, Grit Siegert, Constancia Levertz, Elisabeth Schuette, Sarah Krocker, and Thomas Rossoll of IGB Berlin. At Bwindi we received generous help from the Uganda Wildlife Authority, especially from wardens George Businge, Raymond Kato, Joseph Arinaitwe, and Pontius Ezuma as well as from all staff of the Institute of Tropical Forest Conservation, Chrispino Mugarura, the Uganda Police Force and from many porters at Ruhija. This research was conducted using the Smithsonian Institution High Performance Cluster (SI/HPC): https://doi.org/10.25572/SIHPC.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2020.00218/full#supplementary-material

## REFERENCES

Aagaard, K., Petrosino, J., Keitel, W., Watson, M., Katancik, J., Garcia, N., et al. (2013). The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. *FASEB J.* 27, 1012–1022. doi: 10.1096/fj.12-220806

Adams, C. I. M., Knapp, M., Gemmell, N. J., Jeunen, G.-J., Bunce, M., Lamare, M. D., et al. (2019). Beyond biodiversity: Can environmental DNA (eDNA) cut it as a population genetics tool? *Genes* 10:192. doi: 10.3390/genes10030192

Adler, C. J., Haak, W., Donlon, D., Cooper, A., and Genographic Consortium. (2011). Survival and recovery of DNA from ancient teeth and bones. *J. Archaeol. Sci.* 38, 956–964. doi: 10.1016/j.jas.2010.11.010

Ahmed, E., Parducci, L., Unneberg, P., Ågren, R., Schenk, F., Rattray, J. E., et al. (2018). Archaeal community changes in Lateglacial lake sediments: Evidence from ancient DNA. *Quat. Sci. Rev.* 181, 19–29. doi: 10.1016/j.quascirev.2017.11.037

Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L., Hale, M. L., et al. (2012). The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. B Biol. Sci.* 279, 4724–4733. doi: 10.1098/rspb.2012.1745

Andersen, J. (1976). An ignition method for determination of total phosphorus in lake sediments. *Water Res.* 10, 329–331. doi: 10.1016/0043-1354(76)90175-5

Andersen, K., Bird, K. L., Rasmussen, M., Haile, J., Bruening-Madsen, H., Kjær, K. H., et al. (2012). Meta-barcoding of 'dirt'. DNA from soil reflects vertebrate biodiversity. *Mol. Ecol.* 21, 1966–1979. doi: 10.1111/j.1365-294X.2011.05261.x

Andrews, S. (2016). *FastQC: A Quality Control Tool for high Throughput Sequence Data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed July 15, 2016).

Armbrecht, L., Herrando-Pérez, S., Eisenhofer, R., Hallegraeff, G. M., Bolch, C. J. S., and Cooper, A. (2020). An optimized method for the extraction of ancient eukaryote DNA from marine sediments. *Mol. Ecol. Resour.* 1–14. doi: 10.1111/1755-0998.13162

Armbrecht, L. H., Coolen, M. J. L., Lejzerowicz, F., George, S. C., Negandhi, K., Suzuki, Y., et al. (2019). Ancient DNA from marine sediments: precautions and considerations for seafloor coring, sample handling and data generation. *Earth Sci. Rev.* 196:102887. doi: 10.1016/j.earscirev.2019.102887

Babaasa, D. (2000). Habitat selection by elephants in Bwindi Impenetrable National Park, south-western Uganda. *Afr. J. Ecol.* 38, 116–122. doi: 10.1046/j.1365-2028.2000.00226.x

Barlow, J., França, F., Gardner, T. A., Hicks, C. C., Lennox, G. D., Berenguer, E., et al. (2018). The future of hyperdiverse tropical ecosystems. *Nature* 559, 517–526. doi: 10.1038/s41586-018-0301-1

Blaauw, M., and Christen, J. A. (2011). Flexible paleoclimate age-depth models using an autoregressive gamma process. *Bayesian Anal.* 6, 457–474. doi: 10.1214/11-BA618

Blackburn, D. C., Boix, C., Greenbaum, E., Fabrezi, M., Meirte, D., Pumptre, A. J., et al. (2016). The distribution of the Bururi Long-fingered Frog (Cardioglossa cyaneospila, family Arthroleptidae), a poorly known Albertine Rift endemic. *Zootaxa* 4170, 355–364. doi: 10.11646/zootaxa.4170.2.8

Boessenkool, S., Epp, L. S., Haile, J., Bellemain, E., Edwards, M., Coissac, E., et al. (2012). Blocking human contaminant DNA during PCR allows amplification of rare mammal species from sedimentary ancient DNA. *Mol. Ecol.* 21, 1806–1815. doi: 10.1111/j.1365-294X.2011.05306.x

Boessenkool, S., McGlynn, G., Epp, L. S., Taylor, D., Pimentel, M., Gizaw, A., et al. (2014). Use of ancient sedimentary DNA as a novel conservation tool for high-altitude tropical biodiversity. *Conserv. Biol.* 28, 446–455.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi: 10.1038/s41587-019-0209-9

Brandies, P., Peel, E., Hogg, C. J., and Belov, K. (2019). The value of reference genomes in the conservation of threatened species. *Genes* 10:846. doi: 10.3390/genes10110846

Bremond, L., Favier, C., Ficetola, G. F., Tossou, M. G., Akouégninou, A., Gielly, L., et al. (2017). Five thousand years of tropical lake sediment DNA records from Benin. *Quat. Sci. Rev.* 170, 203–211. doi: 10.1016/j.quascirev.2017.06.025

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., et al. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U.S.A.* 104, 14616–14621. doi: 10.1073/pnas.0704665104

Brooks, T., Hoffmann, M., Burgess, N., Plumptre, A., Williams, S., Gereau, R. E., et al. (2004). "Eastern afromontane," in *Hotspots Revisited: Earth's Biologically Richest and Most Endangered Ecoregions*, 2nd Edn, eds R. A. Mittermeier, P. Robles-Gil, M. Hoffmann, J. D. Pilgrim, T. M. Brooks, C. G. Mittermeier, et al. (Mexico: Cemex), 241–242.

Butynski, T. M. (1984). *Ecological Survey of the Impenetrable (Bwindi) Forest, Uganda, and Recommendations for its Conservation and Management*. New York, NY: New York Zoological Society.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopolous, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Campana, M. G., Bower, M. A., and Crabtree, P. J. (2013). Ancient DNA for the archaeologist: the future of African research. *Afr. Archaeol. Rev.* 30, 21–37. doi: 10.1007/s10437-013-9127-2

Campana, M. G., Robles García, N., Rühli, F. J., and Tuross, N. (2014). False positives complicate ancient pathogen identification using high-throughput shotgun sequencing. *BMC Res. Notes* 7:111. doi: 10.1186/1756-0500-7-111

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S. T., Sinding, M. H. S., Samaniego, J. A., et al. (2018). Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* 9, 410–419. doi: 10.1111/2041-210X.12871

Ceballos, G., Ehrlich, P. R., and Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proc. Natl. Acad. Sci. U.S.A.* 114, E6089–E6096. doi: 10.1073/pnas.1704949114

Crawford, R. M. M. (1983). "Root survival in flooded soils," in *Ecosystems of the World. 4A Mires: Swamp, Bog, Fen and Moor General Studies*, ed. A. J. P. Gore (Amsterdam: Elsevier), 257–283.

Cribdon, B., Ware, R., Smith, O., Gaffney, V., and Allaby, R. G. (2020). PIA: More accurate taxonomic assignment of metagenomic data demonstrated on sedaDNA from the North Sea. *Front. Ecol. Evol.* 8:84. doi: 10.3389/fevo.2020.00084

Dabney, J., Knapp, M., Glocke, I., Gansauge, M. T., Weihmann, A., Nickel, B., et al. (2013a). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15758–15763. doi: 10.1073/pnas.1314445110

Dabney, J., and Meyer, M. (2012). Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques* 52, 87–94. doi: 10.2144/000113809

Dabney, J., Meyer, M., and Pääbo, S. (2013b). Ancient DNA damage. *Cold Spring Harb. Perspect. Biol.* 5:a012567. doi: 10.1101/cshperspect.a012567

Dean, W. E. (1974). Determination of carbonate and organic matter in calcareous sediments and sedimentary rocks by loss on ignition; comparison with other methods. *J. Sediment. Res.* 44, 242–248. doi: 10.1306/74D729D2-2B21-11D7-8648000102C1865D

Decru, E., Vranken, N., Bragança, P. H., Snoeks, J., and Van Steenberge, M. (2019). Where ichthyofaunal provinces meet: the fish fauna of the Lake Edward system, East Africa. *J. Fish Biol.* 96, 1186–1201. doi: 10.1111/jfb.13992

Dickinson, C. H. (1983). "Microorganisms in Peatlands," in *Ecosystems of the World. 4A Mires: Swamp, Bog, Fen and Moor General Studies*, ed. A. J. P. Gore (Amsterdam: Elsevier), 225–243.

Díez-del-Molino, D., Sánchez-Barreiro, F., Barnes, I., Gilbert, M. T. P., and Dalén, L. (2018). Quantifying temporal genomic erosion in endangered species. *Trends Ecol. Evol.* 33, 176–185. doi: 10.1016/j.tree.2017.12.002

Domaizon, I., Winegardner, A., Capo, E., Gauthier, J., and Gregor-Eaves, I. (2017). DNA-based methods in paleolimnology: new opportunities for investigating long-term dynamics of lacustrine biodiversity. *J. Paleolimnol.* 58, 1–21. doi: 10.1007/s10933-017-9958-y

Drewes, R. C., and Vindum, J. V. (1994). Amphibians of the Impenetrable Forest, southwest Uganda. *J. Afr. Zool.* 108, 55–70.

Ebina, J., Tsutsui, T., and Shirai, T. (1983). Simultaneous determination of total nitrogen and total phosphorus in water using peroxodisulfate oxidation. *Water Res.* 17, 1721–1726. doi: 10.1016/0043-1354(83)90192-6

Epp, L. S. (2019). A global perspective for biodiversity history with ancient environmental DNA. *Mol. Ecol.* 28, 2456–2458. doi: 10.1111/mec.15118

Epp, L. S., Stoof, K. R., Trauth, M. H., and Tiedemann, R. (2010). Historical genetics on a sediment core from a Kenyan lake: intraspecific genotype turnover in a tropical rotifer is related to past environmental changes. *J. Paleolimnol.* 43, 939–954. doi: 10.1007/s10933-009-9379-7

Epp, L. S., Stoof-Leichsenring, K. R., Trauth, M. H., and Tiedemann, R. (2011). Molecular profiling of diatom assemblages in tropical lake sediments using taxon-specific PCR and Denaturing High-Performance Liquid Chromatography (PCR-DHPLC). *Mol. Ecol. Resour.* 11, 842–853. doi: 10.1111/j.1755-0998.2011.03022.x

Erlacher, A., Cernava, T., Cardinael, M., Soh, J., Sensen, C. W., Grube, M., et al. (2015). Rhizobiales as functional and enosymbiontic members in the lichen symbiosis of *Lobaria pulmonaria* L. *Front. Microbiol.* 6:53. doi: 10.3389/fmicb.2015.00053

Estrada, O., Breen, J., Richards, S. M., and Cooper, A. (2018). Ancient plant DNA in the genomic era. *Nat. Plants* 4, 394–396. doi: 10.1038/s41477-018-0187-9

Ferrari, G., Lischer, H. E. L., Neukamm, J., Rayo, E., Borel, N., Pospischil, A., et al. (2018). Assessing metagenomic signals recovered from Lyuba, a 42,000-year-old permafrost-preserved woolly mammoth calf. *Genes* 9:436. doi: 10.3390/genes9090436

Ganas, J., Nkurunungi, J. B., and Robbins, M. M. (2009). A preliminary study of the temporal and spatial biomass patterns of herbaceous vegetation consumed by mountain gorillas in an Afromontane rain forest. *Biotropica* 41, 37–46. doi: 10.1111/j.1744-7429.2008.00455.x

Gansauge, M.-T., Gerber, T., Glocke, I., Korlević, P., Lippik, L., Nagel, S., et al. (2017). Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res.* 45:e79. doi: 10.1093/nar/gkx033

Gansauge, M.-T., and Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc.* 8, 737–748. doi: 10.1038/nprot.2013.038

Gaston, K. J. (2000). Global patterns in biodiversity. *Nature* 405, 220–227. doi: 10.1038/35012228

Giguet-Covex, C., Ficetola, G. F., Walsh, K., Poulenard, J., Bajard, M., Fouinat, L., et al. (2019). New insights on lake sediment DNA from the catchment: importance of taphonomic and analytical issues on the record quality. *Sci. Rep.* 9:14676. doi: 10.1038/s41598-019-50339-1

Giguet-Covex, C., Pansu, J., Arnaud, F., Rey, P.-J., Griggo, C., Gielly, L., et al. (2014). Long livestock farming history and human landscape shaping revealed by lake sediment DNA. *Nat. Commun.* 5:3211. doi: 10.1038/ncomms4211

Glaser, P. H., Volin, J. C., Givnish, T. J., Hansen, B. C., and Stricker, C. A. (2012). Carbon and sediment accumulation in the Everglades (USA) during the past 4000 years: rates, drivers, and sources of error. *J. Geophys. Res. Biogeosc.* 117:G03026. doi: 10.1029/2011JG001821

Glenn, T. C., Nilsen, R. A., Kieran, T. J., Sanders, J. G., Bayona-Vásquez, N. J., Finger, J. W. Jr., et al. (2019). Adapterama I: universal stubs and primers for 384 unique dual-indexed or 147,456 combinatorially-indexed Illumina libraries (iTru & iNext). *bioRxiv* [Preprint]. doi: 10.1101/049114

Glocke, I., and Meyer, M. (2017). Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. *Genome Res.* 27, 1230–1237. doi: 10.1101/gr.219675.116

Gomez Cabrera, M. D. C., Young, J. M., Roff, G., Staples, T., Ortiz, J. C., Pandolfi, J. M., et al. (2019). Broadening the taxonomic scope of coral reef palaeoecological studies using ancient DNA. *Mol. Ecol.* 28, 2636–2652. doi: 10.1111/mec.15038

Graham, R. W., Belmecheri, S., Choy, K., Culleton, B. J., Davies, L. J., Froese, D., et al. (2016). Timing and causes of mid-Holocene mammoth extinction on St. Paul Island, Alaska. *Proc. Natl. Acad. Sci. U.S.A.* 113, 9310–9314. doi: 10.1073/pnas.1604903113

Green, R. E., Briggs, A. W., Krause, J., Prüfer, K., Burbano, H. A., Siebauer, M., et al. (2009). The Neandertal genome and ancient DNA authenticity. *EMBO J.* 28, 2494–2502. doi: 10.1038/emboj.2009.222

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., et al. (2010). A draft sequence of the Neandertal genome. *Science* 328, 710–722. doi: 10.1126/science.1188021

Hagan, R. W., Hofman, C. A., Hübner, A., Reinhard, K., Schnorr, S., Lewis, C. M. Jr., et al. (2020). Comparison of extraction methods for recovering ancient microbial DNA from paleofeces. *Am. J. Phys. Anthropol.* 171, 275–284. doi: 10.1002/ajpa.23978

Haile, J., Froese, D. G., MacPhee, R. D. E., Roberts, R. G., Arnold, L. J., Reyes, A. V., et al. (2009). Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proc. Natl. Acad. Sci. U.S.A.* 106, 22352–22357. doi: 10.1073/pnas.0912510106

Haile, J., Holdaway, R., Oliver, K., Bunce, M., Gilbert, M. T. P., Nielsen, R., et al. (2007). Ancient DNA chronology with sediment deposits: Are paleobiological reconstructions possible and is DNA leaching a factor? *Mol. Biol. Evol.* 24, 982–989. doi: 10.1093/molbev/msm016

Hofman, C. A., Rick, T. C., Fleischer, R. C., and Maldonado, J. E. (2015). Conservation archaeogenomics: ancient DNA and biodiversity in the Anthropocene. *Trends Ecol. Evol.* 30, 540–549. doi: 10.1016/j.tree.2015.06.008

Hofreiter, M., Mead, J. I., Martin, P., and Poinar, H. N. (2003). Molecular caving. *Curr. Biol.* 13, R693–R695. doi: 10.1016/j.cub.2003.08.039

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., et al. (2016). MEGAN Community edition—interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput. Biol.* 12:e1004957. doi: 10.1371/journal.pcbi.1004957

Jacobs, B. F., Pan, A. D., and Scotese, C. R. (2010). "A review of the Cenozoic vegetation history of Africa," in *Cenozoic Mammals of Africa*, eds L. Werdelin, and W. J. Sanders (Berkeley, CA: University of California Press), 57–72.

Jenkins, C. N., Pimm, S. L., and Joppa, L. N. (2013). Global patterns of terrestrial vertebrate diversity and conservation. *Proc. Natl. Acad. Sci. U.S.A* 110, E2602–E2610. doi: 10.1073/pnas.1302251110

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684. doi: 10.1093/bioinformatics/btt193

Jousse, H. (2017). *Atlas of Mammal Distribution Through Africa from the LGM (˜18Ka) to Modern times. The Zooarchaeological Record*. Oxford: Archaeopress Publishing Ltd.

Kahindo, C. M., Bates, J. M., and Bowie, R. C. K. (2017). Population genetic structure of Grauer's Swamp Warbler Bradypterus graueri, an Albertine Rift endemic. *Ibis* 159, 415–429. doi: 10.1111/ibi.12453

Kaneko, T., Maita, H., Hirakawa, H., Uchiike, N., Minamisawa, K., Watanabe, A., et al. (2011). Complete genome sequence of the soybean symbiont *Bradyrhizobium japonicum* strain USDA6T. *Genes* 2, 763–787. doi: 10.3390/genes2040763

Kasangaki, A., Bitariho, R., Shaw, P., Robbins, M., and McNeilage, A. (2012). "7. Long-term ecological and socio-economic changes in and around Bwindi Impenetrable National Park, south-western Uganda," in *The Ecological Impact of Long-term Changes in Africa's Rift Valley*, ed. A. J. Plumptre (Hauppauge, NY: Nova Science Publishers), 106–124.

Kasangaki, A., Chapman, L. J., and Balirwa, J. (2008). Land use and the ecology of benthic macroinvertebrate assemblages of high-altitude rainforest streams in Uganda. *Freshw. Biol.* 53, 681–697. doi: 10.1111/j.1365-2427.2007.01925.x

Kasangaki, A., Kityo, R., and Kerbis, J. (2003). Diversity of rodents and shrews along an elevational gradient in Bwindi Impenetrable National Park, south-western Uganda. *Afr. J. Ecol.* 41, 115–123. doi: 10.1046/j.1365-2028.2003.00383.x

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kingdon, J. (1973). Endemic mammals and birds of western Uganda: measuring Uganda's biological wealth and a plea for supra-economic values. *Uganda J.* 37, 1–8.

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acid Res.* 40:e3. doi: 10.1093/nar/gkr771

Kisand, V., Talas, L., Kisand, A., Stivrins, N., Reitalu, T., Alliksaar, T., et al. (2018). From microbial eukaryotes to metazoan vertebrates: wide spectrum paleo-diversity in sedimentary ancient DNA over the last ˜14,500 years. *Geobiology* 16, 628–639. doi: 10.1111/gbi.12307

Kistler, L., Ware, R., Smith, O., Collins, M., and Allaby, R. G. (2017). A new model for ancient DNA decay based on paleogenomic meta-analysis. *Nucleic Acids Res.* 45, 6310–6320. doi: 10.1093/nar/gkx361

Li, H. (2013). *Aligning Sequence Reads, Clone Sequences and Assembly Contigs Using BWA-MEM*. Cambridge, MA: MIT.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature* 362, 709–715. doi: 10.1038/362709a0

Lindahl, T., and Nyberg, B. (1972). Rate of depurination of native deoxyribonucleic acid. *Biochemistry* 11, 3610–3618. doi: 10.1021/bi00769a018

Lorenz, M. G., and Wackernagel, W. (1987). Adsoprtion of DNA to sand and variable degradation rates of adsorbed DNA. *Appl. Environ. Microbiol.* 53, 2948–2952. doi: 10.1128/aem.53.12.2948-2952.1987

Lorenz, M. G., and Wackernagel, W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol. Rev.* 58, 563–602. doi: 10.1128/mmbr.58.3.563-602.1994

Magoč, T., and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957–2963. doi: 10.1093/bioinformatics/btr507

Marchant, R., and Taylor, D. (1998). Dynamics of montane forest in central Africa during the late Holocene: a pollen-based record from western Uganda. *Holocene* 8, 375–381. doi: 10.1191/095968398672993971

Marchant, R., Taylor, D., and Hamilton, A. (1997). Late Pleistocene and Holocene history at Mubwindi Swamp, southwest Uganda. *Quat. Res.* 47, 316–328. doi: 10.1006/qres.1997.1887

Marotz, C., Amir, A., Humphrey, G., Gaffney, J., Gogul, G., and Knight, R. (2017). DNA extraction for streamlined metagenomics of diverse environmental samples. *BioTechniques* 62, 290–293. doi: 10.2144/000114559

Mason, C. F., and Standen, V. (1983). "Aspects of secondary production," in *Ecosystems of the World. 4A Mires: Swamp, Bog, Fen and Moor General Studies*, ed. A. J. P. Gore (Amsterdam: Elsevier), 367–382.

Mergeay, J., Vanoverbeke, J., Verschuren, D., and de Meester, L. (2007). Extinction, recolonization, and dispersal through time in a planktonic crustacean. *Ecology* 88, 3032–3043. doi: 10.1890/06-1538.1

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 2010:pdb.prot5448. doi: 10.1101/pdb.prot5448

Mittelbach, G. G., Schemske, D. W., Cornell, H. V., Allen, A. P., Brown, J. M., Bush, M. B., et al. (2007). Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. *Ecol. Lett.* 10, 315–331. doi: 10.1111/j.1461-0248.2007.01020.x

Mondol, S., Moltke, I., Hart, J., Keigwin, M., Brown, L., Stephens, M., et al. (2015). New evidence for hybrid zones of forest and savanna elephants in Central and West Africa. *Mol. Ecol.* 24, 6134–6147. doi: 10.1111/mec.13472

Morley, R. J. (2000). *Origin and Evolution of Tropical Rain Forests*. Chichester: John Wiley & Sons.

Mugerwa, B., Sheil, D., Ssekiranda, P., van Heist, M., and Ezuma, P. (2013). A camera trap assessment of terrestrial vertebrates in Bwindi Impenetrable National Park, Uganda. *Afr. J. Ecol.* 51, 21–31. doi: 10.1111/aje.12004

Murphy, J., and Riley, J. P. (1962). A modified single solution method for the determination of phosphate in natural waters. *Anal. Chim. Acta* 27, 31–36. doi: 10.1016/S0003-2670(00)88444-5

Nkurunungi, J. B., Ganas, J., Robbins, M. M., and Stanford, C. B. (2004). A comparison of two mountain gorilla habitats in Bwindi Impenetrable National Park, Uganda. *Afr. J. Ecol.* 42, 289–297. doi: 10.1111/j.1365-2028.2004.00523.x

Ogram, A., Sayler, G. S., Gustin, D., and Lewis, R. J. (1988). DNA adsorption to soils and sediments. *Environ. Sci. Technol.* 22, 982–984. doi: 10.1021/es00173a020

Olupot, P., and Plumptre, A. (2010). *Conservation research in Uganda's Forests*. Hauppauge, NY: Nova Science Publishers, Inc.

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rohland, N., et al. (2004). Genetic analyses from ancient DNA. *Annu. Rev. Genet.* 38, 645–679. doi: 10.1146/annurev.genet.37.110801.143214

Parducci, L., Alsos, I. G., Unneberg, P., Pedersen, M. W., Han, L., Lammers, Y., et al. (2019). Shotgun environmental DNA, pollen, and macrofossil analysis of Lateglacial lake sediments from southern Sweden. *Front. Ecol. Evol.* 7:189. doi: 10.3389/fevo.2019.00189

Parducci, L., Bennett, K. D., Ficetola, G. F., Alsos, I. G., Suyama, Y., Wood, J. R., et al. (2017). Ancient plant DNA in lake sediments. *New Phytol.* 214, 924–942. doi: 10.1111/nph.14470

Parducci, L., Nota, K., and Wood, J. (2018). "Reconstructing past vegetation communities using ancient DNA from lake sediments," in *Paleogenomics. Population Genomics*, eds C. Lindqvist, and O. Rajora (Cham: Springer), doi: 10.1007/13836_2018_38

Parducci, L., Suyama, Y., Lascoux, M., and Bennett, K. D. (2005). Ancient DNA from pollen: a genetic record of population history in Scots pine. *Mol. Ecol.* 14, 2873–2882. doi: 10.1111/j.1365-294X.2005.02644.x

Pedersen, M. W., Overballe-Petersen, S., Ermini, L., der Sarkissian, C., Haile, J., Hellstrom, M., et al. (2015). Ancient and modern environmental DNA. *Philos. Trans. R. Soc. B.* 370:20130383.

Pedersen, M. W., Ruter, A., Schweger, C., Friebe, H., Staff, R. A., Kjeldsen, K. K., et al. (2016). Postglacial viability and colonization in North America's ice-free corridor. *Nature* 537, 45–49. doi: 10.1038/nature19085

Pennington, R. T., Hughes, M., and Moonlight, P. W. (2015). The origins of tropical rainforest hyperdiversity. *Trends Plant Sci.* 20, 693–695. doi: 10.1016/j.tplants.2015.10.005

Peterson, B. G., Carl, P., Boudt, K., Bennett, R., Ulrich, J., Zivot, E., et al. (2019). *PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis. R package version* 1.5.3. Available at: https://CRAN.R-project.org/package=PerformanceAnalytics (accessed October 18, 2019).

Plumptre, A. J., Davenport, T. R. B., Behangana, M., Kityo, R., Eilu, G., Ssegawa, P., et al. (2007). The biodiversity of the Albertine Rift. *Biol. Conserv.* 134, 178–194. doi: 10.1016/j.biocon.2006.08.021

Poinar, G. O. (1992). *Life in Amber*. Palo Alto, CA: Stanford University Press.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., et al. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 35, 7188–7196. doi: 10.1093/nar/gkm864

Prüfer, K., Stenzel, U., Hofreiter, M., Pääbo, S., Kelso, J., and Green, R. E. (2010). Computational challenges in the analysis of ancient DNA. *Genome Biol.* 11:R47. doi: 10.1186/gb-2010-11-5-r47

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rawlence, N. J., Lowe, D. J., Wood, J. R., Young, J. M., Churchman, G. J., Huang, Y.-T., et al. (2014). Using palaeoenvironmental DNA to reconstruct past environments: progress and prospects. *J. Quat. Sci.* 29, 610–626. doi: 10.1002/jqs.2740

Reimer, P. J., Bard, E., Baylis, A., Beck, J. W., Blackwell, P. G., Bronk Ramsey, C., et al. (2013). IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* 55, 1869–1887. doi: 10.2458/azu_js_rc.55.16947

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584. doi: 10.7717/peerj.2584

Rohland, N., Glocke, I., Aximu-Petri, A., and Meyer, M. (2018). Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. *Nat. Protoc.* 2, 1756–1762. doi: 10.1038/nprot.2007.247

Rohland, N., and Hofreiter, M. (2007). Comparison and optimization of ancient DNA extraction. *BioTechniques* 42, 343–352. doi: 10.2144/000112383

Rohland, N., and Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* 22, 939–946. doi: 10.1101/gr.128124.111

Rothman, J. M., Nkurunungi, J. B., Shannon, B. F., and Bryer, M. A. (2014). "High altitude diets: implications for the feeding and nutritional ecology of mountain gorillas," in *High Altitude Primates*, eds N. B. Grow, S. Gursky-Doyen, and A. Krzton (New York, NY: Springer), 247–264. doi: 10.1007/978-1-4614-8175-1_14

Roy, J., Arandjelovic, M., Bradley, B. J., Guschanski, K., Stephens, C. R., Bucknell, D., et al. (2014). Recent divergences and size decreases of eastern gorilla populations. *Biol. Lett.* 10:20140811. doi: 10.1098/rsbl.2014.0811

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One* 7:e34131. doi: 10.1371/journal.pone.0034131

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 9, 811–814. doi: 10.1038/nmeth.2066

Seymour, M., Durance, I., Cosby, B. J., Ransom-Jones, E., Deiner, K., Ormerod, S. J., et al. (2018). Acidity promotes degradation of multi-species environmental DNA in lotic mesocosms. *Commun. Biol.* 1:4.

Sherratt, E., del Rosario Castañeda, M., Garwood, R. J., Mahler, D. L., Sanger, T. J., Herrel, A., et al. (2015). Amber fossils demonstrate deep-time stability of Caribbean lizard communities. *Proc. Natl. Acad. Sci. U.S.A.* 112, 9961–9966. doi: 10.1073/pnas.1506516112

Slon, V., Hopfe, C., Weiß, C. L., Mafessoni, F., de la Rasilla, M., Lalueza-Fox, C., et al. (2017). Neandertal and Denisovan DNA from Pleistocene sediments. *Science* 356, 605–608. doi: 10.1126/science.aam9695

Speight, M. C. D., and Blackith, R. E. (1983). "The animals," in *Ecosystems of the World. 4A Mires: Swamp, Bog, Fen and Moor General Studies*, ed. A. J. P. Gore (Amsterdam: Elsevier), 349–365.

Stanford, C. B., and Nkurunungi, J. B. (2003). Behavioral ecology of sympatric chimpanzees and gorillas in Bwindi Impenetrable National Park, Uganda: diet. *Int. J. Primatol.* 24, 901–918. doi: 10.1023/A:1024689008159

Starke, R., and Morais, D. (2019). Gene copy of the 16S rRNA gene cannot outweigh methodological biases of sequencing. *bioRxiv* [Preprint]. doi: 10.1101/813477

Stevens, G. C. (1989). The latitudinal gradient in geographical range: how so many species coexist in the tropics. *Am. Nat.* 133, 240–256. doi: 10.1086/284913

Stoof-Leichsenring, K. R., Epp, L. S., Trauth, M. H., and Tiedemann, R. (2012). Hidden diversity in diatoms of Kenyan Lake Naivasha: a genetic approach detects temporal variation. *Mol. Ecol.* 21, 1918–1930. doi: 10.1111/j.1365-294X.2011.05412.x

Strickler, K. M., Fremier, A. K., and Goldberg, C. S. (2015). Quantifying effects of UV-B, temperature, and pH on eDNA degradation in aquatic microcosms. *Biol. Conserv.* 183, 85–92. doi: 10.1016/j.biocon.2014.11.038

Stuiver, M., and Reimer, P. J. (1993). Extended 14C data base and revised CALIB 3.0 14C age calibration program. *Radiocarbon* 35, 215–230. doi: 10.1017/S0033822200013904

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., et al. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551, 457–463. doi: 10.1038/nature24621

Thomsen, P. F., and Willerslev, E. (2015). Environmental DNA – an emerging tool in conservation for monitoring past and present biodiversity. *Biol. Conserv.* 183, 4–18. doi: 10.1016/j.biocon.2014.11.019

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* 12, 902–903. doi: 10.1038/nmeth.3589

Vågene, Å. J., Herbig, A., Campana, M. G., Robles García, N. M., Warinner, C., Sabin, S., et al. (2018). *Salmonella enterica* genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat. Ecol. Evol.* 2, 520–538. doi: 10.1038/s41559-017-0446-6

Vázquez-Baeza, Y., Gonzalez, A., Smarr, L., McDonald, D., Morton, J. T., Navas-Molina, J. A., et al. (2017). Bringing the dynamic microbiome to life with animations. *Cell Host Microbe* 21, 7–10. doi: 10.1016/j.chom.2016.12.009

Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A., and Knight, R. (2013). EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience* 2:16. doi: 10.1186/2047-217X-2-16

Vuillemin, A., Horn, F., Alawi, M., Henny, C., Wagner, D., Crowe, S. A., et al. (2017). Preservation and significance of extracellular DNA in ferruginous sediments from late Towuti, Indonesia. *Front. Microbiol.* 8:1440. doi: 10.3389/fmicb.2017.01440

Wanek, J., and Rühli, F. (2016). Risks to fragmented DNA in dry, wet, and frozen states from computed tomography: a comparative theoretical study. *Radiat. Environ. Biophys.* 55, 229–241. doi: 10.1007/s00411-016-0637-6

Wanek, J., Speller, R., and Rühli, F. (2013). Direct action of radiation on mummified cells: modeling of computed tomography by Monte Carlo algorithms. *Radiat. Environ. Biophys.* 52, 397–410. doi: 10.1007/s00411-013-0471-z

Willerslev, E., Davison, J., Moora, M., Zobel, M., Coissac, E., Edwards, M. E., et al. (2014). Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature* 506, 47–51. doi: 10.1038/nature12921

Willerslev, E., Hansen, A. J., Binladen, J., Brand, T. B., Gilbert, M. T. P., Shapiro, B., et al. (2003). Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science* 300, 791–795. doi: 10.1126/science.1084114

Wilson, E. O. (1985). Invasion and extinction in the West Indian ant fauna: evidence from the Dominican amber. *Science* 229, 265–267. doi: 10.1126/science.229.4710.265

Wing, S. L., Herrera, F., Jaramillo, C. A., Gómez-Navarro, C., Wilf, P., and Labandeira, C. C. (2009). Late Paleocene fossils from the Cerrejón Formation, Colombia, are the earliest record of Neotropical rainforest. *Proc. Natl. Acad. Sci. U.S.A.* 106, 18627–18632. doi: 10.1073/pnas.0905130106

Wright, H. E., Mann, D. H., and Glaser, P. H. (1984). Piston corers for peat and lake sediments. *Ecology* 65, 657–659. doi: 10.2307/1941430

Ye, S. H., Siddle, K. J., Park, D. J., and Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic classification. *Cell* 178, 779–794. doi: 10.1016/j.cell.2019.07.010

Yoccoz, N. G., Bråthen, K. A., Gielly, L., Haile, J., Edwards, M. E., Goslar, T., et al. (2012). DNA from soil mirrors plant taxonomic and growth from diversity. *Mol. Ecol.* 21, 3647–3655. doi: 10.1111/j.1365-294X.2012.05545.x

Ziesemer, K. A., Mann, A. E., Sankaranarayanan, K., Schroeder, H., Ozga, A. T., Brandt, B. W., et al. (2015). Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. *Sci. Rep.* 5:16498. doi: 10.1038/srep16498

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# Advantages of publishing in Frontiers

**OPEN ACCESS**
Articles are free to read for greatest visibility and readership

**FAST PUBLICATION**
Around 90 days from submission to decision

**HIGH QUALITY PEER-REVIEW**
Rigorous, collaborative, and constructive peer-review

**TRANSPARENT PEER-REVIEW**
Editors and reviewers acknowledged by name on published articles

**Frontiers**
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

**Visit us:** www.frontiersin.org
**Contact us:** frontiersin.org/about/contact

**REPRODUCIBILITY OF RESEARCH**
Support open data and methods to enhance research reproducibility

**DIGITAL PUBLISHING**
Articles designed for optimal readership across devices

**FOLLOW US**
@frontiersin

**IMPACT METRICS**
Advanced article metrics track visibility across digital media

**EXTENSIVE PROMOTION**
Marketing and promotion of impactful research

**LOOP RESEARCH NETWORK**
Our network increases your article's readership