



Expanding the ancient DNA bioinformatics toolbox, and its applications to archeological microbiomes

Dissertation

*in Partial Fulfilment of the Requirements for the Degree of
"Doctor of Philosophy" (PhD)*

Submitted to the Council of the Faculty of Biological Sciences of
Friedrich Schiller University Jena

by B.Sc. M.Sc. **Maxime Borry**

born on September 23rd, 1991 in Boulogne-Billancourt, France

Gutachter

1. Prof. Dr. Christina Warinner (Max Planck Institute for Evolutionary Anthropology, Leipzig, DE / Harvard University, Boston, USA / Friedrich-Schiller-Universität Jena, Jena, DE)
2. Prof. Dr. Gianni Panagiotou (Leibniz Institute for Natural Product Research and Infection Biology, Jena, DE / Friedrich-Schiller-Universität Jena, Jena, DE)
3. Dr. Celine Bon (CNRS-MNHN UMR 7206 - Anthropologie Génétique, Paris, FR)

Beginn der Promotion: 09.09.2018

Dissertation eingereicht am: 10.03.2023

Tag der öffentlichen Verteidigung: 01.08.2023

The important thing is not to stop questioning. Curiosity has its own reason for existing. One cannot help but be in awe when he contemplates the mysteries of eternity, of life, of the marvelous structure of reality. It is enough if one tries merely to comprehend a little of this mystery every day.

Albert Einstein

Contents

Introduction	1
The rise of the microbiome	1
The origin of the microbiome concept	1
The amplicon based era	2
From amplicons to shotgun metagenomics	3
Methods for metagenomics data analysis	6
Taxonomic classifiers	6
Source tracking	8
<i>De novo</i> short read assembly	9
Contig binning	11
Functional annotation	11
Ancient DNA and its challenges for metagenomics	12
History of ancient DNA metagenomics	12
The challenges of aDNA	12
aDNA methods	13
Overview of the manuscripts	16
1 Microbial ecology to the rescue for identifying the origin of paleofeces	18
Manuscript A: <i>Sourcepredict: Prediction of metagenomic sample sources using dimension reduction followed by machine learning classification</i>	18
Manuscript B: <i>CoproID predicts the source of coprolites and paleofeces using microbiome composition and host DNA content</i>	22
2 From alignments to assemblies	47

Manuscript C: <i>sam2lca: Lowest Common Ancestor for SAM, BAM, CRAM alignment files</i>	47
Manuscript D: <i>PyDamage: automated ancient damage identification and estimation for contigs in ancient DNA de novo assembly</i>	52
3 Application to the fermentation microbiome	76
Manuscript E: <i>Fermentation microbiome analysis of biblical king Herod's wine</i>	76
Discussion	102
Machine learning methods to predict the source of microbiome samples	103
The required scalability of metagenomics methods	104
Ancient microbiomes need not to be only human	106
Conclusion	107
References	108
Abbreviations	117
Eidesstattliche Erklärung	118
Detailed contributions	119

Acknowledgements

First and foremost, I would like to thank my supervisors, Prof. Dr. Christina Warinner and Dr. Alexander Herbig for giving the opportunity to work on these exciting and challenging topics. Similarly, I would like to express my gratitude to the different funding bodies who made this work financially possible. I would also like to thank my colleagues at the MPI-SHH and MPI-EVA for the fruitful collaborations and stimulating discussions, and my different collaborators for their respective contributions to my projects. A very special thanks to my family and parents, for enabling me to pursue my studies in the field that stimulated my interests. Last but not least, I would like to thank my partner for her uninterrupted support during this PhD, and her precious comments on this thesis.

Abstract

The 1980s were very prolific years not only for music, but also for molecular biology and genetics, with the first publications on the microbiome and ancient DNA. Several technical revolutions later, the field of ancient metagenomics is now progressing full steam ahead, at a never seen before pace.

While generating sequencing data is becoming cheaper every year, the bioinformatics methods and the compute power needed to analyze them are struggling to catch up. In this thesis, I propose new methods to reduce the sequencing to analysis gap, by introducing scalable and parallelized softwares for ancient DNA metagenomics analysis.

In manuscript A, I first introduce a method for estimating the mixtures of different sources in a sequencing sample, a problem known as source tracking. I then apply this method to predict the original sources of paleofeces in manuscript B.

In manuscript C, I propose a new method to scale the lowest common ancestor calling from sequence alignment files, which brings a solution for the computational intractability of fitting ever growing metagenomic reference database indices in memory.

In manuscript D, I present a method to statistically estimate in parallel the ancient DNA deamination damage, and test it in the context of *de novo* assembly.

Finally, in manuscript E, I apply some of the methods developed in this thesis to the analysis of ancient wine fermentation samples, and present the first ancient genomes of ancient fermentation bacteria.

Taken together, the tools developed in this thesis will help the researchers working in the field of ancient DNA metagenomics to scale their analysis to the massive amount of sequencing data routinely produced nowadays.

Zusammenfassung

Die 1980er Jahre waren nicht nur für die Musikindustrie, sondern auch für Molekularbiologie und Genetik sehr fruchtbare Jahre, in denen die ersten Veröffentlichungen über das Mikrobiom und alte DNA erschienen. Mehrere technische Revolutionen später schreitet das Gebiet der Metagenomics mit Hochdruck voran, und zwar in einem noch nie dagewesenen Tempo.

Während die Erzeugung von Sequenzierdaten jedes Jahr billiger wird, haben die Bioinformatikmethoden und die für ihre Analyse erforderliche Rechenleistung Mühe, mit dieser Entwicklung Schritt zu halten. In dieser Arbeit schlage ich neue Methoden vor, um die Kluft zwischen Sequenzierung und Analyse zu verringern, indem ich skalierbare und parallelisierte Software für die Analyse alter DNA Metagenomics einführe.

In Manuskript A stelle ich zunächst eine Methode zur Schätzung der Mischungen verschiedener Quellen in einer Sequenzierprobe vor, ein Problem, das als "source tracking" bekannt ist. In Manuskript B wende ich diese Methode dann an, um die ursprünglichen Quellen von Paläofäkalien vorherzusagen.

In Manuskript C schlage ich eine neue Methode zur Skalierung des Aufrufs des "lowest common ancestor" aus Sequenzabgleichsdateien vor. Diese Methode bietet eine Lösung für die rechnerische Schwierigkeit der Anpassung immer größer werdender metagenomischer Referenzdatenbankindizes und deren Arbeitsspeicher.

In Manuskript D präsentiere ich eine Methode zur parallelen statistischen Schätzung alter DNA-Desaminationschäden und teste sie im Kontext der Textitde novo-Assemblierung.

In Manuskript E schließlich wende ich einige der in dieser Arbeit entwickelten Methoden auf die Analyse alter Weinfermentationsproben an und präsentiere die ersten Genome alter Fermentationsbakterien.

Insgesamt werden die in dieser Arbeit entwickelten Werkzeuge den Forschern, die auf dem Gebiet der Metagenomics alter DNA arbeiten, helfen, ihre Analysen auf die riesige Menge an Sequenzierdaten zu skalieren, die heutzutage routinemäßig produziert werden.

Introduction

The rise of the microbiome

The origin of the microbiome concept

In 2001, the Physiology or Medicine Nobel prize laureate Joshua Lederberg is said to have coined the term microbiome and gave it the following definition:

"The ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space and have been all but ignored as determinants of health and disease." Lederberg and McCray (2001).

But did we really have to wait until 2001 for the appearance of the microbiome? When digging a bit deeper, one quickly realizes that the origin of the word predates the turn of the century (Prescott, 2017), and that the concept of microbiome had already been formulated in 1988.

"A convenient ecological framework in which to examine biocontrol systems is that of the microbiome. This may be defined as a characteristic microbial community occupying a reasonably well defined habitat which has distinct physio-chemical properties. The term thus not only refers to the microorganisms involved but also encompasses their theatre of activity." Whipps et al. (1988)

In these 35 years of microbiome research, the field has immensely evolved, several technological revolutions have been brought to the world of molecular biology, and the scale of the research questions grew by several orders of magnitude.

The amplicon based era

The first revolution came with the advent of the Polymerase Chain Reaction (PCR) technique by Mullis et al. (1986) (Fig 0.1).

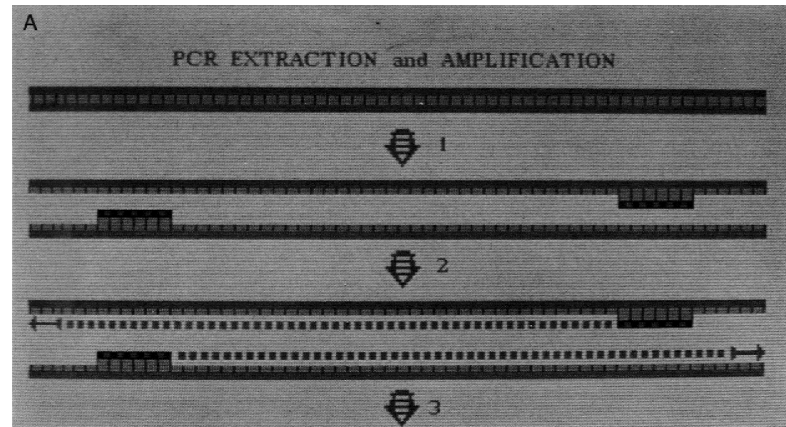


Figure 0.1: Excerpt of the original figure describing the PCR by of Mullis et al. (1986). A double stranded DNA molecule (top) is denaturated, after which primers anneal to their target sequence (middle), followed by elongation where a DNA polymerase synthesizes for each strand a new complementary one (bottom). This operation is repeated throughout multiple cycles to obtain an exponential amplification.

Thanks to the PCR, it became possible to easily generate numerous DNA copies, the amplicons, of target regions of interest, a step necessary for further cloning and sequencing analysis. Coincidentally, the sequence of interest for bacteria were identified concomitantly: the 16S ribosomal RNA (rRNA) markers (Woese, 1987). The combination of 16S rRNA primers and PCR allowed for a rapid expansion of our knowledge on bacterial phylogenetics (Pace, 1997). Primers for other markers were also developed for other clades such as *rbcL* for plants, or *cytB* for vertebrates for example. However, for bacterial community ecology, the impact remained relatively limited: for each 16S rRNA PCR amplification, the PCR products needed to be separated by cloning into competent bacteria, and each clone individually sequenced with Sanger sequencing technique (Sanger et al., 1977). This was both time and cost ineffective, as well as relatively low throughput.

That all changed at the turn of the century with the next technological revolution which became known as high throughput sequencing (Heather and Chain, 2016). It was now possible to directly sequence the DNA after extraction, completely discarding the cloning step (Tringe and Hugenholtz, 2008). And with a constantly decreasing sequencing cost of DNA sequencing (Fig 0.2), the door was thus open to generating data at an

unprecedented pace. The NGS sequencing of 16S rRNA gene amplicons became known as metataxonomics (Marchesi and Ravel, 2015). While metataxonomics contributed on its own to revealing a large part of the earth biodiversity, from the deep sea (Huber et al., 2007; Sogin et al., 2006), to ocean thermal vents (McCliment et al., 2006), antarctic soil (Soo et al., 2009), and human gut microbiome (Human Microbiome Project Consortium, 2012; Yatsunenکو et al., 2012), it also suffered from some limitations, mainly the 16S rRNA gene's lack of fine scale taxonomic resolution.

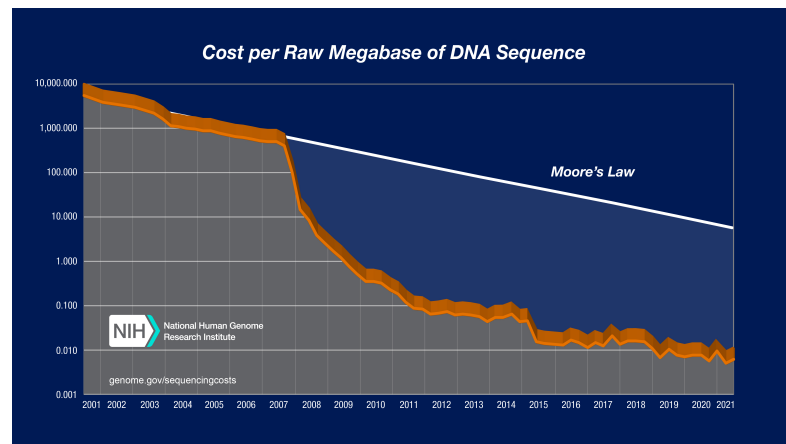


Figure 0.2: **Evolution of the DNA sequencing cost per Megabase.** Adapted from NIH (2022)

From amplicons to shotgun metagenomics

With the sheer drop in DNA sequencing prices, and the limitations of metataxonomics, a new approach was envisioned, the whole genome shotgun (WGS) sequencing of environmental DNA, known as metagenomics. With this method, there is no need to PCR pre-amplify target DNA regions, as all environmental DNA, the so-called metagenome, is turned into sequencing libraries. Metagenomics not only allowed the exploration of the unsampled taxonomic diversity of previously unknown organisms, what would later be given the name of microbial dark-matter (Jiao et al., 2021), but also enriched our knowledge in the functional capacities of metagenomes with the sequencing of all its genes (Tringe and Rubin, 2005).

This drastic cost and time to sequencing reduction allowed for a substantial shift in the scale of microbiome studies, which led to major publications, such as human microbiome project, looking at the diversity of healthy human microbiomes (Human Microbiome Project Consortium, 2012), with a followup a few years later (Proctor et al.,

2019) including individuals with specific diseases. These massive scale studies allowed to study at a population scale the link between microbiomes, diet, lifestyle, and diseases, but also the diversity of the different human microbiomes. Among these different microbiomes, the human gut was identified as the habitat with the most microbial diversity between and within human populations, either living a westernized lifestyle, with easy access to processed foods and medicine, or a non-westernized more traditional lifestyle (Obregon-Tito et al., 2015; Schnorr et al., 2014).

Together with the ever decreasing cost of computational resources, and new algorithms such as metagenomics *de novo* assemblies, the affordable sequencing of an ever increasing number of sample per study paved the way for the reference-free reconstruction of entirely unknown microbial species, such as from the cow rumen (Stewart et al., 2019), or more recently from the human gut microbiome (Pasolli et al., 2019). These studies demonstrated the importance of the microbial dark matter (Fig 0.3), and help to shed a light on these previously unknown and uncharacterized micro-organisms, by reconstructing new genomes from scratch.

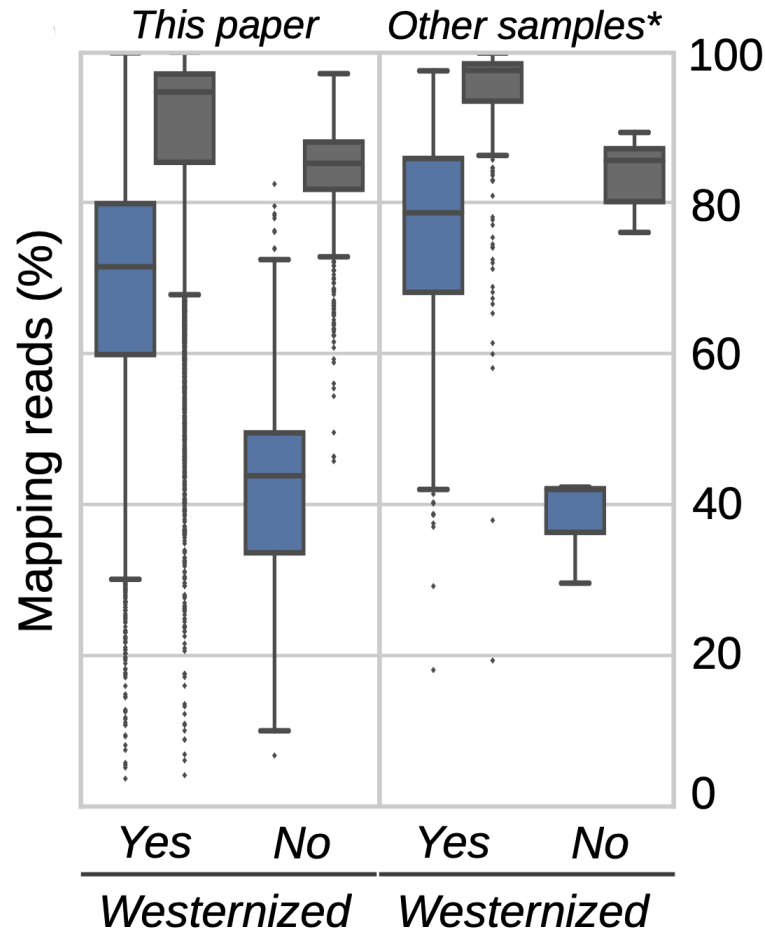


Figure 0.3: **Read mappability of metagenomes.** The proportions of reads mapped to a database containing only known reference genomes are in blue, while the proportions of reads mapped to a database containing known reference genomes, and genomes reconstructed from the dark matter are in gray. With the addition of genomes reconstructed from the microbial dark matter, the proportion of mapped reads increases. This applies for samples coming from both westernized, and non-westernized individuals, even for samples not used to reconstruct the genome of dark matter microbes. Adapted from Pasolli et al. (2019).

Methods for metagenomics data analysis

With the development of shotgun metagenomics arose the necessity of developing new computational methods to process the massive amount of sequencing data (Breitwieser et al., 2017; Sharpton, 2014). These tools mainly help to answer two different questions: "Who is there?", and "What are they doing?" While an introduction to the core-principles behind these different tools follows in the subsequent sections, a benchmark of the performance of many of these methods has been made available with the results of the CAMI challenges (Meyer et al., 2022; Sczyrba et al., 2017).

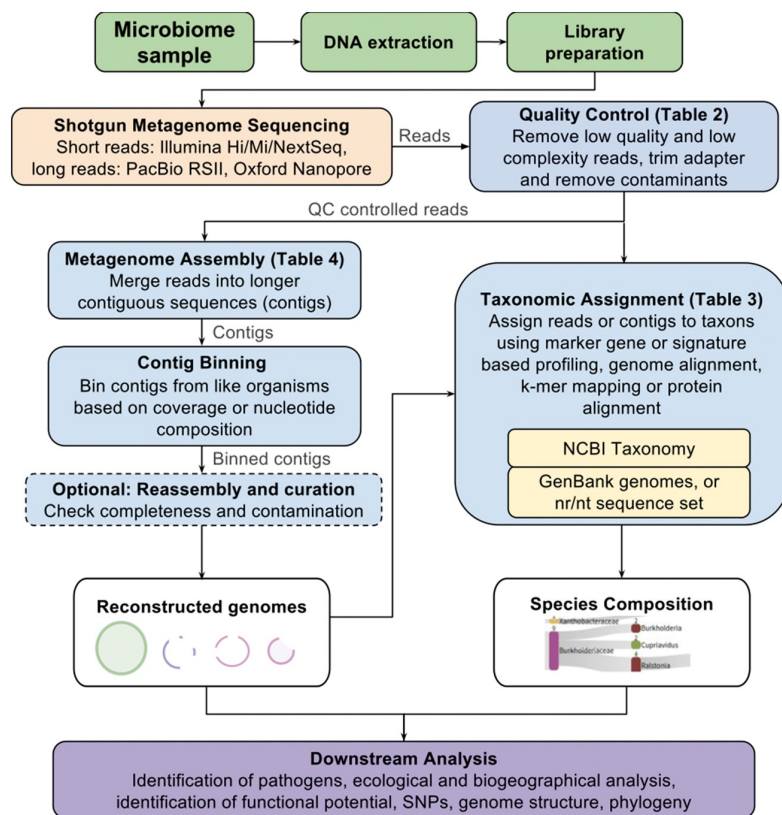


Figure 0.4: **Common analysis procedures for metagenomics data.** Adapted from Breitwieser et al. (2017)

Taxonomic classifiers

One of the first questions to answer when faced with a metagenomic sequencing library, is "Who is there?" or formulated differently, "What are the different taxa present in the library?". The answer to this question is given by taxonomic classifiers, tools that assign

a taxon to each of the sequencing read present in a sequencing library. These tools can broadly be divided into two different categories: tools using alignment free methods, and tools using alignments. Regardless of the category, all these tools possess an algorithm in common: the lowest common ancestor (LCA). This algorithm is needed to perform a disambiguation when a query read weakly aligns to one or more distantly related reference organisms, or when analyzing short DNA sequences, a query DNA read can match equally well to more than one reference organism, posing a challenge for its taxonomic assignation. The LCA algorithm solves this ambiguity by assigning the query higher in a taxonomic tree, at a less precise taxonomic level (Fig 0.5), an idea first implemented for metagenomics by MEGAN (Huson et al., 2007).

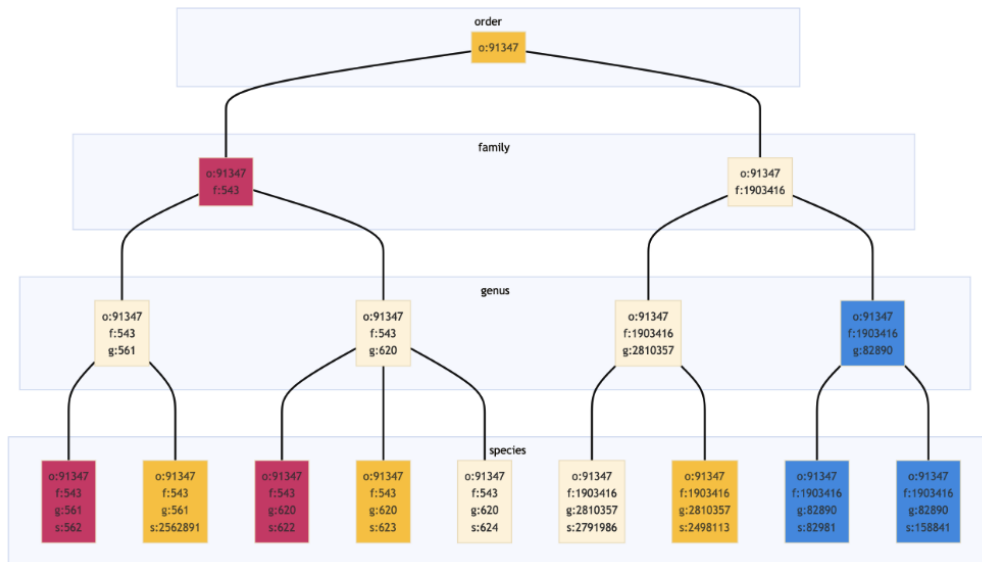


Figure 0.5: **Illustration of the LCA algorithm.** Taxa and their LCA are displayed in the same color. The lineage for each taxon is shown with a one letter code for the rank, and the corresponding taxonomic IDs (TAXID). The LCA of s:562 (*E. coli* species) and s:622 (*S. dysenteriae* species) is f:543 (Enterobacteriaceae family). The LCA of s:82981 (*L. grimontii* species) and s:158841 (*L. richardii* species) is g:82980 (Leminorella genus). The LCA of s:2562891 (*E. alba* species), s:623 (*S. flexneri* species) and s:2498113 (*J. zhutongyuii* species) is o:91347 (Enterobacterales order). Adapted from Borry et al. (2022)

The alignment free methods typically rely on (near) exact matches of shorter DNA fragments of a fixed size k , the k -mers. Typical tools in this category include Kraken (Wood and Salzberg, 2014), KrakenUniq (Breitwieser et al., 2018), Kraken2 (Wood et al., 2019), Clark-S (Ounit and Lonardi, 2016). While these tools usually offer a good compromise between speed and accuracy, some of them by design tend to have an elevated rate of false alignments to favor classification speed.

The alignment based methods can be further subdivided into tools using DNA reference databases, or protein reference databases. While BLAST (Altschul et al., 1990) was originally used to query DNA reference databases, it very quickly became inadequate for the size of metagenomics datasets. To remedy to this, MALT/MEGAN (Herbig et al., 2016; Huson et al., 2007) relies on slightly different algorithmic choices to perform searches of metagenomic libraries against reference databases containing the genome of all sequenced organisms. Nevertheless, with the ever increasing size of reference databases, even cleverer algorithmic choices aren't enough, and these too soon also became computationally intractable. An alternative approach chosen by the metaphlan family of tools (Blanco-Miguez et al., 2022; Segata et al., 2012; Truong et al., 2015) is to rely on a curated set of clade specific marker genes, which keeps the database size in check. This is also the approach taken by GTDB-TK (Chaumeil et al., 2020), a tool relying on a different taxonomic system, the genome taxonomy database (GTDB) (Parks et al., 2022), while all previously mentioned tools rely on the NCBI taxonomy.

Regarding alignment based taxonomic classifiers relying on protein databases, they benefit from the 3 fold reduction of the length of reference sequences due to the translation of DNA sequences in protein sequences, and the higher conservation of protein sequences due to the redundancy of the genetic code, and evolutionary pressure on proteins. Tools in this category include DIAMOND (Buchfink et al., 2015), Kaiju (Menzel et al., 2016), and MMseqs2 (Steinegger and Söding, 2017). In combination with modern indexing algorithm, these tools still manage to classify metagenomics query sequences using databases of all known protein sequences in a timely manner.

Source tracking

A question often related to *"Who is there?"*, is *"Where are these taxons coming from?"*. Even in an ideal situation, contamination of a metagenomic sample by external taxons is a possibility. To check for this contamination, the concept of source tracking was developed. Akin to the genetic concept of admixture, it aims to identify the mixture proportions of different sources in the test sample, also known as the sink. The most established software to perform source tracking, SourceTracker (Knights et al., 2011) relies on a markov chain monte carlo (MCMC) approach to determine the mixing proportions. However, because the convergence of the MCMC can take a long time, alternative approaches have been developed relying on expectation-maximization such as FEAST (Shenhav et al., 2019), or the recently published DECOM (González et al., 2023), using k-mer counts matrix operations.

De novo short read assembly

While previously mentioned strategies all relied on a comparison with an already existing reference database, the *de novo* assembly allows the exploration of metagenomics data in a reference free manner. The basic idea of *de novo* assembly is to re-assemble the sequencing reads, short DNA fragments ranging from 30 to 400 bp, into longer ones, called contigs, by using the overlaps between reads. While there are different algorithms available to perform assembly, the main strategy for metagenomics assembly relies on de Bruijn graphs. De Bruijn graphs model the overlap between reads by first dividing them into shorter sequences of length k , known as k -mer. The unique k -mers are then represented in a graph as edges, connecting nodes made up of all unique k -mer prefixes and suffixes. After the graph has been built, an Eulerian path is then computed to reconstruct the original DNA sequence (Fig 0.6).

While the choice of k originally played an important role, with smaller k values leading to more assemblies, and larger k allowing for a better handling of the repeated regions, modern metagenomics de Bruijn *de novo* assemblers, like IDBA-UD (Peng et al., 2012), metaSPADES (Nurk et al., 2017), or MEGAHIT (Li et al., 2015) use an iterative approach with increasing k , replacing the reads with assembled contigs at each iteration (Breitwieser et al., 2017). In theory, bacterial genomes could be directly assembled in a single contig, however in practice, due to repeated sequences and sequencing errors, multiple contigs will be created, when there are more than one Eulerian cycles (Compeau et al., 2011). The assembly graph will therefore have more than one connected component, each representing a single contig. Trying to overlap these contigs is the so-called scaffolding step, which relies on the paired-end information of sequencing reads. When a read from a read pair is found on a contig, and the other member of the pair on another contig, then two contigs can be gathered in scaffold. Another scaffolding strategy relies on adding additional information, with the use of long read sequencing technologies to bridge gaps between contigs.

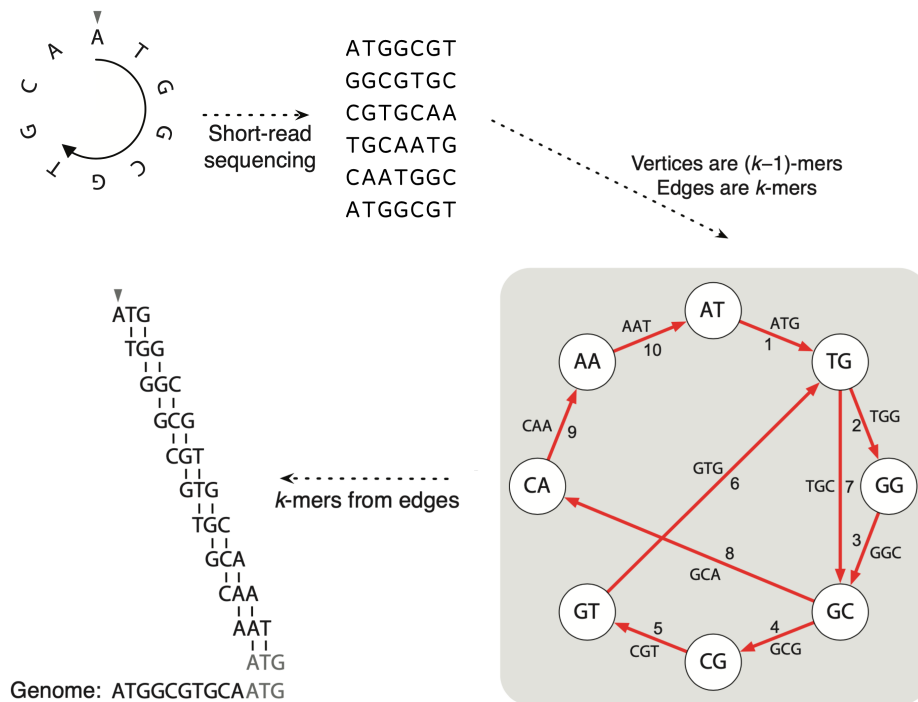


Figure 0.6: **Illustration of de Bruijn graph based *de novo* assembly.** Sequencing reads are first divided in all possible k -mers (here $k = 3$). A node is then formed for each unique k -mer prefix and suffix ($k - 1$ -mer), and nodes are connected with directed edges if a k -mer contains both prefix and suffix. The next step is to find an Eulerian cycle, meaning a path that visits all edges of the graph exactly once. The Eulerian cycle is the reconstructed contig. Adapted from Compeau et al. (2011)

Contig binning

Whether the scaffolding step was successful or not, there are still numerous cases where multiple shorter contigs or scaffolds were assembled from a sequencing library. Furthermore, in a metagenomic context, these assemblies are potentially coming from multiple organisms. The next task is then to cluster them into taxon bins, or so called metagenome assembled genomes (MAGs). While one possibility is to use taxonomic classifier to regroup contigs by taxons, this approach is often disfavored because it can not cluster contigs assembled from previously unknown organisms. Reference-free clustering therefore relies on nucleotide composition, such as tetranucleotide (k -mer with $k = 4$) frequencies, and/or coverage information of reads mapped to contigs. Popular binning approaches include CONCOCT (Alneberg et al., 2014), MaxBin2 (Wu et al., 2016), or MetaBAT2 (Kang et al., 2019).

Functional annotation

Once the question "*Who is there?*" has been answered, the question "*What are they doing?*" still remains.

While it is possible to directly infer the functional capacity from the sequencing reads by querying them against a protein, or a gene coding sequence (CDS) database, in practice, it is often more informational to first reconstruct MAGs, and then annotate them with regional and functional information. Tools to perform such annotations, such as Prokka (Seemann, 2014) and Bakta (Schwengers et al., 2021), typically rely on a series of external feature prediction tools, like Prodigal for identifying CDS (Hyatt et al., 2010), protein alignments with blastp (Camacho et al., 2009) or DIAMOND (Buchfink et al., 2015), and a variety of hidden markov models (HMM) (Eddy, 2011) and protein domain profiles to assign functions to more distantly related proteins.

Ancient DNA and its challenges for metagenomics

History of ancient DNA metagenomics

The first ever ancient DNA (aDNA) sequence was published in 1984 with a segment of the quagga mitochondrial genome, retrieved from dried muscle tissues of a museum specimen of this extinct zebra species (Higuchi et al., 1984). However, it is only 20 years later, in 2003 that the first ancient metataxonomic study was published, using 16S, *rbcL*, and *cytB* markers to study the flora, fauna, and microbiota of permafrost and cave sediments (Willerslev et al., 2003). Three years later, in 2006, came the first metagenomics shotgun sequencing dataset, from a 29 000 years old mammoth mandible retrieved in Siberia (Poinar et al., 2006). Since then, many metagenomics studies have been conducted including samples from a variety of environments, such as dental pulp (Bos et al., 2011), dental calculus (Warinner et al., 2014), paleofeces (Tito et al., 2008), ancient chewing gum (Jensen et al., 2019), and more. While the study of different environments gives us more information about their respective microbiomes, they are also used to answer different research questions. For example, dental pulp tissues may contain traces of blood, which can contain blood-borne pathogens such as *Yersinia pestis*, and is therefore used to study ancient epidemics such as the black death (Bos et al., 2011). Other environments such as paleofeces not only allow us to study the gut microbiome, and its associated diseases, but also the diet of the individuals. Finally, the study of ancient microbiomes should not be limited to human samples. Humans have been living in a symbiotic relationship not only with intra-corporal micro-organisms, but also with extra-corporal microbes, helping to produce their fermented foods and beverages. The study of ancient fermentation vessels and artifacts will also bring extremely valuable insights on past culinary and cultural practices.

The challenges of aDNA

While aDNA metagenomics borrows a lot of wetlab and computational method to its modern counterparts, the characteristics of ancient nucleic acids pose their own set of challenges due to their post-mortem degradation.

One of these challenges is the fragmentation of DNA in very short segments (most often $< 100bp$), due a hydrolytic depurination and β -excision (Orlando et al., 2021), a phenomena naturally occurring several thousand times per day in every living cell (Lindahl, 1993), but normally repaired through base excision repair (BER) pathway by DNA polymerases (Fig 0.7).

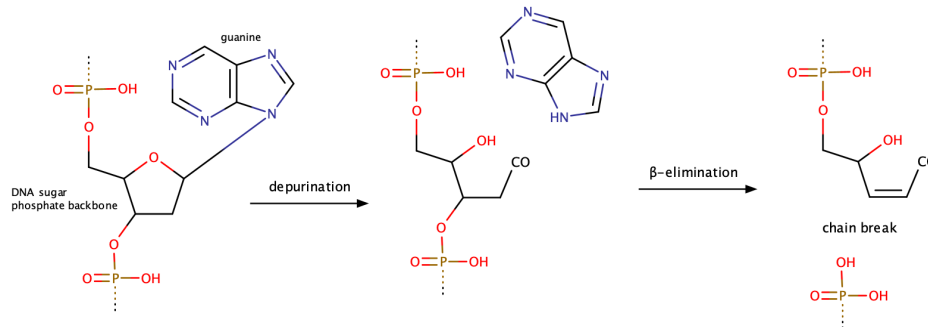


Figure 0.7: **DNA depurination, followed by β -elimination.** Which leads to a "nick", or break, in the DNA backbone.

The other challenge is the cytosine deamination into uracils, read as thymine by DNA sequencers, the so called C to T misincorporation (or G to A on reverse strand). This process is mostly happening at the end of aDNA fragments, because of the above-mentioned single-stranded DNA breaks, leaving exposed overhanging DNA termini. While this specific process can lead to an artificial increase of mutations when aligning the aDNA sequences to a reference genome, it also creates a damage pattern characteristic to aDNA (Fig 0.8), which helps validate the archaeological authenticity of the nucleic acids (Orlando et al., 2021).

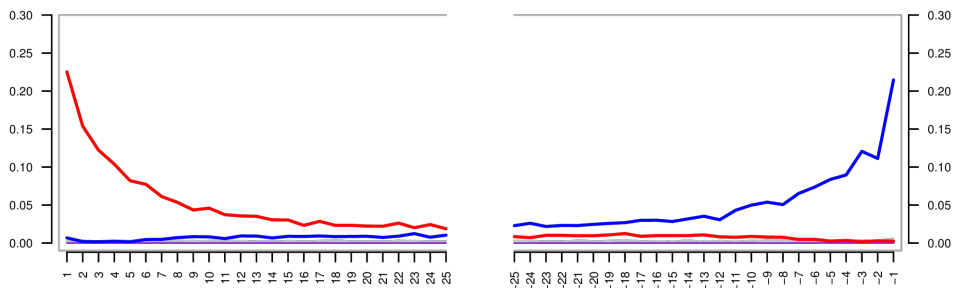


Figure 0.8: **Characteristic aDNA damage pattern, colloquially known as the "smiley plot".** The red line represents the C to T misincorporation rate from the 5' end on the forward read, while the blue line represents the G to A misincorporation rate from the 3' end on the reverse read. Adapted from MapDamage plots (Jónsson et al., 2013).

aDNA methods

Because of these characteristics, both molecular and computational methods were developed to deal with the specificities of aDNA, namely very short DNA sequences, and

ancient DNA damage patterns. For an extensive review of the molecular biology methods specific to ancient DNA, see the recent review of Orlando et al. (2021).

Computational methods for aDNA

The short length of aDNA molecules lead to the development, or the adaptation of a variety of tools. First and foremost, forward and reverse sequencing from paired end libraries often need to be merged on their overlap, because of the negative inner distance (Fig 0.9), with programs such as leehom (Renaud et al., 2014), AdapterRemoval (Schubert et al., 2016), or fastp (Chen et al., 2018). Merging the reads has the added benefit of increasing base calling accuracy on the otherwise more error prone 3' end of the sequencing reads.

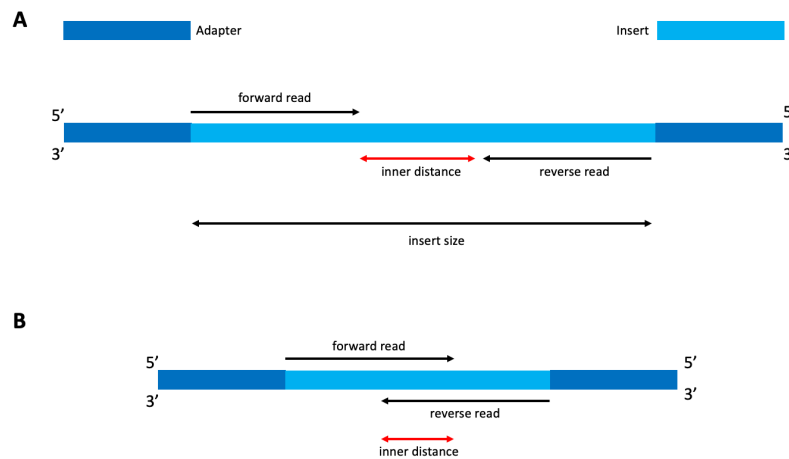


Figure 0.9: **Inner distance difference between modern and ancient DNA.** In modern DNA sequencing libraries (A), the length of the DNA molecule, or the insert size, is often greater than the cumulated length of both the forward and reverse reads, leaving an unsequenced segment of DNA in between, the inner distance. For ancient DNA libraries (B), the length of the DNA molecules is often shorter than the minimal read length, which leads to a negative inner distance, and an overlap of the forward with the reverse read.

To align the pre-processed reads to a single reference, general short read aligners are used, for instance BWA, or Bowtie2, with so-called "ancient DNA parameters". These parameters imply a higher sensitivity at the seeding step, by allowing mismatches, which can happen due to C to T misincorporation.

To visualize and quantify the damage of reads aligned to a single reference, tools such as mapDamage (Jónsson et al., 2013) and DamageProfiler (Neukamm et al., 2021a) have been developed, while tools like PMDtools (Skoglund et al., 2014) serve to filter reads containing aDNA damage.

As the C to T misincorporation pattern is characteristic to ancient DNA, and can only be observed via a sequence alignment, alignment based approaches are often preferred to alignment free approaches in aDNA metagenomics. This led to the development of the MALT metagenomic aligner (Herbig et al., 2016), which in combination with HOPS (Hübler et al., 2019), enable the visualization of the characteristic aDNA damage patterns of a selection of taxons in a metagenomic community.

More recently, *de novo* assembly has been successfully applied to aDNA on single genomes (Seitz and Nieselt, 2017), and metagenomes (Granehall et al., 2021; Wibowo et al., 2021) leading to the reconstruction of hundreds of ancient MAGs after binning.

Overview of the manuscripts

The work presented in this thesis can be divided into three different parts.

- Use and develop microbial ecology techniques, including sourcetracking methods, to identify the host origin of paleofeces from metagenomics shotgun sequencing data (manuscript A and B, chapter 1).
- Develop new methods to address the specificities of ancient DNA taxonomic classification and *de novo* assembly (manuscript C and D, chapter 2).
- Apply these methods to shotgun metagenomics data from wine fermentation microbiome (manuscript E, chapter 3).

The chapters of this thesis are composed of the following manuscripts

Chapter 1: Microbial ecology to the rescue for identifying the origin of paleofeces

- **Manuscript A:** After explaining the limitations of currently available sourcetracking methods, I present a new source tracking and source prediction method. This method combines explainability, scalability, and visualization of the metagenomics samples in a lower dimensional space, and comes with unit and integration tested code.
- **Manuscript B:** I use the above mentioned method, in combination with host endogenous content, to predict the host origin of paleofeces shotgun sequencing samples, and integrate all these steps into a self contained reproducible and scalable data analysis pipeline. We apply this pipeline to published and newly sequenced data, to tell apart dog from human paleofeces.

Chapter 2: From alignments to assemblies

- **Manuscript C:** After pointing the limitations of current alignment based metagenomics classifiers, I present a more scalable method to apply a LCA algorithm to the output of any short read aligner producing a SAM alignment format file. This method comes with unit and integration tested code.

- **Manuscript D:** After introducing the need for a aDNA damage estimation software for metagenomics data, I present a new scalable method to statistically assess C to T aDNA misincorporation damage of multiple references, in parallel. This method comes with unit and integration tested code, and we demonstrated its performances in the context of aDNA *de novo* assembly.

Chapter 3: Application to the fermentation microbiome

- **Manuscript E:** I first assessed the conservation of the fermentation microbiome in ancient wine samples from biblical times. After having identified the remaining microbes involved in the wine fermentation process, I selectively captured the sequencing libraries to enrich these fermentation microbes, and reconstructed their genomes using *de novo* assembly. Finally, I conducted a functional and phylogenetic analysis of these ancient fermentation MAGs.

Microbial ecology to the rescue for identifying the origin of paleofeces

Manuscript A: Sourcepredict: Prediction of metagenomic sample sources using dimension reduction followed by machine learning classification

Maxime Borry

Published in The Journal of Open Source Software, 2019 June 28; DOI: 10.21105/joss.01540

In Manuscript A, I introduce a new method for predicting the origin of a metagenomic sample based on its microbiome composition. The most established method (at the time of the publication of Sourcepredict), SourceTracker (Knights et al., 2011) was designed at the beginning of the expansion of metagenomics, with the goal of assigning source proportions to a test sample, also called sink. While it has been widely adopted by the field of microbiome research, it suffered from two main drawbacks. First, as it relies on a MCMC approach, it often suffers from very long convergence time, sometimes days, especially when using a multiplicity of sources. Furthermore, because of the stochasticity of the MCMC approach, results differ between runs, conferring it a black-box aspect, with hard to interpret outcomes. To circumvent these issues, I proposed the Sourcepredict method, which relies on a faster fuzzy clustering source prediction method, operating on a β -diversity pairwise distance matrix embedded, and visualized, in a lower dimensional space. When benchmarked against SourceTracker, Sourcepredict showed similar or better performances for both the tasks of source prediction (Borry, 2019a) and source tracking (mixture of sources) (Borry, 2019b).

Sourcepredict: Prediction of metagenomic sample sources using dimension reduction followed by machine learning classification

Maxime Borry¹

¹ Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, 07745, Germany

DOI: [10.21105/joss.01540](https://doi.org/10.21105/joss.01540)

Software

- [Review](#) ↗
- [Repository](#) ↗
- [Archive](#) ↗

Submitted: 28 June 2019

Published: 04 September 2019

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC-BY](#)).

Summary

SourcePredict is a Python package distributed through Conda, to classify and predict the origin of metagenomic samples, given a reference dataset of known origins, a problem also known as source tracking.

DNA shotgun sequencing of human, animal, and environmental samples has opened up new doors to explore the diversity of life in these different environments, a field known as metagenomics (Hugenholtz & Tyson, 2008). One aspect of metagenomics is investigating the community composition of organisms within a sequencing sample with tools known as taxonomic classifiers, such as Kraken (Wood & Salzberg, 2014).

In cases where the origin of a metagenomic sample, its source, is unknown, it is often part of the research question to predict and/or confirm the source. For example, in microbial archaeology, it is sometimes necessary to rely on metagenomics to validate the source of paleofaeces. Using samples of known sources, a reference dataset can be established with the taxonomic composition of the samples, i.e., the organisms identified in the samples as features, and the sources of the samples as class labels.

With this reference dataset, a machine learning algorithm can be trained to predict the source of unknown samples (sinks) from their taxonomic composition.

Other tools used to perform the prediction of a sample source already exist, such as SourceTracker (Knights et al., 2011), which employs Gibbs sampling.

However, the Sourcepredict results are more easily interpreted since the samples are embedded in a human observable low-dimensional space. This embedding is performed by a dimension reduction algorithm followed by K-Nearest-Neighbours (KNN) classification.

Method

Starting with a numerical organism count matrix (samples as columns, organisms as rows, obtained by a taxonomic classifier) of merged references and sinks datasets, samples are first normalized relative to each other, to correct for uneven sequencing depth using the geometric mean of pairwise ratios (GMPR) method (default) (L. Chen et al., 2018).

After normalization, Sourcepredict performs a two-step prediction algorithm. First, it predicts the proportion of unknown sources, i.e., which are not represented in the reference dataset. Second, it predicts the proportion of each known source of the reference dataset in the sink samples.

Organisms are represented by their taxonomic identifiers (TAXID).

Prediction of the proportion of unknown sources

Let $S_i \in \{S_1, \dots, S_n\}$ be a sample from the normalized sinks dataset D_{sink} , $o_j^i \in \{o_1^i, \dots, o_{n_o^i}^i\}$ an organism in S_i , and n_o^i the total number of organisms in S_i , with $o_j^i \in \mathbb{Z}^+$. Let m be the mean number of samples per source in the reference dataset, such that $m = \frac{1}{O} \sum_{i=1}^O S_i$. For each S_i sample, I define $\|m\|$ derivative samples $U_k^{S_i} \in \{U_1^{S_i}, \dots, U_{\|m\|}^{S_i}\}$ to add to the reference dataset to account for the unknown source proportion in a test sample. Separately for each S_i , a proportion denoted $\alpha \in [0, 1]$ (default = 0.1) of each o_j^i organism of S_i is added to each $U_k^{S_i}$ sample such that $U_k^{S_i}(o_j^i) = \alpha \cdot x_{i,j}$, where $x_{i,j}$ is sampled from a Gaussian distribution $\mathcal{N}(S_i(o_j^i), 0.01)$. The $\|m\|$ $U_k^{S_i}$ samples are then added to the reference dataset D_{ref} , and labeled as *unknown*, to create a new reference dataset denoted $^{unk}D_{ref}$. To predict the proportion of unknown sources, a Bray-Curtis (Bray & Curtis, 1957) pairwise dissimilarity matrix of all S_i and $U_k^{S_i}$ samples is computed using scikit-bio (Rideout et al., 2018). This distance matrix is then embedded in two dimensions (default) with the scikit-bio implementation of PCoA. This sample embedding is divided into three subsets: $^{unk}D_{train}$ (64%), $^{unk}D_{test}$ (20%), and $^{unk}D_{validation}$ (16%). The scikit-learn (Pedregosa et al., 2011) implementation of KNN algorithm is then trained on $^{unk}D_{train}$, and the training accuracy is computed with $^{unk}D_{test}$. This trained KNN model is then corrected for probability estimation of the unknown proportion using the scikit-learn implementation of Platt's scaling method (Platt & others, 1999) with $^{unk}D_{validation}$. The proportion of unknown sources in S_i , $p_u \in [0, 1]$ is then estimated using this trained and corrected KNN model. Ultimately, this process is repeated independently for each sink sample S_i of D_{sink} .

Prediction of the proportion of known sources

First, only organism TAXIDs corresponding to the species taxonomic level are retained using the ETE toolkit (Huerta-Cepas, Serra, & Bork, 2016). A weighted Unifrac (default) (Lozupone, Hamady, Kelley, & Knight, 2007) pairwise distance matrix is then computed on the merged and normalized training dataset D_{ref} and test dataset D_{sink} with scikit-bio, using the NCBI taxonomy as a reference tree. This distance matrix is then embedded in two dimensions (default) using the scikit-learn implementation of t-SNE (Maaten & Hinton, 2008). The 2-dimensional embedding is then split back to training $^{tsne}D_{ref}$ and testing dataset $^{tsne}D_{sink}$. The KNN algorithm is then trained on the train subset, with a five (default) cross validation to look for the optimum number of K-neighbors. The training dataset $^{tsne}D_{ref}$ is further divided into three subsets: $^{tsne}D_{train}$ (64%), $^{tsne}D_{test}$ (20%), and $^{tsne}D_{validation}$ (16%). The training accuracy is then computed with $^{tsne}D_{test}$. Finally, this second trained KNN model is also corrected for source proportion estimation using the scikit-learn implementation of the Platt's method with $^{tsne}D_{validation}$. The proportion $p_{c_s} \in [0, 1]$ of each of the n_s sources $c_s \in \{c_1, \dots, c_{n_s}\}$ in each sample S_i is then estimated using this second trained and corrected KNN model.

Combining unknown and source proportions

For each sample S_i of the test dataset D_{sink} , the predicted unknown proportion p_u is then combined with the predicted proportion p_{c_s} for each of the n_s sources c_s of the training dataset such that $\sum_{c_s=1}^{n_s} s_c + p_u = 1$ where $s_c = p_{c_s} \cdot p_u$.

Finally, a summary table gathering the estimated sources proportions is returned as a csv file, as well as the t-SNE embedding sample coordinates.

Acknowledgements

Thanks to Dr. Christina Warinner, Dr. Alexander Herbig, Dr. AB Rohrlach, and Alexander Hübner for their valuable comments and for proofreading this manuscript. This work was funded by the Max Planck Society and the Deutsche Forschungsgemeinschaft, project code: EXC 2051 #390713860.

References

- Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological monographs*, 27(4), 325–349. doi:[10.2307/1942268](https://doi.org/10.2307/1942268)
- Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X., & Chen, J. (2018). GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ*, 6, e4600. doi:[10.7717/peerj.4600](https://doi.org/10.7717/peerj.4600)
- Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, 33(6), 1635–1638. doi:[10.1093/molbev/msw046](https://doi.org/10.1093/molbev/msw046)
- Hugenholtz, P., & Tyson, G. W. (2008). Microbiology: Metagenomics. *Nature*, 455(7212), 481. doi:[10.1038/455481a](https://doi.org/10.1038/455481a)
- Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., Bushman, F. D., et al. (2011). Bayesian community-wide culture-independent microbial source tracking. *Nature methods*, 8(9), 761. doi:[10.1038/nmeth.1650](https://doi.org/10.1038/nmeth.1650)
- Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, 73(5), 1576–1585. doi:[10.1128/AEM.01996-06](https://doi.org/10.1128/AEM.01996-06)
- Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov), 2579–2605.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825–2830.
- Platt, J., & others. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3), 61–74.
- Rideout, J. R., Caporaso, G., Bolyen, E., McDonald, D., Baeza, Y. V., Alastuey, J. C., Pitman, A., et al. (2018, December). biocore/scikit-bio: scikit-bio 0.5.5: More compositional methods added. doi:[10.5281/zenodo.2254379](https://doi.org/10.5281/zenodo.2254379)
- Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3), R46. doi:[10.1186/gb-2014-15-3-r46](https://doi.org/10.1186/gb-2014-15-3-r46)

Manuscript B: CoproID predicts the source of coprolites and paleofeces using microbiome composition and host DNA content

Maxime Borry, Bryan Cordova, Angela Perri, Marsha Wibowo, Tanvi Prasad Honap, Jada Ko, Jie Yu, Kate Britton, Linus Girdland-Flink, Robert C. Power, Ingelise Stuijts, Domingo C. Salazar-García, Courtney Hofman, Richard Hagan, Thérèse Samdapawindé Kagoné, Nicolas Meda, Helene Carabin, David Jacobson, Karl Reinhard, Cecil Lewis, Aleksandar Kostic, Choongwon Jeong, Alexander Herbig, Alexander Hübner, Christina Warinner

Published in PeerJ, 2020 April 17; DOI: 10.7717/peerj.9001

In manuscript B, I introduce a new method combining the Sourcepredict approach, and endogenous DNA content to predict the original source host of paleofeces samples.

When the first ancient human gut metagenomics studies were published, there was a keen interest from the broader scientific community to use these results and compare them with our current gut microbiomes. However, distinguishing human paleofeces, from paleofeces of other species had proven to be challenging from a morphological and microscopical perspective (Tito et al., 2012). Furthermore, the tight relationship between humans, and their canine companions made it especially complicated to distinguish human from canine coprolites (Poinar et al., 2009a). To address this issue, we used both the amount of host endogenous DNA, and the microbiome community based host prediction by Sourcepredict, which we both computed from shotgun metagenomics sequencing data. We combined all these steps in a self contained scalable and reproducible data analysis pipeline written with nf-core. (Ewels et al., 2020) and Nextflow (Tommaso et al., 2017). Our findings confirmed the general ambiguity between human and dog paleofeces in the archeological records, and allowed us to lift it for some of them.

Authors contributions

- **The candidate is:** first author.
- **Status:** published

Authors' contributions (in %, from 10%) to the given categories of the publication

Author	Conceptual	Data analysis	Experimental	Writing the manuscript	Provision of material
Maxime Borry	30	80	-	40	-
Bryan Cordova	-	-	50	-	-
Angela Perri	-	-	-	-	50
Richard Hagan	-	-	40	-	-
Choongwon Jeong	-	15	-	-	-
Alexander Herbig	15	-	-	10	-
Alexander Hübner	15	-	-	10	-
Christina Warinner	30	-	-	30	-
Others	10	5	10	10	50
Total:	100%	100%	100%	100%	100%

CoproID predicts the source of coprolites and paleofeces using microbiome composition and host DNA content

Maxime Borry¹, Bryan Cordova¹, Angela Perri^{2,3}, Marsha Wibowo^{4,5,6}, Tanvi Prasad Honap^{7,8}, Jada Ko⁹, Jie Yu¹⁰, Kate Britton^{3,11}, Linus Girdland-Flink^{11,12}, Robert C. Power^{3,13}, Ingelise Stuijts¹⁴, Domingo C. Salazar-García^{15,16}, Courtney Hofman^{7,8}, Richard Hagan¹, Thérèse Samdapawindé Kagoné¹⁷, Nicolas Meda¹⁷, Helene Carabin¹⁸, David Jacobson^{7,8}, Karl Reinhard¹⁹, Cecil Lewis^{7,8}, Aleksandar Kostic^{4,5,6}, Choongwon Jeong^{1,20}, Alexander Herbig¹, Alexander Hübner¹ and Christina Warinner^{1,9,21}

¹ Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany

² Department of Archaeology, Durham University, Durham, UK

³ Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

⁴ Section on Pathophysiology and Molecular Pharmacology, Joslin Diabetes Center, Boston, MA, USA

⁵ Section on Islet Cell and Regenerative Biology, Joslin Diabetes Center, Boston, MA, USA

⁶ Department of Microbiology, Harvard Medical School, Boston, MA, USA

⁷ Department of Anthropology, University of Oklahoma, Norman, OK, USA

⁸ Laboratories of Molecular Anthropology and Microbiome Research (LMAMR), University of Oklahoma, Norman, OK, USA

⁹ Department of Anthropology, Harvard University, Cambridge, MA, USA

¹⁰ Department of History, Wuhan University, Wuhan, China

¹¹ Department of Archaeology, University of Aberdeen, Aberdeen, Scotland, UK

¹² School of Natural Sciences and Psychology, Liverpool John Moores University, Liverpool, UK

¹³ Institut für Vor- und Frühgeschichtliche Archäologie und Provinzialrömische Archäologie, Ludwig-Maximilians-Universität München, München, Germany

¹⁴ The Discovery Programme, Dublin, Ireland

¹⁵ Grupo de Investigación en Prehistoria IT-1223-19 (UPV-EHU), IKERBASQUE-Basque Foundation for Science, Vitoria-Gasteiz, Spain

¹⁶ Departament de Prehistòria, Arqueologia i Història Antiga, Universitat de València, València, Spain

¹⁷ Centre Muraz, Bobo-Dioulasso, Burkina Faso

¹⁸ Département de pathologie et de microbiologie, Faculté de Médecine vétérinaire, Université de Montréal, Saint-Hyacinthe, QC, Canada

¹⁹ School of Natural Resources, University of Nebraska, Lincoln, NE, USA

²⁰ School of Biological Sciences, Seoul National University, Seoul, South Korea

²¹ Faculty of Biological Sciences, Friedrich-Schiller Universität Jena, Jena, Germany

Submitted 15 January 2020

Accepted 26 March 2020

Published 17 April 2020

Corresponding authors

Maxime Borry, borry@shh.mpg.de

Christina Warinner,

warinner@fas.harvard.edu

Academic editor

Antonio Amorim

Additional Information and
Declarations can be found on
page 17

DOI 10.7717/peerj.9001

© Copyright

2020 Borry et al.

Distributed under

Creative Commons CC-BY 4.0

OPEN ACCESS

ABSTRACT

Shotgun metagenomics applied to archaeological feces (paleofeces) can bring new insights into the composition and functions of human and animal gut microbiota from the past. However, paleofeces often undergo physical distortions in archaeological sediments, making their source species difficult to identify on the basis of fecal morphology or microscopic features alone. Here we present a reproducible and scalable pipeline using both host and microbial DNA to infer the host source of fecal material. We apply this pipeline to newly sequenced archaeological

specimens and show that we are able to distinguish morphologically similar human and canine paleofeces, as well as non-fecal sediments, from a range of archaeological contexts.

Subjects Anthropology, Bioinformatics, Genomics, Microbiology, Data Mining and Machine Learning

Keywords Coprolite, Paleofeces, Microbiome, Endogenous DNA, Archeology, Machine learning, Nextflow, Gut, Human, Dog

INTRODUCTION

The gut microbiome, located in the distal colon and primarily studied through the analysis of feces, is the largest and arguably most influential microbial community within the body (*The Human Microbiome Project Consortium, 2012*). Recent investigations of the human microbiome have revealed that it plays diverse roles in health and disease, and gut microbiome composition has been linked to a variety of human health states, including inflammatory bowel diseases, diabetes, and obesity (*Kho & Lal, 2018*). To investigate the gut microbiome, metagenomic sequencing is typically used to reveal both the taxonomic composition (i.e., which bacteria are there) and the functions the microbes are capable of performing (i.e., their potential metabolic activities) (*Sharpton, 2014*). Given the importance of the gut microbiome in human health, there is great interest in understanding its recent evolutionary and ecological history (*Warinner & Lewis, 2015; Davenport et al., 2017*).

Paleofeces, either in an organic or partially mineralized (coprolite) state, present a unique opportunity to directly investigate changes in the structure and function of the gut microbiome through time (*Warinner et al., 2015*). Paleofeces are found in a wide variety of archaeological contexts around the world and are generally associated with localized processes of desiccation, freezing, or mineralization. Paleofeces can range in size from whole, intact fecal pieces (*Jiménez et al., 2012*) to millimeter-sized sediment inclusions identifiable by their high phosphate and fecal sterol content (*Sistiaga et al., 2014*). Although genetic approaches have long been used to investigate dietary DNA found within human (*Gilbert et al., 2008; Poinar et al., 2001*) and animal (*Poinar et al., 1998; Hofreiter et al., 2000; Bon et al., 2012; Wood et al., 2016*) paleofeces, it is only recently that improvements in metagenomic sequencing and bioinformatics have enabled detailed characterization of their microbial communities (*Tito et al., 2008, 2012; Warinner et al., 2017*).

However, before evolutionary studies of the gut microbiome can be conducted, it is first necessary to confirm the host source of the paleofeces under study. Feces can be difficult to taxonomically assign by morphology alone (*Supplemental Text; Reinhard & Bryant, 1992*), and human and canine feces can be particularly difficult to distinguish in archaeological contexts (*Poinar et al., 2009*). Since their initial domestication more than 12,000 years ago (*Frantz et al., 2016*), dogs have often lived in close association with humans, and it is not uncommon for human and dog feces to co-occur at archaeological sites. Moreover, dogs often consume diets similar to humans because of provisioning or

refuse scavenging (Guiry, 2012), making their feces difficult to distinguish based on dietary contents. Even well-preserved fecal material degrades over time, changing in size, shape, and color (Fig. 1; Reinhard & Bryant, 1992). The combined analysis of host and microbial ancient DNA (aDNA) within paleofeces presents a potential solution to this problem.

Previously, paleofeces host source has been genetically inferred on the basis of PCR-amplified mitochondrial DNA sequences alone (Hofreiter et al., 2000); however, this is problematic in the case of dogs, which, in addition to being pets and working animals, were also eaten by many ancient cultures (Clutton-Brock & Hammond, 1994; Rosenswig, 2007; Kirch & O'Day, 2003; Podberscek, 2009), and thus trace amounts of dog DNA may be expected to be present in the feces of humans consuming dogs. Additionally, dogs often scavenge on human refuse, including human excrement (Butler & Du Toit, 2002), and thus ancient dog feces could also contain trace amounts of human DNA, which could be further inflated by PCR-based methods.

A metagenomics approach overcomes these issues by allowing a quantitative assessment of eukaryotic DNA at a genome-wide scale, including the identification and removal of modern human contaminant DNA that could potentially arise during excavation or subsequent curation or storage. It also allows for the microbial composition of the feces to be taken into account. Gut microbiome composition differs among mammal species (Ley et al., 2008), and thus paleofeces microbial composition could be used to confirm and authenticate host assignment. Available microbial tools, such as SourceTracker (Knights et al., 2011) and FEAST (Shenhav et al., 2019), can be used to perform the source prediction of microbiome samples from uncertain sources (sinks) using a reference dataset of source-labeled microbiome samples and, respectively, Gibbs sampling or an Expectation-Maximization algorithm. However, although SourceTracker has been widely used for modern microbiome studies and has even been applied to ancient gut microbiome data (Tito et al., 2012; Hagan et al., 2020), it was not designed to be a host species identification tool for ancient microbiomes.

In this work we present a bioinformatics method to infer and authenticate the host source of paleofeces from shotgun metagenomic DNA sequencing data: coproID (coprolite IDentification). coproID combines the analysis of putative host ancient DNA with a machine learning prediction of the feces source based on microbiome taxonomic composition. Ultimately, coproID predicts the host source of a paleofeces specimen from the shotgun metagenomic data derived from it. We apply coproID to previously published modern fecal datasets and show that it can be used to reliably predict their host. We then apply coproID to a set of newly sequenced paleofeces specimens and non-fecal archaeological sediments and show that it can discriminate between feces of human and canine origin, as well as between fecal and non-fecal samples.

MATERIALS AND METHODS

Gut microbiome reference datasets

Previously published modern reference microbiomes were chosen to represent the diversity of potential paleofeces sources and their possible contaminants, namely human fecal microbiomes from Non-Westernized Human/Rural (NWHHR) and Westernized

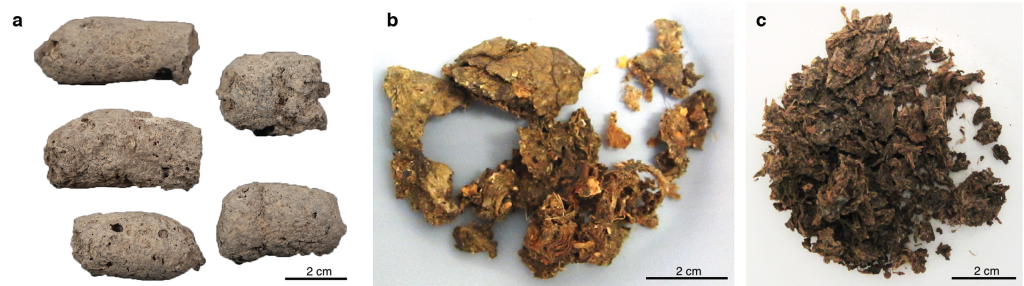


Figure 1 Examples of archaeological paleofeces analyzed in this study. (A) H29-3, from Anhui Province, China, Neolithic period; (B) Zape 2, from Durango, Mexico, ca. 1300 BP; (C) Zape 28, from Durango, Mexico, ca. 1300 BP. Paleofeces ranged from slightly mineralized intact pieces (A) to more fragmentary organic states (B and C), and color ranged from pale gray (A) to dark brown (C).

Full-size  DOI: 10.7717/peerj.9001/fig-1

Table 1 Modern reference microbiome datasets.

Metagenome source	Food production	N	Analysis	Source
<i>Homo sapiens</i> , USA	WHU	36	microbiome	<i>The Human Microbiome Project Consortium (2012)</i>
<i>Homo sapiens</i> , India (Bhopal and Kerala)	WHU and NWHR	19	microbiome	<i>Dhakan et al. (2019)</i>
<i>Homo sapiens</i> , Fiji (agrarian villages)	NWHR	20	microbiome	<i>Brito et al. (2019)</i>
<i>Homo sapiens</i> , Madagascar	NWHR	110	microbiome	<i>Pasolli et al. (2019)</i>
<i>Homo sapiens</i> , Brazil (Yanomami)	NWHR	3	microbiome	<i>Pasolli et al. (2019)</i>
<i>Homo sapiens</i> , Peru (Tunapuco)	NWHR	12	microbiome	<i>Obregon-Tito et al. (2015)</i>
<i>Homo sapiens</i> , Tanzania (Hadza)	NWHR	38	microbiome	<i>Rampelli et al. (2015)</i>
<i>Homo sapiens</i> , Peru (Matses)	NWHR	24	microbiome	<i>Obregon-Tito et al. (2015)</i>
<i>Homo sapiens</i> , USA (Boston)	WHU	49	host DNA	This study
<i>Homo sapiens</i> , Burkina Faso	NWHR	69	host DNA	This study
<i>Canis familiaris</i>	–	150	microbiome and host DNA	<i>Coelho et al. (2018)</i>
Soil	–	16	microbiome	<i>Fierer et al. (2012)</i>
Soil	–	2	microbiome	<i>CSIR-Central Institute of Medicinal & Aromatic Plants (2016)</i>
Soil	–	2	microbiome	<i>Orellana et al. (2018)</i>

Human/Urban (WHU) communities, dog fecal microbiomes, and soil samples (Table 1). Because the human datasets had been filtered to remove human genetic sequences prior to database deposition, we additionally generated new sequencing data from 118 fecal specimens from both NWHR and WHU populations (Table S5) in order to determine the average proportion and variance of host DNA in human feces. The Joslin Diabetes Center granted Ethical approval (CHS# 2017-25) to sample the WHU individuals. The Centre MURAZ Research Institute granted Ethical approval (No. 31/2016/CE-CM) to sample the NWHR individuals.

Table 2 Archaeological samples.

Archeological ID	Laboratory ID	Site Name	Region	Period	Sample type	Archaeologically suspected species	Plot ID
Zape 2*	ZSM002	Cueva de los Muertos Chiquitos	Mexico	1300 BP	Paleofeces	HUMAN	01
Zape 5*	ZSM005	Cueva de los Muertos Chiquitos	Mexico	1300 BP	Paleofeces	HUMAN	02
Zape 23	ZSM023	Cueva de los Muertos Chiquitos	Mexico	1300 BP	Paleofeces	HUMAN or CANID	03
Zape 25	ZSM025	Cueva de los Muertos Chiquitos	Mexico	1300 BP	Paleofeces	HUMAN	04
Zape 27	ZSM027	Cueva de los Muertos Chiquitos	Mexico	1300 BP	Paleofeces	HUMAN	05
Zape 28*	ZSM028	Cueva de los Muertos Chiquitos	Mexico	1300 BP	Paleofeces	HUMAN	06
Zape 29	ZSM029	Cueva de los Muertos Chiquitos	Mexico	1300 BP	Paleofeces	HUMAN	07
Zape 31	ZSM031	Cueva de los Muertos Chiquitos	Mexico	1300 BP	Paleofeces	HUMAN	08
H29-1	AHP001	Xiaosungang	China	Neolithic 7200–6800 BP	Paleofeces	CANID or CERVID	09
H35-1	AHP002	Xiaosungang	China	Neolithic 7200–6800 BP	Paleofeces	CANID or CERVID	10
H29-2	AHP003	Xiaosungang	China	Neolithic 7200–6800 BP	Paleofeces	CANID or CERVID	11
H29-3	AHP004	Xiaosungang	China	Neolithic 7200–6800 BP	Paleofeces	CANID or CERVID	12
LG 4560.69	YRK001	Surrey	UK	Post-Medieval	Paleofeces	HUMAN	13
AP3-C197S163	DRL001.A	Derragh	Ireland	Mesolithic	Midden Sediment	–	14
AP4-A6-2860	CBA001.A	CabeĂşo das Amoreiras	Portugal	Mesolithic	Midden Sediment	–	15
AP5-798-162	BRF001.A	Binchester Roman Fort	England	Roman	Midden Sediment	–	16
AP6-LPZ702	LEI010.A	Leipzig	Germany	10th–11th century AD	Midden Sediment	–	17
AP7-6-28353	ECO004.D	El Collado	Spain	Mesolithic	Pelvic Sediment	–	18
AP8-CMN-M1	CMN001.D	Cingle del Mas Nou	Spain	Mesolithic	Pelvic Sediment	–	19
AP9-17590	MLP001.A	Molpir	Slovakia	7th century BC	Pelvic Sediment	–	20

Note:

* Metagenomic data were previously published in [Hagan et al. \(2020\)](#).

Archaeological samples

A total of 20 archaeological samples, originating from 10 sites ([Fig. S3](#)) and spanning periods from 7200 BP to the medieval era, were selected for this study. Among these 20 samples, of which 17 are newly sequenced, 13 are paleofeces, 4 are midden sediments, and 3 are sediments obtained from human pelvic bone surfaces ([Table 2](#)).

Sampling

Paleofeces specimens from Mexico were sampled in a dedicated aDNA cleanroom in the Laboratories for Molecular Anthropology and Microbiome Research (LMAMR) at the University of Oklahoma, USA. Specimens from China were sampled in a dedicated aDNA cleanroom at the Max Planck Institute for the Science of Human History (MPI-SHH) in Jena, Germany. All other specimens were first sampled at the Max Planck Institute for Evolutionary Anthropology (MPI-EVA) in Leipzig, Germany before being transferred to the MPI-SHH for further processing. Sampling was performed using a sterile stainless steel spatula or scalpel, followed by homogenization in a mortar and pestle, if necessary. Because the specimens from Xiaosungang, China were very hard and dense, a rotary drill was used to section the coprolite prior to sampling. Where possible, fecal material was sampled from the interior of the specimen rather than the surface. Specimens from Molphir and Leipzig were received suspended in a buffer of trisodium phosphate, glycerol, and formyl following screening for parasite eggs using optical microscopy. For each paleofeces specimen, a total of 50–200 mg was analyzed.

Modern feces were obtained under written informed consent from Boston, USA (WHU) from a long-term (>50 years) type 1 diabetes cohort, and from villages in Burkina Faso (NWHR) as part of broader studies on human gut microbiome biodiversity and health-associated microbial communities. Feces were collected fresh and stored frozen until analysis. A total of 250 mg was analyzed for each fecal specimen.

DNA extraction

For paleofeces and sediment samples, DNA extractions were performed using a silica spin column protocol ([Dabney et al., 2013](#)) with minor modifications in dedicated aDNA cleanrooms located at LMAMR (Mexican paleofeces) and the MPI-SHH (all other paleofeces). At LMAMR, the modifications followed those of method D described in [Hagan et al. \(2020\)](#). DNA extractions at the MPI-SHH were similar, but omitted the initial bead-beating step, and a single silica column was used per sample instead of two. Additionally, to reduce centrifugation errors, DNA extractions performed at the MPI-SHH substituted the column apparatus from the High Pure Viral Nucleic Acid Large Volume Kit (Roche, Switzerland) in place of the custom assembled Zymo-reservoirs coupled to MinElute (Qiagen, Hilden, Germany) columns described in [Dabney et al. \(2013\)](#). Samples processed at the MPI-SHH were also partially treated with uracil-DNA-glycosylase (UDG) enzyme to confine DNA damage to the ends of the DNA molecules ([Rohland et al., 2015](#)).

For modern feces, DNA was extracted from Burkina Faso fecal samples using the AllPrep PowerViral DNA/RNA Qiagen kit at Centre MURAZ Research Institute in Burkina Faso. DNA was extracted from the Boston fecal material using the ZymoBIOMICS DNA Miniprep Kit (D4303) at the Joslin Diabetes Center.

Library preparation and sequencing

For paleofeces and sediment samples, double-stranded, dual-indexed shotgun Illumina libraries were constructed following ([Meyer & Kircher, 2010](#)) using either the NEBNext DNA Library Prep Master Set (E6070) kit ([Hagan et al., 2020](#); [Mann et al., 2018](#)) for the

Mexican paleofeces or individually purchased reagents ([Mann et al., 2018](#)) for all other samples. Following library amplification using Phusion HotStart II (ZSM023, ZSM025, ZSM027, ZSM029), KAPA HiFi Uracil+ (ZSM002, ZSM005, ZSM028), or Agilent Pfu Turbo Cx Hotstart (all other paleofeces) polymerase, the libraries were purified using a Qiagen MinElute PCR Purification kit and quantified using either a BioAnalyzer 2100 with High Sensitivity DNA reagents or an Agilent Tape Station D1000 Screen Tape kit. The Mexican libraries were pooled in equimolar amounts and sequenced on an Illumina HiSeq 2000 using 2×100 bp paired-end sequencing. All other libraries were pooled in equimolar amounts and sequenced on an Illumina HiSeq 4000 using 2×75 bp paired-end sequencing.

For modern NWHR feces, double-stranded, dual-indexed shotgun Illumina libraries were constructed in a dedicated modern DNA facility at LMAMR. Briefly, after DNA quantification using a Qubit dsDNA Broad Range Assay Kit, DNA was sheared using a QSonica Q800R in 1.5 mL 4 °C cold water at 50% amplitude for 12 min to aim for a fragment size between 400 and 600 bp. Fragments shorter than 150 bp were removed using Sera-Mag SpeedBeads and a Alpaqua 96S Super Magnet Plate. End-repair and A-tailing was performed using the Kapa HyperPrep EndRepair and A-Tailing Kit, and Illumina sequencing adapters were added. After library quantification, libraries were dual-indexed in an indexing PCR over four replicates, pooled, and purified using the SpeedBeads. Libraries were quantified using the Agilent Fragment Analyzer, pooled in equimolar ratios, and size-selected using the Pippin Prep to a target size range of 400–600 bp. Libraries were sequenced on an Illumina NovaSeq S1 using 2×150 bp paired-end sequencing at the Oklahoma Medical Research Foundation Next-Generation Sequencing Core facility. Modern WHU libraries were generated using the NEBNext DNA library preparation kit following manufacturer's recommendations, after fragmentation by shearing for a target fragment size of 350 bp. The libraries were then pooled and sequenced by Novogene on a NovaSeq S4 using 2×150 bp paired-end sequencing.

Proportion of host DNA in gut microbiome

Because it is standard practice to remove human DNA sequences from metagenomics DNA sequence files before data deposition into public repositories, we were unable to infer the proportion of human DNA in human feces from publicly available data. To overcome this problem, we measured the proportion of human DNA in two newly generated fecal metagenomics datasets from Burkina Faso (NWHR) and Boston, U.S.A. (WHU) ([Table S5](#)). To measure the proportion of human DNA in each fecal dataset, we used the Anonymap pipeline ([Borry, 2019a](#)) to perform a mapping with Bowtie 2 ([Langmead & Salzberg, 2012](#)) with the parameters `--very-sensitive -N 1` after adapter cleaning and reads trimming for ambiguous and low-quality bases with a QScore below 20 by AdapterRemoval v2 ([Schubert, Lindgreen & Orlando, 2016](#)). To preserve the anonymity of the donors, the sequences of mapped reads were then replaced by Ns thus anonymizing the alignment files. We obtained the proportion of host DNA per sample by dividing the number of mapped reads by the total number of reads in the sample. The proportion of

host DNA in dog feces was determined from the published dataset [Coelho et al. \(2018\)](#) as described above, but without the anonymization step.

Visualization and statistical analysis

The statistical analyses were performed in Python v3.7.6 using Scipy v1.4.1, and the figures were generated using Plotnine v0.6.0.

coproID pipeline

Data were processed using the coproID pipeline v1.0 ([Fig. 2](#)) ([DOI 10.5281/zenodo.2653757](#)) written using Nextflow ([Di Tommaso et al., 2017](#)) and made available through nf-core ([Ewels et al., 2019](#)). Nextflow is a Domain Specific Language designed to ensure reproducibility and scalability for scientific pipelines, and nf-core is a community-developed set of guidelines and tools to promote standardization and maximum usability of Nextflow pipelines. CoproID consists of 5 different steps:

Preprocessing

Fastq sequencing files are given as an input. After quality control analysis with FastQC ([Andrews, 2010](#)), raw sequencing reads are cleaned from sequencing adapters and trimmed from ambiguous and low-quality bases with a QScore below 20, while reads shorter than 30 base pairs are discarded using AdapterRemoval v2. By default, paired-end reads are merged on overlapping base pairs.

Mapping

The preprocessed reads are then aligned to each of the target species genomes (source species) by Bowtie2 with the `--very-sensitive` preset while allowing for a mismatch in the seed search (`-N 1`). When running coproID with the ancient DNA mode (`--adna`), alignments are filtered by PMDtools ([Skoglund et al., 2014](#)) to only retain reads showing post-mortem damages (PMD). PMDtools default settings are used, with specified library type, and only reads with a PMDScore greater than three are kept.

Computing host DNA content

Next, filtered alignments are processed in Python using the Pysam library ([Pysam Developers, 2018](#)). Reads matching above the identity threshold of 0.95 to multiple host genomes are flagged as common reads $reads_{commons}$ whereas reads mapping above the identity threshold to a single host genome are flagged as genome-specific host reads $reads_{spec\ g}$ to each genome g . Each source species host DNA is normalized by genome size and gut microbiome host DNA content such as:

$$NormalizedHostDNA(source\ species) = \frac{\sum length(reads_{spec\ g})}{genome_g\ length \cdot endo_g} \quad (1)$$

where for each species of genome g , $\sum length(reads_{spec\ g})$ is the total length of all $reads_{spec\ g}$, $genome_g\ length$ is the size of the genome, and $endo_g$ is the host DNA proportion in the species gut microbiome.

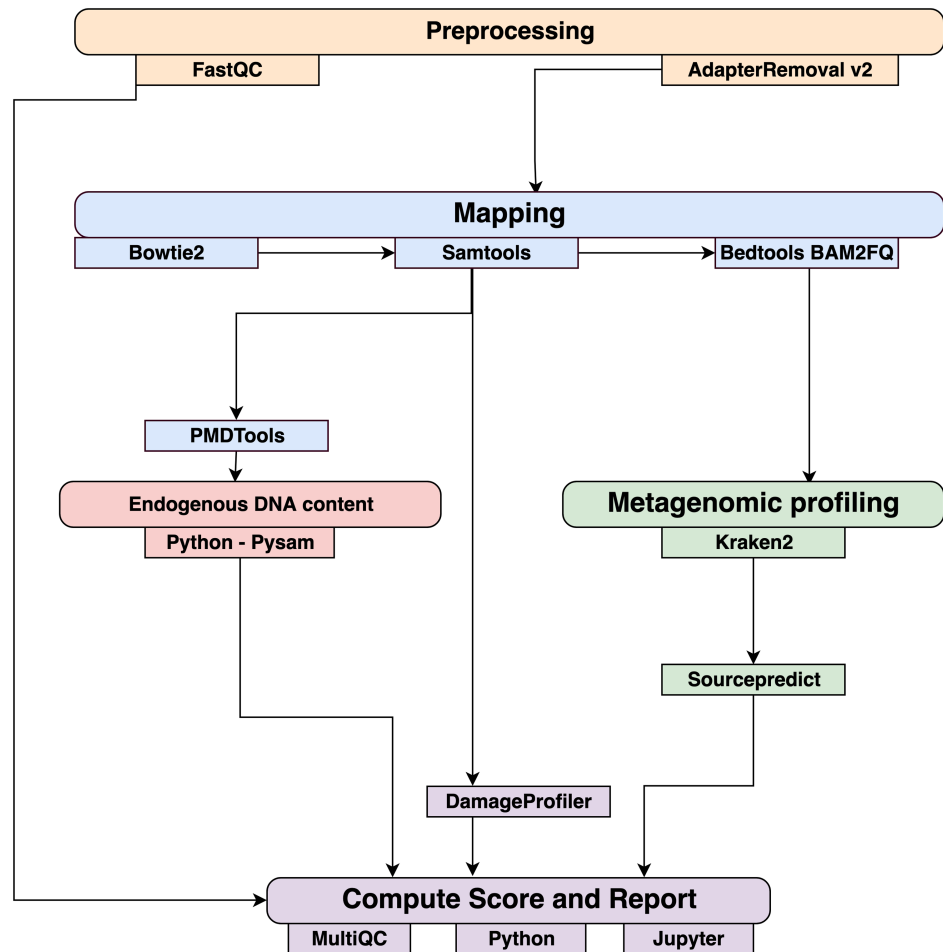


Figure 2 Workflow schematic of the coproID pipeline. CoproID consists of five steps: *Preprocessing* (orange), *Mapping* (blue), *Computing host DNA content for each metagenome* (red), *Metagenomic profiling* (green), and *Reporting* (violet). Individual programs (squared boxes) are colored by category (rounded boxes). [Full-size !\[\]\(bc797a94aafc1f7587775d98c87011d1_img.jpg\) DOI: 10.7717/peerj.9001/fig-2](https://doi.org/10.7717/peerj.9001/fig-2)

Afterwards, an host DNA ratio is computed for each source species such as:

$$\text{NormalizedRatio}(\text{source species}) = \frac{\text{NormalizedHostDNA}(\text{source species})}{\sum \text{NormalizedHost DNA}(\text{source species})} \quad (2)$$

where $\sum \text{NormalizedHost DNA}(\text{source species})$ is the sum of all source species Normalized Host DNA.

Metagenomic profiling

Adapter clipped and trimmed reads are given as an input to Kraken 2 (Wood & Salzberg, 2014). Using the MiniKraken2_v2_8GB database (2019/04/23 version), Kraken 2 performs the taxonomic classification to output a taxon count per sample report file. All samples' taxon counts are pooled together in a taxon counts matrix with samples in columns, and taxons in rows. Next, Sourcepredict (Borry, 2019b) is used to predict the source based on each microbiome sample taxon composition. Using dimension reduction and K-Nearest

Neighbors (KNN) machine learning trained with reference modern gut microbiomes samples (Table 1), Sourcepredict estimates a proportion $prop_{microbiome}(source\ species)$ of each potential source species, here Human or Dog, for each sample.

Reporting

For each filtered alignment file, the DNA damage patterns are estimated with DamageProfiler (Peltzer & Neukamm, 2019). The information from the host DNA content and the metagenomic profiling are gathered for each source in each sample such as:

$$proportion(source\ species) = \frac{NormalizedRatio(source\ species)}{prop_{microbiome}(source\ species)}$$

Finally, a summary report is generated including the damage plots, a summary table of the coproID metrics, and the embedding of the samples in two dimensions by Sourcepredict. coproID is available on GitHub at the following address: github.com/nf-core/coproID.

RESULTS

We analyzed 20 archaeological samples with coproID v1.0 to estimate their source using both host DNA and microbiome composition.

Host DNA in reference gut microbiomes

Before analyzing the archaeological samples, we first tested whether there is a per-species difference in host DNA content in modern reference human and dog feces. With Anonymap, we computed the amount of host DNA in each reference gut microbiome (Table S1). We found that the median percentages of host DNA in NWHR, WHU, and Dog (Fig. 3) are significantly different at $\alpha = 0.05$ (Kruskal–Wallis H -test = 117.40, p value < 0.0001). We confirmed that there is a significant difference of median percentages of host DNA between dogs and NWHR, as well as dogs and WHU, with Mann–Whitney U tests (Table 3) and therefore corrected each sample by the mean percentage of gut host DNA found in each species, 1.24% for humans ($\mu_{NWHR} = 0.85$, $\sigma_{NWHR} = 2.33$, $\mu_{WHU} = 1.67$, $\sigma_{WHU} = 0.81$), and 0.11% for dogs ($\sigma_{dog} = 0.16$) (Eq. (1); Table S1). This information was used to correct for the amount of host DNA found in paleofeces.

The effect of PMD filtering on host species prediction

Because aDNA accumulates damage over time (Briggs et al., 2007), we could use this characteristic to filter for reads carrying these specific damage patterns using PMDtools, and therefore reduce modern contamination in the dataset. We applied PMD filtering to our archaeological datasets, and for each, compared the predicted host source before and afterwards. The predicted host sources did not change after the DNA damage read filtering, but some became less certain (Fig. 4). Most samples are confidently assigned to one of the two target species, however some samples previously categorized as humans now lie in the uncertainty zone. This suggests that PMDtools filtering lowered the modern

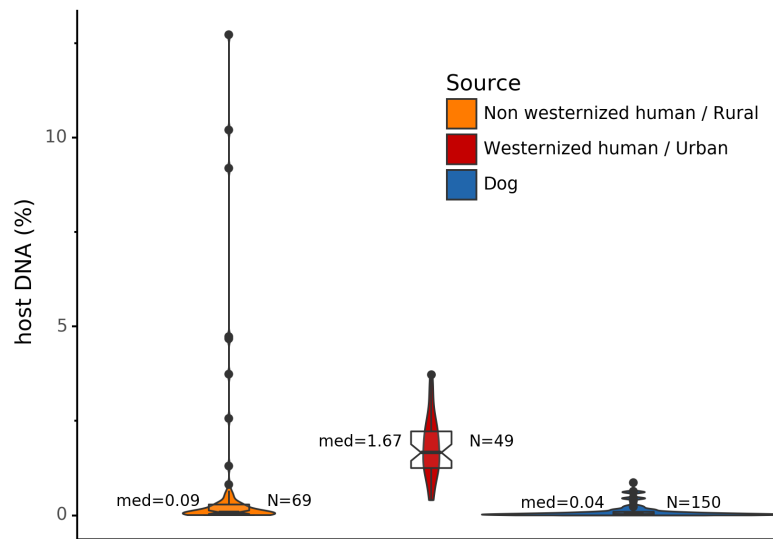


Figure 3 Gut microbiome host DNA content. The median percentage of host DNA in the gut microbiome and the number of samples in each group are displayed besides each boxplot.

Full-size [DOI: 10.7717/peerj.9001/fig-3](https://doi.org/10.7717/peerj.9001/fig-3)

Table 3 Statistical comparison of reference gut host DNA content. Mann-Whitney U test for independent observations. H_0 : the distributions of both populations are equal.

Comparison	Mann-Whitney U test	p Value
Dog vs NWHR	3327.0	<0.0001
Dog vs WHU	41.0	<0.0001
NWHR vs WHU	370.0	<0.0001
Dog vs Human	3368.0	<0.0001

human contamination which might have originated from sample excavation and manipulation.

The trade-off of PMDtools filtering is that it reduces the assignment power by lowering the number of reads available for host DNA-based source prediction by only keeping PMD-bearing reads. This loss is greater for well-preserved samples, which may have relatively few damaged reads (<15% of total). Ultimately, applying damage filtering can make it more difficult to categorize samples on the sole basis of host DNA content, but it also makes source assignments more reliable by removing modern contamination.

Source microbiome prediction of reference samples by Sourcepredict

To help resolve ambiguities related to the host aDNA present within a sample, we also investigated gut microbiome composition as an additional line of evidence to better predict paleofeces source. After performing taxonomic classification using Kraken2, we computed a sample pairwise distance matrix from the species counts. With the t -SNE dimension reduction method, we embedded this distance matrix in two dimensions to visualize the

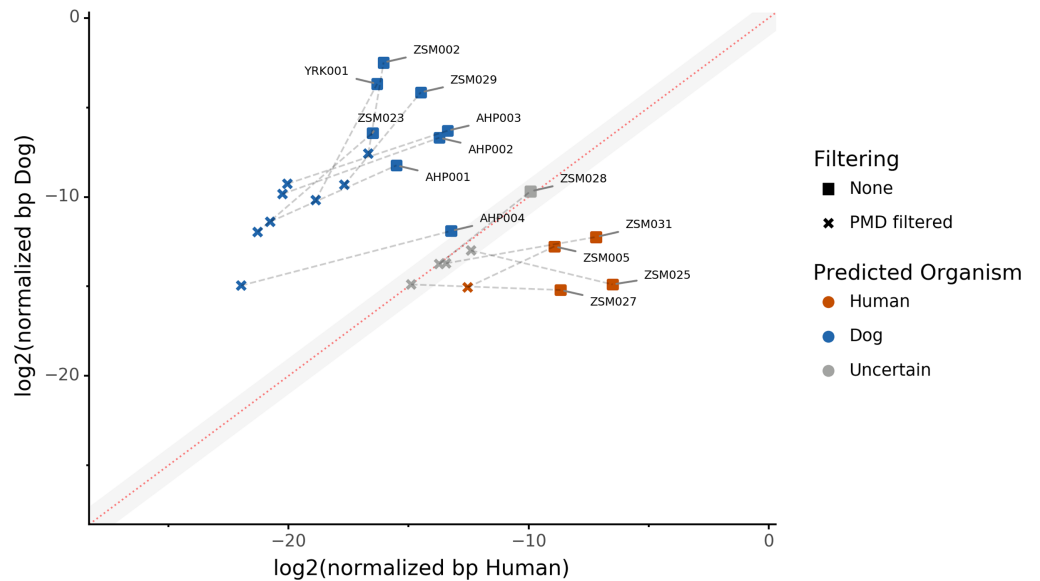


Figure 4 The effect of filtering for damaged reads using PMD. The \log_2 of the human *NormalizedHostDNA* is graphed against the \log_2 of the dog *NormalizedHostDNA*. Squares represent samples before filtering by PMD, whereas crosses represent samples after filtering by PMD. Dotted lines show the correspondence between samples. The red diagonal line marks the boundary between the two species, and the grey shaded area indicates a zone of species uncertainty ($\pm 1\log_2FC$) due to insufficient genetic information. [Full-size !\[\]\(429fa903b72fda6689f4e2eacafe6305_img.jpg\) DOI: 10.7717/peerj.9001/fig-4](https://doi.org/10.7717/peerj.9001/fig-4)

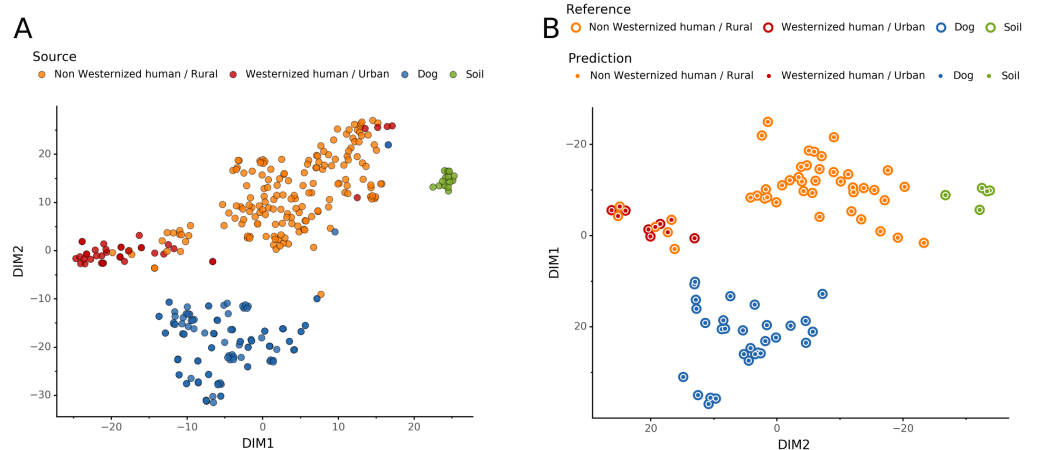


Figure 5 Embedding of reference modern gut microbiomes. (A) *t*-SNE embedding of the species composition based on sample pairwise Weighted Unifrac distances for training modern gut microbiomes training samples. Samples are colored by their actual source. (B) *t*-SNE embedding of the species composition based on sample pairwise Weighted Unifrac distances for source prediction of modern test samples. The outer circle color is the actual source of a sample, while the inner circle color is the predicted sample source by Sourcepredict. [Full-size !\[\]\(17ad878ff18720bfa5633be96f8af173_img.jpg\) DOI: 10.7717/peerj.9001/fig-5](https://doi.org/10.7717/peerj.9001/fig-5)

sample positions and sources (Fig. 5A). We then used a KNN machine learning classifier on this low dimension embedding to predict the source of gut microbiome samples. This trained KNN model reached a test accuracy of 0.94 on previously unseen data (Fig. 5B).

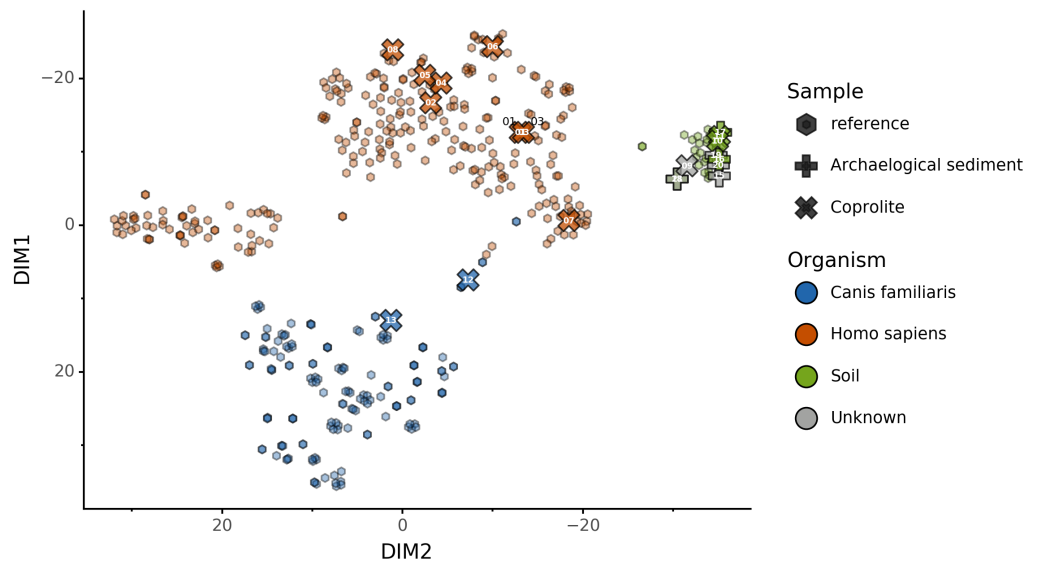


Figure 6 Prediction of archaeological samples sources and *t*-SNE embedding by Sourcepredict. *t*-SNE embedding of archaeological (crosses) and modern (hexagons) samples. The color of the modern samples is based on their actual source while the color of the archaeological samples is based on their predicted source by Sourcepredict. Archaeological sample are labelled with their *Plot ID* (Table 2).

Full-size [DOI: 10.7717/peerj.9001/fig-6](https://doi.org/10.7717/peerj.9001/fig-6)

Embedding of archaeological samples by Sourcepredict

We used this trained KNN model to predict the sources of the 20 paleofeces and archaeological sediment samples, after embedding them in a two-dimensional space (Fig. 6). Based on their microbiome composition data, Sourcepredict predicted 2 paleofeces samples as dogs, 8 paleofeces samples as human, 2 paleofeces samples and 4 archaeological sediments as soil, while the rest were predicted as unknown (Table S2).

coproID prediction

Combining both PMD-filtered host DNA information and microbiome composition, coproID was able to reliably categorize 7 of the 13 paleofeces samples, as 5 human paleofeces and 2 canine paleofeces, whereas all of the non-fecal archaeological sediments were flagged as unknown (Fig. 7). This confirms the original archaeological source hypothesis for five samples (ZSM005, ZSM025, ZSM027, ZSM028, ZSM031) and specifies or rejects the original archaeological source hypothesis for the two others (YRK001, AHP004). The 6 paleofeces samples not reliably identified by coproID have a conflicting source proportion estimation between host DNA and microbiome composition (Fig. 8; Table S3). Specifically, paleofeces AHP001, AHP002 and AHP003 show little predicted gut microbiome preservation, and thus have likely been altered by taphonomic (decomposition) processes. Paleofeces ZSM002, ZSM023 and ZSM029, by contrast, show good evidence of both host and microbiome preservation, but have conflicting source predictions based on host and microbiome evidence. Given that subsistence is associated with gut microbiome composition, this conflict may be related to insufficient gut microbiome datasets available for non-Westernized dog populations (Hagan et al., 2020).

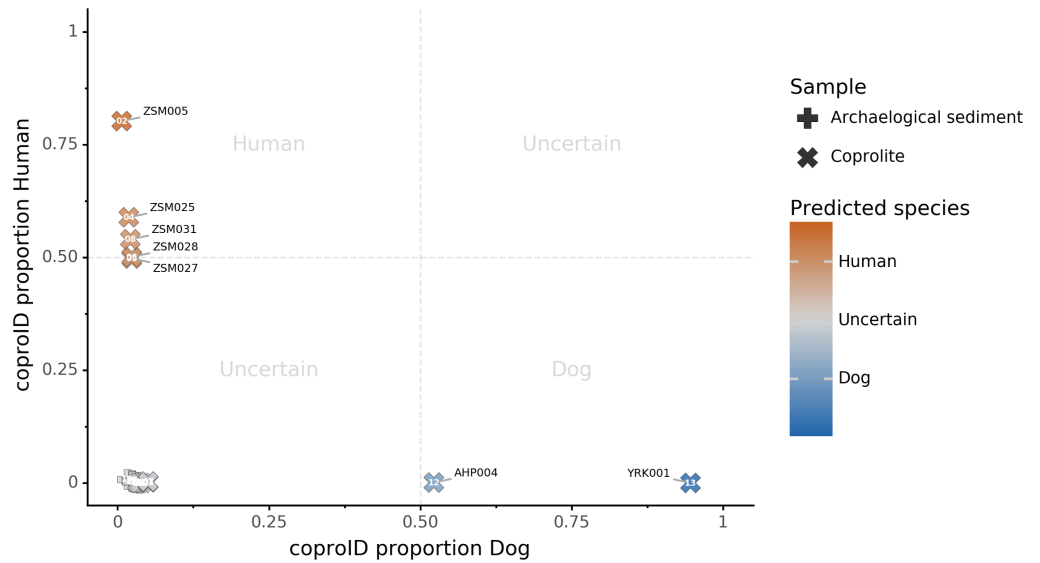


Figure 7 coproID source prediction. Predicted human proportion graphed versus predicted canine proportion. Samples are colored by their predicted sources proportions. Samples with a low canine and human proportion are not annotated. [Full-size !\[\]\(d05e99f54f2116973a3261aa569ffd8a_img.jpg\) DOI: 10.7717/peerj.9001/fig-7](https://doi.org/10.7717/peerj.9001/fig-7)

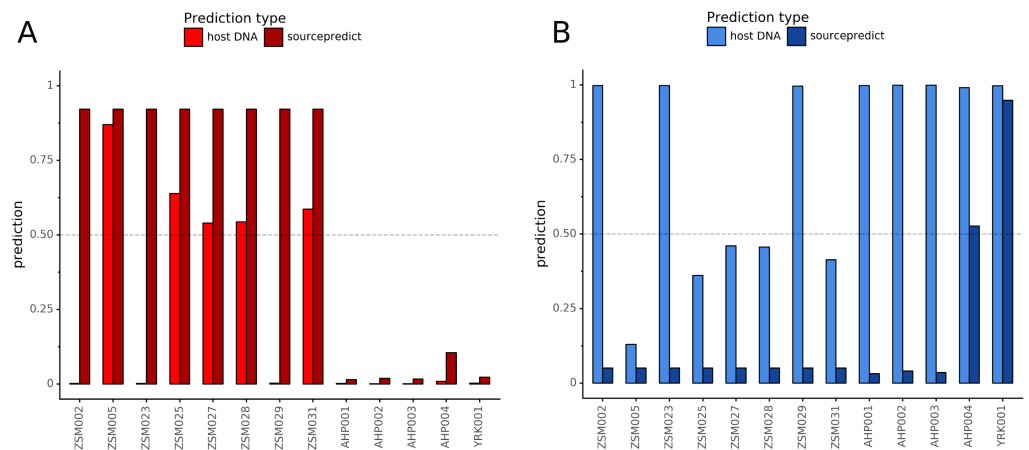


Figure 8 Host DNA and Sourcepredict source prediction for paleofeces samples. For human (A) and canine (B). The vertical bar represents the predicted proportion by host DNA (lighter fill) or by Sourcepredict (darker fill). The horizontal dashed line represents the confidence threshold to assign a source to a sample. [Full-size !\[\]\(ffa6a1c610eba82371ae6fa91d718d09_img.jpg\) DOI: 10.7717/peerj.9001/fig-8](https://doi.org/10.7717/peerj.9001/fig-8)

DISCUSSION

Paleofeces are the preserved remains of human or animal feces, and although they typically only preserve under highly particular conditions, they are nevertheless widely reported in the paleontological and archaeological records and include specimens ranging in age from the Paleozoic era (*Dentzien-Dias et al., 2013*) to the last few centuries. Paleofeces can provide unprecedented insights into animal health and diet, parasite biology and evolution, and the changing ecology and evolution of the gut microbiome. However,

because many paleofeces lack distinctive morphological features, determining the host origin of a paleofeces can be a difficult problem (Poinar *et al.*, 2009). In particular, distinguishing human and canine paleofeces can be challenging because they are often similar in size and shape, they tend to co-occur at archaeological sites and in midden deposits, and humans and domesticated dogs tend to eat similar diets (Guiry, 2012). We developed coproID to aid in identifying the source organism of archaeological paleofeces and coprolites by applying a combined approach relying on both ancient host DNA content and gut microbiome composition.

coproID addresses several shortcomings of previous methods. First, we have included a DNA damage-filtering step that allows for the removal of potentially contaminating modern human DNA, which may otherwise skew host species assignment. We have additionally measured and accounted for significant differences in the mean proportion of host DNA found in dog and human feces, and we also accounted for differences in host genome size between humans and dogs when making quantitative comparisons of host DNA. Then, because animal DNA recovered from paleofeces may contain a mixture of host and dietary DNA, we also utilize gut microbiome compositional data to estimate host source. We show that humans and dogs have distinct gut microbiome compositions, and that their feces can be accurately distinguished from each other and from non-feces using a machine learning classifier after data dimensionality reduction. Taken together, these approaches allow a robust determination of paleofeces and coprolite host source, that takes into account both modern contamination, microbiome composition, and postmortem degradation.

In applying coproID to a set of 20 archaeological samples of known and/or suspected origin, all 7 non-fecal sediment samples were accurately classified as “uncertain” and were grouped with soil by Sourcepredict. For the 13 paleofeces and coprolites under study, 7 exhibited matching host and microbiome source assignments and were confidently classified as either human ($n = 5$) or canine ($n = 2$). Importantly, one of the samples confidently identified as canine was YRK001, a paleofeces that had been recovered from an archaeological chamber pot in the United Kingdom, but which showed an unusual diversity of parasites inconsistent with human feces, and therefore posed issues in host assignment.

For the remaining six unidentified paleofeces, three exhibited poor microbiome preservation and were classified as “uncertain”, while the other three were well-preserved but yielded conflicting host DNA and microbiome assignments. These three samples, ZSM002, Z023 and ZSM029, all from prehistoric Mexico, all contain high levels of canine DNA, but have gut microbiome profiles within the range of NWHHR humans. Classified as “uncertain”, there are two possible explanations for these samples. First, these feces could have originated from a human who consumed a recent meal of canine meat. Dogs were consumed in ancient Mesoamerica (Clutton-Brock & Hammond, 1994; Santley & Rose, 1979; Rosenswig, 2007; Wing, 1978), but further research on the expected proportion of dietary DNA in human feces is needed to determine whether this is a plausible explanation for the very high amounts of canine DNA (and negligible amounts of human DNA) observed.

Alternatively, these feces could have originated from a canine whose microbiome composition is shifted relative to that of the reference metagenomes used in our training set. It is now well-established that subsistence mode strongly influences gut microbiome composition in humans (*Obregon-Tito et al., 2015*), with NWHR and WHU human populations largely exhibiting distinct gut microbiome structure (*Fig. 5A*). To date, no gut microbiome data is available from non-Westernized dogs, and all reference dog metagenome data included as training data for coproID originated from a single study of labrador retrievers and beagles (*Coelho et al., 2018*). Future studies of non-Westernized rural dogs are needed to establish the full range of gut microbial diversity in dogs and to more accurately model dog gut microbiome diversity in the past. Given that all confirmed human paleofeces in this study falls within the NWHR cluster (*Fig. 6*), we anticipate that our ability to accurately classify dog paleofeces and coprolites as canine (as opposed to “uncertain”) will improve with the future addition of non-Westernized rural dog metagenomic data.

In addition to archaeological applications, coproID may also have useful applications in the field of forensic genetic sciences, where it may assist with the identification of human or other feces. As with the investigation of paleofeces, coproID works best when sufficient comparative reference materials or datasets are available. Until a more exhaustive catalog of the human and dog gut microbiome composition is established, not all samples submitted to the coproID analysis will be able to be accurately classified. However, as microbiome reference datasets expand and methods become more standardized in the field, gut microbiome analyses will have increasing applications in the fields of archaeology and forensics (*Hampton-Marcell, Lopez & Gilbert, 2017*).

CONCLUSIONS

We developed an open-source, documented, tested, scalable, and reproducible method to perform the identification of archaeological paleofeces and coprolite source. By leveraging the information from host DNA and microbiome composition, we were able to identify and/or confirm the source of newly sequenced paleofeces. We demonstrated that coproID can provide useful assistance to archaeologists in identifying authentic paleofeces and inferring their host. Future work on dog gut microbiome diversity, especially among rural, non-Westernized dogs, may help improve the tool’s sensitivity even further.

ACKNOWLEDGEMENTS

We thank David Petts, Zdeněk Tvrďý, Susanne Stegmann-Rajtár, and Zuzana Rajtarova for contributing archaeological samples to this study. We thank the Guildford Museum (Guildford Borough Council Heritage Service) and Catriona Wilson for allowing us to analyze the chamber pot paleofeces sample from Surrey, UK. The sample from Derragh, Ireland was excavated by Discovery Programme, an all-Ireland public center of archaeological research supported by the Heritage Council, during field work in 2003–2005 as part of the Lake Settlement Project. Thanks to the Servei d’Investigació Prehistòrica of València and Museu de la Valltorta of Catelló for access to material.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the US National Institutes of Health R01GM089886 (to Christina Warinner and Cecil Lewis), the Deutsche Forschungsgemeinschaft EXC 2051 #390713860 (to Christina Warinner), and the Max Planck Society. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
US National Institutes of Health R01GM089886.
Deutsche Forschungsgemeinschaft: EXC 2051 #390713860.
Max Planck Society.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Maxime Borry conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Bryan Cordova performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Angela Perri performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Marsha Wibowo performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Tanvi Prasad Honap performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Jada Ko performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Jie Yu performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Kate Britton performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Linus Girdland-Flink performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Robert C. Power performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Ingelise Stuijts performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

- Domingo C. Salazar-García performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Courtney Hofman performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Richard Hagan performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Thérèse Samdapawindé Kagoné performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Nicolas Meda performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Helene Carabin performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- David Jacobson performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Karl Reinhard performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Cecil Lewis conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Aleksandar Kostic conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Choongwon Jeong analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Alexander Herbig conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Alexander Hübner conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Christina Warinner conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Human Ethics

The following information was supplied relating to ethical approvals (i.e., approving body and any reference numbers):

The Joslin Diabetes Center granted ethical approval to sample the WHU individuals (Study No. 2017-25). The Centre MURAZ Research Institute granted ethical approval to sample the NWHR individuals (31/2016/CE-CM).

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

Genetic data are available in the European Nucleotide Archive (ENA): [PRJEB33577](#) and [PRJEB35362](#).

Data Availability

The following information was supplied regarding data availability:

The code for the analysis is available at GitHub: <https://github.com/maxibor/coproid-article> and Zenodo:

Maxime Borry, Alexander Peltzer, & James A. Fellows Yates. (2019, April 29). nf-core/coproid: coproID v1.0 - Dioptra Walrus (Version 1.0). Zenodo.

DOI [10.5281/zenodo.2653757](https://doi.org/10.5281/zenodo.2653757).

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.9001#supplemental-information>.

REFERENCES

- Andrews S. 2010.** Fastqc: a quality control tool for high throughput sequence data. Available at <https://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
- Bon C, Berthouaud V, Maksud F, Labadie K, Poulain J, Artiguenave F, Wincker P, Aury J-M, Elalouf J-M. 2012.** Coprolites as a source of information on the genome and diet of the cave hyena. *Proceedings of the Royal Society B: Biological Sciences* **279(1739)**:2825–2830 DOI [10.1098/rspb.2012.0358](https://doi.org/10.1098/rspb.2012.0358).
- Borry M. 2019a.** maxibor/anonymap: Anonymap v1.0. Available at <https://doi.org/10.5281/zenodo.2669470>.
- Borry M. 2019b.** Sourcepredict: prediction of metagenomic sample sources using dimension reduction followed by machine learning classification. *Journal of Open Source Software* **4(41)**:1540.
- Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S. 2007.** Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of The United States of America* **104(37)**:14616–14621 DOI [10.1073/pnas.0704665104](https://doi.org/10.1073/pnas.0704665104).
- Brito IL, Gurry T, Zhao S, Huang K, Young SK, Shea TP, Naisilisili W, Jenkins AP, Jupiter SD, Gevers D, Alm EJ. 2019.** Transmission of human-associated microbiota along family and social networks. *Nature Microbiology* **4(6)**:964–971 DOI [10.1038/s41564-019-0409-6](https://doi.org/10.1038/s41564-019-0409-6).
- Butler J, Du Toit J. 2002.** Diet of free-ranging domestic dogs (*canis familiaris*) in rural Zimbabwe: implications for wild scavengers on the periphery of wildlife reserves. *Animal Conservation Forum* **5(1)**:29–37 DOI [10.1017/S136794300200104X](https://doi.org/10.1017/S136794300200104X).
- Clutton-Brock J, Hammond N. 1994.** Hot dogs: comestible canids in preclassic maya culture at cuello, belize. *Journal of Archaeological Science* **21(6)**:819–826 DOI [10.1006/jasc.1994.1079](https://doi.org/10.1006/jasc.1994.1079).
- Coelho LP, Kultima JR, Costea PI, Fournier C, Pan Y, Czarnecki-Maulden G, Hayward MR, Forslund SK, Schmidt TSB, Descombes P, Jackson JR, Li Q, Bork P. 2018.** Similarity of the dog and human gut microbiomes in gene content and response to diet. *Microbiome* **6(1)**:72 DOI [10.1186/s40168-018-0450-3](https://doi.org/10.1186/s40168-018-0450-3).
- CSIR-Central Institute of Medicinal and Aromatic Plants. 2016.** Chrysopogon zizanioides (ID 322597) - BioProject - NCBI. Available at <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA322597>.
- Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, Valdiosera C, García N, Pääbo S, Arsuaga J-L, Meyer M. 2013.** Complete mitochondrial genome sequence of a middle pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National*

- Academy of Sciences of the United States of America* **110(39)**:15758–15763
DOI [10.1073/pnas.1314445110](https://doi.org/10.1073/pnas.1314445110).
- Davenport ER, Sanders JG, Song SJ, Amato KR, Clark AG, Knight R. 2017. The human microbiome in evolution. *BMC Biology* **15(1)**:127 DOI [10.1186/s12915-017-0454-7](https://doi.org/10.1186/s12915-017-0454-7).
- Dentzien-Dias PC, Poinar G Jr, De Figueiredo AEQ, Pacheco ACL, Horn BLD, Schultz CL, Turrens JF. 2013. Tapeworm eggs in a 270 million-year-old shark coprolite. *PLOS ONE* **8(1)**:e55007 DOI [10.1371/journal.pone.0055007](https://doi.org/10.1371/journal.pone.0055007).
- Dhakan DB, Maji A, Sharma AK, Saxena R, Pulikkan J, Grace T, Gomez A, Scaria J, Amato KR, Sharma VK. 2019. The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. *GigaScience* **8(3)**:804 DOI [10.1093/gigascience/giz004](https://doi.org/10.1093/gigascience/giz004).
- Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017. Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35(4)**:316–319 DOI [10.1038/nbt.3820](https://doi.org/10.1038/nbt.3820).
- Ewels P, Peltzer A, Fillinger S, Alneberg J, Patel H, Wilm A, Garcia M, Di Tommaso P, Nahnsen S. 2019. nf-core: Community curated bioinformatics pipelines. *bioRxiv* 610741 DOI [10.1101/610741](https://doi.org/10.1101/610741).
- Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH, Caporaso JG. 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences of the United States of America* **109(52)**:21390–21395 DOI [10.1073/pnas.1215210110](https://doi.org/10.1073/pnas.1215210110).
- Frantz LAF, Mullin VE, Pionnier-Capitan M, Lebrasseur O, Ollivier M, Perri A, Linderholm A, Mattiangeli V, Teasdale MD, Dimopoulos EA, Tresset A, Duffraisse M, McCormick F, Bartosiewicz L, Gál E, Nyerges Éva A, Sablin MV, Bréhard S, Mashkour M, Bălăşescu A, Gillet B, Hughes S, Chassaing O, Hitte C, Vigne J-D, Dobney K, Hänni C, Bradley DG, Larson G. 2016. Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science* **352(6290)**:1228–1231 DOI [10.1126/science.aaf3161](https://doi.org/10.1126/science.aaf3161).
- Gilbert MTP, Jenkins DL, Götherstrom A, Naveran N, Sanchez JJ, Hofreiter M, Thomsen PF, Binladen J, Higham TFG, Yohe RM, Parr R, Cummings LS, Willerslev E. 2008. Dna from pre-clovis human coprolites in oregon, north america. *Science* **320(5877)**:786–789 DOI [10.1126/science.1154116](https://doi.org/10.1126/science.1154116).
- Guiry EJ. 2012. Dogs as analogs in stable isotope-based human paleodietary reconstructions: a review and considerations for future use. *Journal of Archaeological Method and Theory* **19(3)**:351–376 DOI [10.1007/s10816-011-9118-z](https://doi.org/10.1007/s10816-011-9118-z).
- Hagan RW, Hofman CA, Hübner A, Reinhard K, Schnorr S, Lewis CM, Sankaranarayanan K, Warinner CG. 2020. Comparison of extraction methods for recovering ancient microbial DNA from paleofeces. *American Journal of Physical Anthropology* **171(2)**:275–284 DOI [10.1002/ajpa.23978](https://doi.org/10.1002/ajpa.23978).
- Hampton-Marcell JT, Lopez JV, Gilbert JA. 2017. The human microbiome: an emerging tool in forensics. *Microbial Biotechnology* **10(2)**:228–230 DOI [10.1111/1751-7915.12699](https://doi.org/10.1111/1751-7915.12699).
- Hofreiter M, Poinar HN, Spaulding WG, Bauer K, Martin PS, Possnert G, Pääbo S. 2000. A molecular analysis of ground sloth diet through the last glaciation. *Molecular Ecology* **9(12)**:1975–1984 DOI [10.1046/j.1365-294X.2000.01106.x](https://doi.org/10.1046/j.1365-294X.2000.01106.x).
- Jiménez FA, Gardner SL, Araújo A, Fugassa M, Brooks RH, Racz E, Reinhard KJ. 2012. Zoonotic and human parasites of inhabitants of cueva de los muertos chiquitos, Rio Zape Valley, Durango, Mexico. *Journal of Parasitology* **98(2)**:304–310 DOI [10.1645/GE-2915.1](https://doi.org/10.1645/GE-2915.1).

- Kho ZY, Lal SK. 2018.** The human gut microbiome: a potential controller of wellness and disease. *Frontiers in Microbiology* 9:215 DOI 10.3389/fmicb.2018.01835.
- Kirch P, O'Day SJ. 2003.** New archaeological insights into food and status: a case study from pre-contact Hawaii. *World Archaeology* 34(3):484–497 DOI 10.1080/0043824021000026468.
- Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, Bushman FD, Knight R, Kelley ST. 2011.** Bayesian community-wide culture-independent microbial source tracking. *Nature Methods* 8(9):761–763 DOI 10.1038/nmeth.1650.
- Langmead B, Salzberg SL. 2012.** Fast gapped-read alignment with bowtie 2. *Nature methods* 9(4):357–359 DOI 10.1038/nmeth.1923.
- Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R, Gordon JI. 2008.** Evolution of mammals and their gut microbes. *Science* 320(5883):1647–1651 DOI 10.1126/science.1155725.
- Mann AE, Sabin S, Ziesemer K, Vågane s J, Schroeder H, Ozga AT, Sankaranarayanan K, Hofman CA, Yates JAF, Salazar-García DC, Frohlich B, Aldenderfer M, Hoogland M, Read C, Milner GR, Stone AC, Lewis CM, Krause J, Hofman C, Bos KI, Warinner C. 2018.** Differential preservation of endogenous human and microbial DNA in dental calculus and dentin. *Scientific Reports* 8(1):9822 DOI 10.1038/s41598-018-28091-9.
- Meyer M, Kircher M. 2010.** Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols* 2010(6):pdb.prot5448 DOI 10.1101/pdb.prot5448.
- Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, Zech Xu Z, Van Treuren W, Knight R, Gaffney PM, Spicer P, Lawson P, Marin-Reyes L, Trujillo-Villarreal O, Foster M, Gujja-Poma E, Troncoso-Corzo L, Warinner C, Ozga AT, Lewis CM. 2015.** Subsistence strategies in traditional societies distinguish gut microbiomes. *Nature Communications* 6(1):6505 DOI 10.1038/ncomms7505.
- Orellana LH, Chee-Sanford JC, Sanford RA, Löffler FE, Konstantinidis KT. 2018.** Year-Round Shotgun Metagenomes Reveal Stable Microbial Communities in Agricultural Soils and Novel Ammonia Oxidizers Responding to Fertilization. *Applied and Environmental Microbiology* 84(2):e0164617.
- Passoli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. 2019.** Extensive unexplored human microbiome diversity revealed by Over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 176(3):649–662.e20 DOI 10.1016/j.cell.2019.01.001.
- Peltzer A, Neukamm J. 2019.** Integrative-Transcriptomics/DamageProfiler: DamageProfiler v0.4.7. Available at <https://doi.org/10.5281/zenodo.1064062>.
- Podberscek AL. 2009.** Good to pet and eat: the keeping and consuming of dogs and cats in South Korea. *Journal of Social Issues* 65(3):615–632 DOI 10.1111/j.1540-4560.2009.01616.x.
- Poinar H, Fiedel S, King CE, Devault AM, Bos K, Kuch M, Debruyne R. 2009.** Comment on DNA from pre-clovis human coprolites in Oregon, North America. *Science* 325(5937):148 DOI 10.1126/science.1168182.
- Poinar HN, Hofreiter M, Spaulding WG, Martin PS, Stankiewicz BA, Bland H, Evershed RP, Possnert G, Pääbo S. 1998.** Molecular coproscopy: dung and diet of the extinct ground sloth *nothrotheriops shastensis*. *Science* 281(5375):402–406 DOI 10.1126/science.281.5375.402.
- Poinar HN, Kuch M, Sobolik KD, Barnes I, Stankiewicz AB, Kuder T, Spaulding WG, Bryant VM, Cooper A, Pääbo S. 2001.** A molecular analysis of dietary diversity for three

- archaic native Americans. *Proceedings of the National Academy of Sciences of the United States of America* **98**(8):4317–4322 DOI [10.1073/pnas.061014798](https://doi.org/10.1073/pnas.061014798).
- Pysam Developers. 2018.** Pysam: a python module for reading and manipulating files in the sam/bam format. Available at <https://github.com/pysam-developers/pysam>.
- Rampelli S, Schnorr S, Consolandi C, Turrone S, Severgnini M, Peano C, Brigidi P, Crittenden A, Henry A, Candela M. 2015.** Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Current Biology* **25**(13):1682–1693 DOI [10.1016/j.cub.2015.04.055](https://doi.org/10.1016/j.cub.2015.04.055).
- Reinhard KJ, Bryant VM. 1992.** Coprolite analysis: a biological perspective on archaeology. *Archaeological Method and Theory* **4**:245–288.
- Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D. 2015.** Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**(1660):20130624 DOI [10.1098/rstb.2013.0624](https://doi.org/10.1098/rstb.2013.0624).
- Rosenswig RM. 2007.** Beyond identifying elites: feasting as a means to understand early middle formative society on the pacific coast of mexico. *Journal of Anthropological Archaeology* **26**(1):1–27 DOI [10.1016/j.jaa.2006.02.002](https://doi.org/10.1016/j.jaa.2006.02.002).
- Santley RS, Rose EK. 1979.** Diet, nutrition and population dynamics in the basin of mexico. *World Archaeology* **11**(2):185–207 DOI [10.1080/00438243.1979.9979760](https://doi.org/10.1080/00438243.1979.9979760).
- Schubert M, Lindgreen S, Orlando L. 2016.** Adapterremoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes* **9**(1):88 DOI [10.1186/s13104-016-1900-2](https://doi.org/10.1186/s13104-016-1900-2).
- Sharpton TJ. 2014.** An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science* **5**(e1002358):209 DOI [10.3389/fpls.2014.00209](https://doi.org/10.3389/fpls.2014.00209).
- Shenhav L, Thompson M, Joseph TA, Briscoe L, Furman O, Bogumil D, Mizrahi I, Pe'er I, Halperin E. 2019.** FEAST: fast expectation-maximization for microbial source tracking. *Nature Methods* **16**(7):627–632 DOI [10.1038/s41592-019-0431-x](https://doi.org/10.1038/s41592-019-0431-x).
- Sistiaga A, Mallol C, Galván B, Summons RE, Hardy K. 2014.** The neanderthal meal: a new perspective using faecal biomarkers. *PLOS ONE* **9**(6):e101045 DOI [10.1371/journal.pone.0101045](https://doi.org/10.1371/journal.pone.0101045).
- Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Pääbo S, Krause J, Jakobsson M. 2014.** Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences of the United States of America* **111**(6):2229–2234 DOI [10.1073/pnas.1318934111](https://doi.org/10.1073/pnas.1318934111).
- The Human Microbiome Project Consortium. 2012.** Structure, function and diversity of the healthy human microbiome. *Nature* **486**(7402):207–214 DOI [10.1038/nature11234](https://doi.org/10.1038/nature11234).
- Tito RY, Knights D, Metcalf J, Obregon-Tito AJ, Cleeland L, Najjar F, Roe B, Reinhard K, Sobolik K, Belknap S, Foster M, Spicer P, Knight R, Lewis CM. 2012.** Insights from characterizing extinct human gut microbiomes. *PLOS ONE* **7**(12):e51146.
- Tito RY, Macmil S, Wiley G, Najjar F, Cleeland L, Qu C, Wang P, Romagne F, Leonard S, Ruiz AJ, Reinhard K, Roe BA, Lewis CM Jr, Ahmed N. 2008.** Phylotyping and functional analysis of two ancient human microbiomes. *PLOS ONE* **3**(11):e3703 DOI [10.1371/journal.pone.0003703](https://doi.org/10.1371/journal.pone.0003703).
- Warinner C, Herbig A, Mann A, Fellows Yates JA, Weiß CL, Burbano HA, Orlando L, Krause J. 2017.** A robust framework for microbial archaeology. *Annual Review of Genomics and Human Genetics* **18**(1):321–356 DOI [10.1146/annurev-genom-091416-035526](https://doi.org/10.1146/annurev-genom-091416-035526).
- Warinner C, Lewis CM Jr. 2015.** Microbiome and health in past and present human populations. *American Anthropologist* **117**(4):740–741 DOI [10.1111/aman.12367](https://doi.org/10.1111/aman.12367).

- Warinner C, Speller C, Collins MJ, Lewis CM Jr. 2015.** Ancient human microbiomes. *Journal of Human Evolution* 79:125–136 DOI [10.1016/j.jhevol.2014.10.016](https://doi.org/10.1016/j.jhevol.2014.10.016).
- Wing ES. 1978.** *Use of dogs for food: an adaptation to the coastal environment*. Amsterdam: Elsevier.
- Wood DE, Salzberg SL. 2014.** Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 15(3):R46 DOI [10.1186/gb-2014-15-3-r46](https://doi.org/10.1186/gb-2014-15-3-r46).
- Wood JR, Crown A, Cole TL, Wilmshurst JM. 2016.** Microscopic and ancient DNA profiling of polynesian dog (kur) coprolites from northern New Zealand. *Journal of Archaeological Science: Reports* 6:496–505 DOI [10.1016/j.jasrep.2016.03.020](https://doi.org/10.1016/j.jasrep.2016.03.020).

From alignments to assemblies

Manuscript C: *sam2lca*: Lowest Common Ancestor for SAM, BAM, CRAM alignment files

Maxime Borry, Alexander Hübner, and Christina Warinner

Published in The Journal of Open Source Software, 2022 June 01; DOI: 10.21105/joss.04360

In manuscript C, we propose a new method to apply the LCA algorithm on files in SAM/BAM/CRAM format. While the field of microbiome research has favored alignment free methods for taxonomic classification, in aDNA metagenomics, the need for alignment remained to be able to assess the deamination damage patterns. The main tool to perform this kind of classification was MALT (Herbig et al., 2016), but with the ever growing reference database size, it became computationally intractable to fit an up-to-date index in memory. To circumvent this issue, we propose the *sam2lca* method. By integrating *sam2lca* in the nextflow computational pipeline *adnamap* (Borry, 2023), employing an divide and conquer approach, we can map sequencing reads to many reference indices in parallel, potentially located on different machines of a cluster or cloud computing service. After all reads have been mapped to all reference indices, all SAM/BAM/CRAM alignment are gathered and merged thanks to the `merge` command from the `samtools` utility (Li et al., 2009). *sam2lca* then applies the LCA algorithm on the merged alignment file. This approach allows for a much easier scaling of the reference databases, with improved index flexibility: adding a new reference to an index does not require to rebuilt it entirely, but only to add a new index for this reference index. Finally, because *sam2lca* uses the common SAM/BAM/CRAM, it is agnostic to the software used to generate the alignments.

Authors contributions

- **The candidate is:** first author.
- **Status:** published

Authors' contributions (in %, from 10%) to the given categories of the publication

Author	Conceptual	Data analysis	Experimental	Writing the manuscript	Provision of material
Maxime Borry	70	80	-	70	-
Alexander Hübner	20	20	-	15	-
Christina Warinner	10	-	-	15	-
Total:	100%	100%	-	100%	-



sam2lca: Lowest Common Ancestor for SAM/BAM/CRAM alignment files

Maxime Borry ¹, Alexander Hübner ^{1,2}, and Christina Warinner ^{1,2,3}

1 Microbiome Sciences Group, Max Planck Institute for Evolutionary Anthropology, Department of Archaeogenetics, Leipzig, Germany **2** Faculty of Biological Sciences, Friedrich-Schiller Universität Jena, Jena, Germany **3** Department of Anthropology, Harvard University, Cambridge, MA, United States of America

DOI: [10.21105/joss.04360](https://doi.org/10.21105/joss.04360)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Jacob Schreiber](#)  

Reviewers:

- [@fasnicar](#)
- [@marouenbg](#)

Submitted: 21 April 2022

Published: 01 June 2022

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Summary

sam2lca is a program performing reference sequence disambiguation for reads mapping to multiple reference sequences in a shotgun metagenomics sequencing dataset. To do so, it takes as input the common SAM sequence alignment format and applies the lowest common ancestor algorithm.

Statement of need

The rapidly decreasing cost of massively parallel short-read DNA sequencing technologies has enabled the genetic characterization of entire ecological communities, a technique known as shotgun metagenomics.

In a typical shotgun metagenomics approach, after the DNA of an ecological community has been sequenced, it is compared to a genetic reference database of organisms with known taxonomy. Even though the number of DNA sequences and genomes in reference databases is constantly growing, there are still instances where a query sequence will not have a direct match in a reference database, and it will instead weakly align to one or more distantly related reference organisms. Furthermore, when analyzing short DNA sequences, a query DNA sequence will often match equally well to more than one reference organism, posing a challenge for its taxonomic assignment.

One solution to this problem is to apply a lowest common ancestor algorithm (LCA) ([Figure 1](#)) during taxonomic profiling to place such ambiguous assignments higher in a taxonomic tree, where they can be more confidently assigned. This idea was first implemented for metagenomics with the MEGAN program ([Huson et al., 2007](#)).

Many programs have since been developed to perform LCA during taxonomic profiling. For example, MALT ([Herbig et al., 2017](#)) and MetaPhlan ([Segata et al., 2012](#)) perform LCA and taxonomic profiling after DNA sequence alignment, while other programs, such as Kraken2 ([Wood et al., 2019](#)) and Centrifuge ([Kim et al., 2016](#)), are alignment-free methods that apply LCA after k-mer matching. While combining the steps of database matching and LCA into one program can be useful, it also limits user choice for the selection of different alignment or k-mer matching programs.

With sam2lca, we propose to decouple the LCA step from the alignment step to allow the end-user to freely choose from one of the many DNA sequence aligner programs available, such as Bowtie2 ([Langmead & Salzberg, 2012](#)), bwa ([Li & Durbin, 2009](#)), bmap ([Bushnell, 2014](#)), or minimap2 ([Li, 2018](#)). Each of these aligners exports the sequence alignments in

the widely adopted Sequence Alignment Map format (SAM) (Li et al., 2009), or in its binary (BAM), or compressed representation (CRAM), which sam2lca uses as an input.

The use of the SAM file format enables easier integration of sam2lca in a wide variety of analysis workflows, which often already contain steps generating or using SAM/BAM/CRAM files, and allows for an easy subsequent analysis using well-established programs, such as SAMtools (Li et al., 2009).

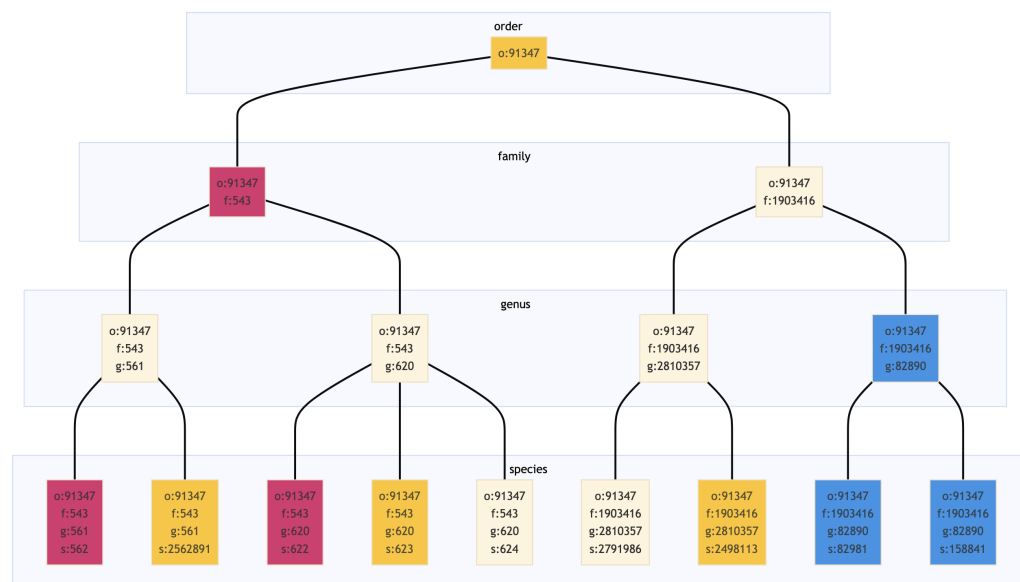


Figure 1: Example of the LCA algorithm with NCBI TAXIDs. Taxons and their LCA are displayed in the same color. The lineage for each taxon is shown with a one letter code for the rank, and the corresponding TAXID. The LCA of s:562 (*E. coli* species) and s:622 (*S. dysenteriae* species) is f:543 (*Enterobacteriaceae* family). The LCA of s:82981 (*L. grimontii* species) and s:158841 (*L. richardii* species) is g:82890 (*Leminorella* genus). The LCA of s:2562891 (*E. alba* species), s:623 (*S. flexneri* species) and s:2498113 (*J. zhutongyuui* species) is o:91347 (Enterobacterales order)

Implementation

sam2lca is a program written in Python, which takes as an input an indexed and sorted SAM/BAM/CRAM alignment file. Broadly, the program consists of four main steps. First, reference sequence accessions, present in the BAM file header section, are converted to taxonomic identifiers (TAXID) using a RocksDB persistent key-value store (Dong et al., 2021). The alignment section of the BAM file is then parsed with Pysam (pysam-developers, 2022) and a dictionary is created to match single and multi-mapping query sequences/reads to the TAXID(s) of their matching reference sequence(s). Next, if a read has been matched to multiple TAXIDs, the LCA implementation of Taxopy (Camargo, 2022) is used to attribute it to the lowest common ancestor, using the NCBI taxonomy by default. Finally, each TAXID is used to retrieve its associated taxon's scientific name and taxonomic lineage, and results are saved in a JSON and CSV file. Optionally, a BAM file, similar to the input file, can be generated. This BAM file contains for each read an additional XT tag added to report the TAXID of the LCA for each read, an XN tag for the taxon's scientific name, and finally an XR tag for the taxon's rank. sam2lca is distributed through pip and conda, and the documentation and tutorials are available at sam2lca.readthedocs.io

Acknowledgements

This research was supported by the Werner Siemens Stiftung (M.B. and C.W.) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2051, Project-ID 390713860 (A.H. and C.W.).

References

- Bushnell, B. (2014). *BBMap: A fast, accurate, splice-aware aligner*. <https://www.osti.gov/biblio/1241166>
- Camargo, A. (2022). Taxopy: A python package for manipulating NCBI-formatted taxonomic databases. In *GitHub repository*. GitHub. <https://github.com/apcamargo/taxopy>
- Dong, S., Kryczka, A., Jin, Y., & Stumm, M. (2021). RocksDB: Evolution of development priorities in a key-value store serving large-scale applications. *ACM Transactions on Storage (TOS)*, 17(4), 1–32. <https://doi.org/10.1145/3483840>
- Herbig, A., Maixner, F., Bos, K. I., Zink, A., Krause, J., & Huson, D. H. (2017). MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the tyrolean iceman. *BioRxiv*. <https://doi.org/10.1101/050559>
- Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, 17(3), 377–386. <https://doi.org/10.1101/gr.5969107>
- Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research*, 26(12), 1721–1729. <https://doi.org/10.1101/gr.210641.116>
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4), 357–359. <https://doi.org/10.1038/nmeth.1923>
- Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- pysam-developers. (2022). Pysam: A python module for reading and manipulating files in the SAM/BAM format. In *GitHub repository*. GitHub. <https://github.com/pysam-developers/pysam>
- Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8), 811–814. <https://doi.org/10.1038/nmeth.2066>
- Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome Biology*, 20(1), 1–13. <https://doi.org/10.1186/s13059-019-1891-0>

Manuscript D: *PyDamage: automated ancient damage identification and estimation for contigs in ancient DNA de novo assembly*

Maxime Borry, Alexander Hübner, Adam B. Rohrlach, and Christina Warinner

Published in PeerJ, 2021 July 27; DOI: 10.7717/peerj.11845

In Manuscript D, we introduce a new method to statistically assess the deamination damages of multiple aDNA sequences in parallel.

Tools to assess the deamination damages of aDNA sequences, originally described by Briggs et al. (2007), were first implemented in 2011 with mapDamage (Ginolhac et al., 2011), later refined into mapDamage2 (Jónsson et al., 2013), or the more recent DamageProfiler (Neukamm et al., 2021b). While mapDamage2 already integrated a statistical model of the aDNA deamination patterns, it was designed for single-genome alignment applications, and its statistical model did not scale well when applied to many different references. With the adoption of *de novo* assembly by the aDNA community (Wibowo et al., 2021), there was a nascent need to be able to assess the aDNA damage patterns for each assembled contig and/or MAG, resulting in potentially many thousands of references to check for deamination. To address this new challenge, we extended the approach of deamination assessment with a likelihood ratio test (LRT) first introduced in PMDTools (Skoglund et al., 2014). To do so, we first both a null and a damage model, then compare them with a LRT to compute a damage test statistic with a for each reference sequence, in parallel. To validate the performances of pyDamage, we simulated data and built a linear model to assess the accuracy of pyDamage predictions for varying amount of available sequencing data.

Authors contributions

- **The candidate is:** first author.
- **Status:** published

Authors' contributions (in %, from 10%) to the given categories of the publication

Author	Conceptual	Data analysis	Experimental	Writing the manuscript	Provision of material
Maxime Borry	60	30	-	50	-
Alexander Hübner	10	30	-	15	-
Adam B. Rohrlach	20	30	-	15	-
Christina Warinner	10	10	-	20	-
Total:	100%	100%	-	100%	-



PyDamage: automated ancient damage identification and estimation for contigs in ancient DNA *de novo* assembly

Maxime Borry¹, Alexander Hübner^{1,2}, Adam B. Rohrlach^{3,4} and Christina Warinner^{1,2,5}

¹ Microbiome Sciences Group, Max Planck Institute for the Science of Human History, Department of Archaeogenetics, Jena, Germany

² Faculty of Biological Sciences, Friedrich-Schiller Universität Jena, Jena, Germany

³ Population Genetics Group, Max Planck Institute for the Science of Human History, Department of Archaeogenetics, Jena, Germany

⁴ ARC Centre of Excellence for Mathematical and Statistical Frontiers, The University of Adelaide, Adelaide, Australia

⁵ Department of Anthropology, Harvard University, Cambridge, MA, United States of America

ABSTRACT

DNA *de novo* assembly can be used to reconstruct longer stretches of DNA (contigs), including genes and even genomes, from short DNA sequencing reads. Applying this technique to metagenomic data derived from archaeological remains, such as paleofeces and dental calculus, we can investigate past microbiome functional diversity that may be absent or underrepresented in the modern microbiome gene catalogue. However, compared to modern samples, ancient samples are often burdened with environmental contamination, resulting in metagenomic datasets that represent mixtures of ancient and modern DNA. The ability to rapidly and reliably establish the authenticity and integrity of ancient samples is essential for ancient DNA studies, and the ability to distinguish between ancient and modern sequences is particularly important for ancient microbiome studies. Characteristic patterns of ancient DNA damage, namely DNA fragmentation and cytosine deamination (observed as C-to-T transitions) are typically used to authenticate ancient samples and sequences, but existing tools for inspecting and filtering aDNA damage either compute it at the read level, which leads to high data loss and lower quality when used in combination with *de novo* assembly, or require manual inspection, which is impractical for ancient assemblies that typically contain tens to hundreds of thousands of contigs. To address these challenges, we designed PyDamage, a robust, automated approach for aDNA damage estimation and authentication of *de novo* assembled aDNA. PyDamage uses a likelihood ratio based approach to discriminate between truly ancient contigs and contigs originating from modern contamination. We test PyDamage on both on simulated aDNA data and archaeological paleofeces, and we demonstrate its ability to reliably and automatically identify contigs bearing DNA damage characteristic of aDNA. Coupled with aDNA *de novo* assembly, Pydamage opens up new doors to explore functional diversity in ancient metagenomic datasets.

Submitted 29 March 2021

Accepted 1 July 2021

Published 27 July 2021

Corresponding authors

Maxime Borry, borry@shh.mpg.de

Christina

Warinner, warinner@fas.harvard.edu

Academic editor

Rodolfo Aramayo

Additional Information and
Declarations can be found on
page 16

DOI 10.7717/peerj.11845

© Copyright
2021 Borry et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Anthropology, Bioinformatics, Computational Biology, Genomics, Paleontology

Keywords metagenomics, aDNA, ancient DNA, assembly, damage, *de novo*, automated

INTRODUCTION

Ancient DNA (aDNA) is highly fragmented (*Orlando et al., 2021; Warinner et al., 2017*). Although genomic DNA molecules within a living organism can be millions to hundreds of millions of base pairs (bp) long, postmortem enzymatic and chemical degradation after death quickly reduces DNA to fragment lengths of less than 150 bp, typically with medians less than 75 bp and modes less than 50 bp (*Mann et al., 2018; Hansen et al., 2017*). Within the field of metagenomics, many approaches require longer stretches of DNA for adequate analysis, a requirement that particularly applies to functional profiling, which often involves *in silico* translation steps (*Seemann, 2014*). For example, in our experiments we observed that FragGeneScan (*Rho, Tang & Ye, 2010*), a tool designed for gene prediction from short read data, failed to predict open-reading frames in any DNA sequences shorter than 60 bp. If applied directly to highly fragmented ancient metagenomic datasets, such data filtering can introduce biases that interfere with functional analyses when preservation is variable across samples or when comparing ancient samples to modern ones.

Because very short (<100 bp) and ultrashort (<50 bp) DNA molecules pose many downstream analytical challenges, there is a long-standing interest in leveraging the approach of *de novo* assembly to computationally reconstruct longer stretches of DNA for analysis. With *de novo* assembly, longer contiguous DNA sequences (contigs), and sometimes entire genes or gene clusters, can be reconstructed from individual sequencing reads (*Compeau, Pevzner & Tesler, 2011*), which can then be optionally binned into metagenome-assembled genomes (MAGs) (*Kang et al., 2015*). Such contigs are more amenable to functional profiling, and applying this technique to microbial metagenomics datasets derived from archaeological remains, such as paleofeces and dental calculus, has the potential to reveal ancient genes and functional diversity that may be absent or underrepresented in modern microbiomes (*Tett et al., 2019; Wibowo et al., 2021; Brealey et al., 2020*). However, because ancient samples generally contain a mixture of ancient bacterial DNA and modern bacterial contaminants, it is essential to distinguish, among the thousands of contigs generated by assembly, truly ancient contigs from contigs that may originate from the modern environment, such as the excavation site, storage facility, or other exogenous sources.

In addition to being highly fragmented, aDNA also contains other forms of characteristic molecular decay, namely cytosine deamination (observed as C → T transitions in aDNA datasets) (*Dabney, Meyer & Pääbo, 2013*), which can be measured and quantified to indicate the authenticity of an ancient sample, or even an individual sequence (*Hofreiter et al., 2001; Briggs et al., 2007*). However, tools for inspecting and filtering aDNA damage were primarily designed for genomic and not metagenomic applications, and they are largely unsuited or impractical for use in combination with *de novo* assembly. For example, PMDTools (*Skoglund et al., 2014*) operates at the read level, and when subsequently combined with *de novo* assembly leads to higher data loss and lower overall assembly quality. MapDamage (*Ginolhac et al., 2011*) and DamageProfiler (*Neukamm, Peltzer & Nieselt, 2020*) are tools that can be applied to assembled contigs, but require manual contig inspection by the user, which is infeasible for *de novo* assemblies yielding tens to hundreds

of thousands of contigs. Other tools, such as MapDamage (Jónsson *et al.*, 2013), do provide an estimation of damage, but use slower algorithms that do not scale well to the analysis of many thousands of contigs. Even tools, such as HOPS (Hübler *et al.*, 2019), designed for aDNA metagenomics, can not easily scale for the analysis of the sheer number of unknown contigs generated by the assembly process. A faster, automated approach with a better sensitivity for distinguishing truly ancient contigs from modern environmental contigs is needed.

Here, we present PyDamage, a software tool to automate the process of contig damage identification and estimation. PyDamage models aDNA damage from deamination data (C → T transitions), and tests for damage significance using a likelihood ratio test to discriminate between truly ancient contigs and contigs originating from modern contaminants. Testing PyDamage on *in silico* simulated data, we show that it is able to accurately distinguish ancient and modern contigs. We then apply PyDamage to *de novo* assembled DNA from ancient paleofeces from the site of Cueva de los Muertos Chiquitos, Mexico (ca. 1300 BP) and find that the contigs PyDamage identifies as ancient are consistent with taxa known to be members of the human gut microbiome. Among the ancient contigs, PyDamage authenticated multiple functional genes of interest, including a multidrug and bile salt resistance gene cluster from the gut microbe *Treponema succinifacians*, a species that is today only found in societies practicing traditional forms of subsistence. Using PyDamage, *de novo* assembled contigs from aDNA datasets can be rapidly and robustly authenticated for a variety of downstream metagenomics applications.

MATERIAL AND METHODS

Simulated sequencing data

In order to evaluate the performance of PyDamage with respect to the GC content of the assembled genome, the sequencing depth along the genome, the amount of observed aDNA damage on the DNA fragments, and the mean length of these DNA fragments, we simulated short-read sequencing data using gargammel (Renaud *et al.*, 2017) varying these four parameters. We chose three microbiome-associated microbial taxa with low (*Methanobrevibacter smithii*, 31%), medium (*Tannerella forsythia*, 47%), and high (*Actinomyces dentalis*, 72%) GC content, following Mann *et al.* (2018) (Fig. 1A). Using three different read length distributions (Fig. 1B), we generated short-read sequencing data from each reference genome using gargammel's *fragSim*. To the resulting short-read sequences we added different amounts of aDNA damage using gargammel's *deamSim* so that ten levels of damage ranging from 0% to 20% were observed, which were measured as the amount of observed C → T substitutions on the terminal base at the 5' end of the DNA fragments (Fig. 1C). Finally, each of these 90 simulated datasets was subsampled to generate nine coverage bins ranging from 1-fold to 500-fold genome coverage by randomly drawing a coverage value from the uniform distribution defining each bin (Fig. 1D) and these were aligned to their respective reference genome using BWA *aln* (Li & Durbin, 2009) with the non-default parameters optimized for aDNA `-n 0.01 -o 2 -l 16500` (Meyer *et al.*, 2012).

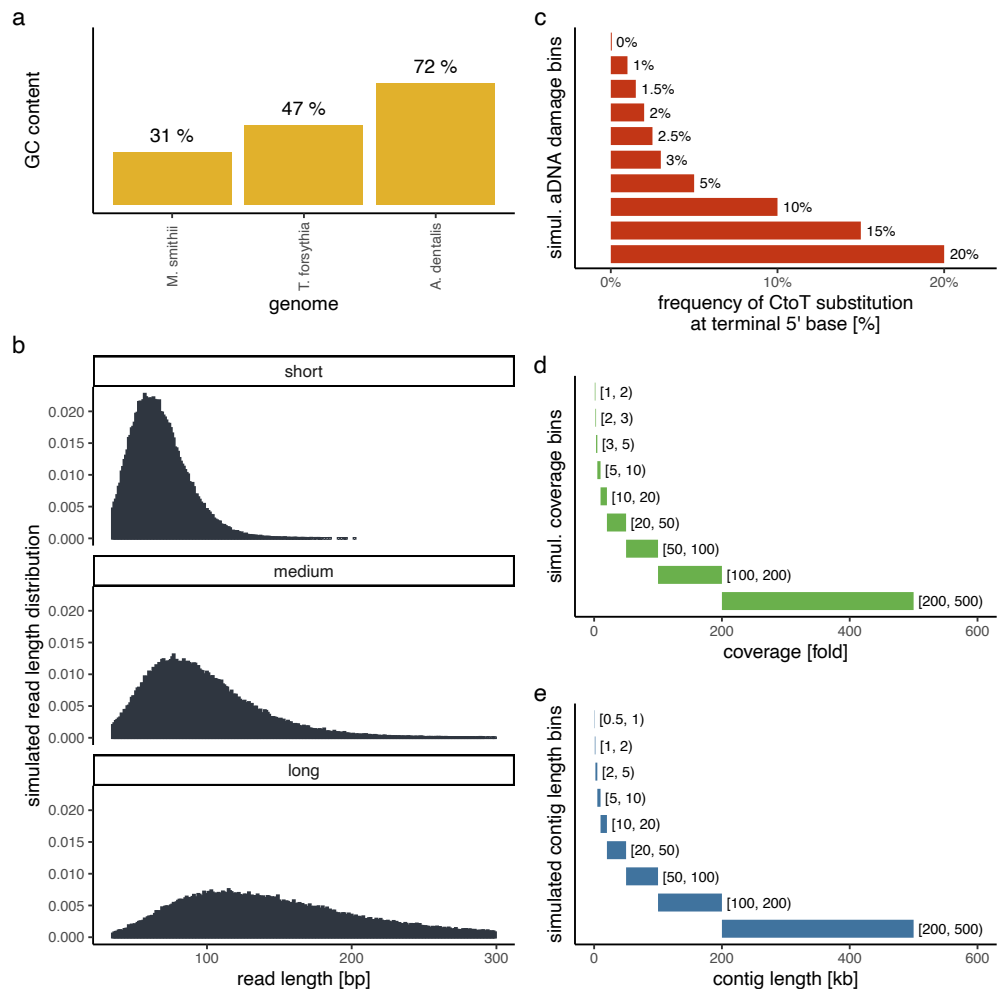


Figure 1 Simulation scheme for evaluating the performance of PyDamage. (A) The GC content of the three microbial reference genomes. (B) The read length distributions used as input into gargammel *fragSim*. (C) The amount of aDNA damage as observed as the frequency of C → T substitutions on the terminal 5' end of the DNA fragments that was added using gargammel *deamSim*. (D) Nine coverage bins from which the exact coverage was sampled by randomly drawing a number from the uniform distribution defining the bin. (E) Nine contig length bins from which the exact contig length was sampled by randomly drawing a number from the uniform distribution defining the bin.

Full-size DOI: [10.7717/peerj.11845/fig-1](https://doi.org/10.7717/peerj.11845/fig-1)

Test contigs of different length were simulated by defining nine contig length bins ranging from 0.5 kb to 500 kb length (Fig. 1E) and randomly drawing 100 contig lengths from the respective uniform distribution defining each bin. Next, we chose the location of these test contigs by randomly selecting a contig from all contigs of sufficient length. We determined the exact location on the selected test contig from the reference genome by randomly drawing the start position from the uniform distribution defined by the length of the selected reference contig. This resulted in 900 test contigs per reference genome. Using these test contigs, we selected the aligned DNA fragments of the simulated sequencing data that overlapped the region defined by the contig and evaluated them using PyDamage.

In total, we evaluated 702,900 test contigs (243,000 contigs for both *M. smithii* and *T. forsythia*, and 216,000 contigs for *A. dentalis*, for which no reference contig longer than 200 kb was available).

Archaeological sample Preparation and sequencing

We re-analyzed ancient metagenomic data from the archaeological paleofeces sample ZSM028 (Zape 28) dating to ca. 1300 BP from the site of Cueva de los Muertos Chiquitos, in Mexico, previously published in [Borry et al. \(2020\)](#) (ENA run accession codes ERR3678595, ERR3678598, ERR3678602, ERR3678603, and ERR3678613).

Bioinformatic processing

The ZSM028 sample was first trimmed to remove adapters, low quality sequences with Q-scores below 20, and short sequences below 30 bp using AdapterRemoval ([Schubert, Lindgreen & Orlando, 2016](#)) v2.3.1. The reads were *de novo* assembled into contigs using MetaSPAdes [Nurk et al. \(2017\)](#) v3.13.1 using the non-default k-mer lengths 21, 33, and 45. The set of k-mer lengths was adapted to consider the on-average short length of the DNA molecules of this sample (mode: 37 bp). We selected a k-mer length of 45 as the longest one since this was the next longer uneven k-mer length of the median DNA molecule length (median: 44 bp). Reads were then mapped back to the contigs with length > 1,000 bp using Bowtie2 ([Langmead & Salzberg, 2012](#)), in the very-sensitive mode, while allowing up to 1 mismatch in the seeding process. The alignment files were then given as an input to PyDamage v0.50. Contigs passing filtering thresholds were functionally annotated with Prokka v1.14.6 ([Seemann, 2014](#)), using the `--metagenome` flag.

Contig Taxonomic Profiling

To investigate the taxonomic profile of the contigs that passed the PyDamage filtering, we ran Kraken2 v2.1.1 ([Wood, Lu & Langmead, 2019](#)) using the PlusPFP database (<https://benlangmead.github.io/aws-indexes/k2>) from 27/1/2021. We then generated the Sankey plot using Pavian ([Breitwieser & Salzberg, 2016](#)).

PyDamage implementation

PyDamage takes alignment files of reads (in SAM, BAM, or CRAM format) mapped against reference sequences (i.e., contigs, a MAG, a genome, or any other reference sequences of DNA). For each read mapping to each reference sequence j , using pysam ([pysam developers, 2018](#)), we count the number of apparent C \rightarrow T transitions at each position which is i bases from the 5' terminal end, $i \in \{0, 1, \dots, k\}$, denoted N_i^j (by default, we set $k = 35$). Similarly we denote the number of observed conserved 'C-to-C' sites M_i^j , thus

$$M^j = (M_0^j, \dots, M_k^j) \quad \text{and} \quad N^j = (N_0^j, \dots, N_k^j).$$

Finally, we calculate the proportion of C \rightarrow T transitions occurring at each position, denoted \hat{p}_i^j , in the following way:

$$\hat{p}_i^j = \frac{N_i^j}{M_i^j + N_i^j}.$$

For D_i , the event that we observe a $C \rightarrow T$ transition i bases from the terminal end, we define two models: a null model \mathcal{M}_0 (Eq. (1)) which assumes that damage is independent of the position from the 5' terminal end, and a damage model \mathcal{M}_1 (Eq. (2)) which assumes a decreasing probability of damage the further a the position from the 5' terminal end. For the damage model, we re-scale the curve to the interval defined by parameters $[d_{pmin}^j, d_{pmax}^j]$.

$$P_0(D_i|p_0, j) = p_0 = \mathcal{M}_0 \pi^j \quad (1)$$

$$\begin{aligned} P_1(D_i|p_d^j, d_{pmin}^j, d_{pmax}^j, j) &= \frac{\left([(1-p_d^j)^i \times p_d^j] - \hat{p}_{min}^j \right)}{\hat{p}_{max}^j - \hat{p}_{min}^j} \times (d_{pmax}^j - d_{pmin}^j) + d_{pmin}^j \\ &= \mathcal{M}_1 \pi_i^j, \end{aligned} \quad (2)$$

where

$$\hat{p}_{min}^j(p_d^j) = (1-p_d^j)^k \times p_d^j \quad \text{and} \quad \hat{p}_{max}^j(p_d^j) = (1-p_d^j)^0 \times p_d^j.$$

Using the curve fitting function of Scipy (Virtanen et al., 2020), with a trf (Branch, Coleman & Li, 1999) optimization and a Huber loss (Huber, 1992), we optimize the parameters of both models using p_i^j , by minimising the sum of squares, giving us the optimized set of parameters

$$\hat{\theta}_0 = \{\hat{p}_0\} \quad \text{and} \quad \hat{\theta}_1 = \{\hat{p}_d^j, \hat{d}_{pmin}^j, \hat{d}_{pmax}^j\}$$

for \mathcal{M}_0 and \mathcal{M}_1 respectively. Under \mathcal{M}_0 and \mathcal{M}_1 we have the following likelihood functions

$$\mathcal{L}_0(\hat{\theta}_0 | \mathbf{M}^j, \mathbf{N}^j) = \prod_{i=0}^k \binom{M_i^j + N_i^j}{N_i^j} (\mathcal{M}_0 \hat{\pi}^j)^{N_i^j} (1 - \mathcal{M}_0 \hat{\pi}^j)^{M_i^j},$$

$$\mathcal{L}_1(\hat{\theta}_1 | \mathbf{M}^j, \mathbf{N}^j) = \prod_{i=0}^k \binom{M_i^j + N_i^j}{N_i^j} (\mathcal{M}_1 \hat{\pi}_i^j)^{N_i^j} (1 - \mathcal{M}_1 \hat{\pi}_i^{1,j})^{M_i^j},$$

where $\mathcal{M}_0 \hat{\pi}^j$ and $\mathcal{M}_1 \hat{\pi}_i^j$ are calculated using Eqs. (1) and (2). Note that if $d_{pmax}^j = d_{pmin}^j = p_0$, then $\mathcal{M}_0 \pi^j = \mathcal{M}_1 \pi_i^j$ for $i = 0, \dots, k$. Hence to compare the goodness-of-fit for models \mathcal{M}_0 and \mathcal{M}_1 for each reference, we calculate a likelihood-ratio test-statistic of the form

$$\lambda_j = -2 \ln \left[\frac{\mathcal{L}_0(\hat{\theta}_0 | \mathbf{M}^j, \mathbf{N}^j)}{\mathcal{L}_1(\hat{\theta}_1 | \mathbf{M}^j, \mathbf{N}^j)} \right],$$

from which we compute a p -value using the fact that $\lambda_j \sim \chi_2^2$, asymptotically (Neyman & Pearson, 1933). Finally, we adjust the p -values for multiple testing of all references, using the StatsModels (Seabold & Perktold, 2010) implementation of the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995).

RESULTS

Statistical analysis and model selection

To test the performance of PyDamage in recognizing metagenome-assembled contigs with ancient DNA damage, we used the simulated short-read sequencing data aligned against simulated contigs of different lengths. Our method correctly identified contigs as not significantly damaged for simulations with no damage in 100% of cases. However, our model only correctly identified contigs as significantly damaged in 87.71% of cases where the contigs were simulated to have damage. To assess the performance of our method, and to determine the simulation parameters that most affected model accuracy, we analysed the simulated data using logistic regression via the `glm` function as implemented in the `stats` package using R (*R Core Team, 2018*). We included as potential explanatory variables the median read length, the simulated coverage, the simulated contig length, the simulated level of damage, and the GC content of each of the reference contigs, yielding 32 candidate logistic regression models.

We separated the data into two data sets: half of our data was used as ‘fit data’, data for performing model fit and parameter estimation, and the remaining half was reserved as ‘test data’, data that is used to assess model accuracy on data not used in fitting the model ($n = 206,831$ in both cases). Unfortunately, with so many observations in our model, classical model selection methods such as AIC and ANOVA tend to overfit (*Babyak, 2004*). Similarly, we also performed ten-fold down-sampling of the data for each model such that we had equal numbers of damaged and undamaged simulations so as not to bias the predictive model. Hence, for each of the fitted 32 logistic regression models (with $\epsilon = 1 \times 10^{-14}$ and maximum iterations 10^3) we instead report the mean F_1 and Nagelkerke’s R^2 values for each candidate model.

Of the 32 candidate models, four models had both F_1 and R^2 values greater than 0.6 (see [Table 1](#)). Each these four models contained at least the following predictor variables: contig length, mean coverage, and the simulated level of damage. However, the full model with GC content and read length as additional predictor variables had similar F_1 and R^2 values, and so we consider all four models (see [Fig. 2](#)). Because it is possible that there is correlation between some of our predictor variables (i.e., increased levels of simulated damage could lead to a reduced median read length), we then performed a Relative Weights Analysis (RWA) to further estimate predictor variable importance in an uncorrelated setting (*Chan, 2020*). In essence, RWA calculates the proportion of the overall R^2 for the model that can be attributed to each variable. We performed RWA on both the full model and our best performing model. We found that the median read length and GC content accounted for only 0.31% and 2.75% of the R^2 value in the full model respectively. However, we found that contig length, mean coverage and the simulated level of damage all accounted for approximately one third of the R^2 value in our best performing model, indicating that these are the predictor variables of importance.

Our final logistic regression model identified mean coverage, the level of damage, and the contig length as significant predictor variables for model accuracy. Each of these variables had positive coefficients, meaning that an increase in damage, genome coverage, or contig

Table 1 The F_1 score and Nagelkerke's R^2 mean values for the top ten models (ranked by F_1). The model we retained is highlighted in bold.

Variables	F_1	R^2
readlength	0.791	0.001
GCcontent/readlength	0.642	0.005
damage/contiglength/coverage	0.624	0.607
damage/contiglength/readlength/coverage	0.624	0.610
damage/contiglength/GCcontent/readlength/coverage	0.623	0.619
damage/contiglength/GCcontent/coverage	0.622	0.618
damage/contiglength	0.600	0.432
damage/contiglength/readlength	0.593	0.434
damage/coverage	0.592	0.385
damage/readlength/coverage	0.588	0.387

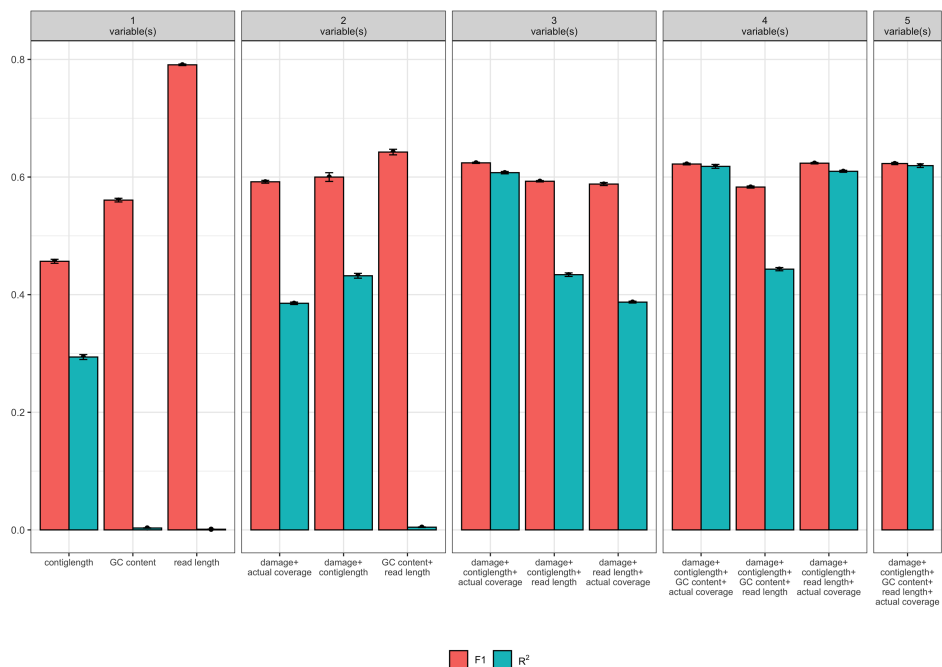


Figure 2 Measures of model fit calculated on the test data for the top 3 models with one, two, three, four, and five variables, where red is the F_1 score and blue is Nagelkerke's R^2 . Error bars indicate two standard deviations calculated from ten-fold cross validation.

Full-size DOI: 10.7717/peerj.11845/fig-2

length all lead to improved model accuracy. Each variable contributed about one third weight to the R^2 value in the model, indicating roughly equal importance in the accuracy of PyDamage. We integrated the best logistic regression model in PyDamage, with the StatsModels (Seabold & Perktold, 2010) implementation of GLM to provide an estimation of PyDamage ancient contig prediction accuracy given the amount of damage, coverage, and length for each reference (Fig. 3), and found these predictions to adequately match the observed model accuracy for our simulated data set (Fig. 4).

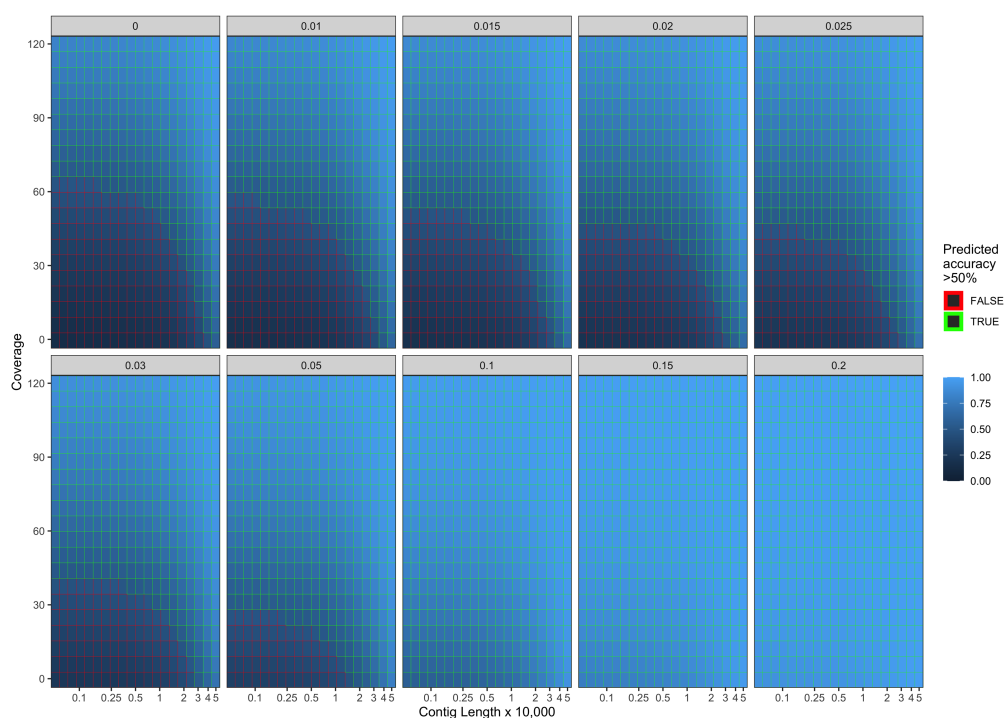


Figure 3 Predicted model accuracy of simulated data. The grey title box above each panel is the simulated damage frequency on the 5' end. Light blue indicates improved model accuracy, with parameter combinations resulting in better than 50% accuracy are outlined in green.

Full-size [DOI: 10.7717/peerj.11845/fig-3](https://doi.org/10.7717/peerj.11845/fig-3)

Application of PyDamage to archeological samples

To test PyDamage on empirical data, we assembled metagenomic data from the paleofeces sample ZSM028 with the metaSPAdes *de novo* assembler. We obtained a total of 359,807 contigs, with an N50 of 429 bp. Such assemblies, consisting of a large number of relatively short contigs, are typical for *de novo* assembled aDNA datasets (Wibowo *et al.*, 2021). After removing sequences shorter than 1,000 bp, 17,103 contigs were left. PyDamage (revision 099fd34) was able to perform a damage estimation for 99.75% of these contigs (17,061 contigs). Because the ZSM028 sequencing library was not treated with uracil-DNA-glycosylase (Rohland *et al.*, 2015), nor amplified with a damage suppressing DNA polymerase, we expect a relatively shallow DNA damage decay curve, and thus filtered for this using the p_d^j parameter. We chose a prediction accuracy threshold of 0.67 after locating the knee point on Fig. 5 with the kneedle method (Satopaa *et al.*, 2011). After filtering PyDamage results with a q -value ≤ 0.05 , $p_d^j \leq 0.6$, and $prediction\ accuracy \geq 0.67$, 1,944 contigs remain. The 5' damage for these contigs ranges from 4.0% to 45.1% with a mean of 14.3% (Fig. 6). Their coverage spans 6.1X to 1,579.8X with a mean of 65.6X, while their length ranges from 1,002 bp to 90,306 bp with a mean of 5,212 bp and an N50 of 10,805 bp.

The Kraken2 taxonomic profile of the microbial contigs identified by PyDamage identified as ancient (Fig. 7) is consistent with bacteria known to be members of the

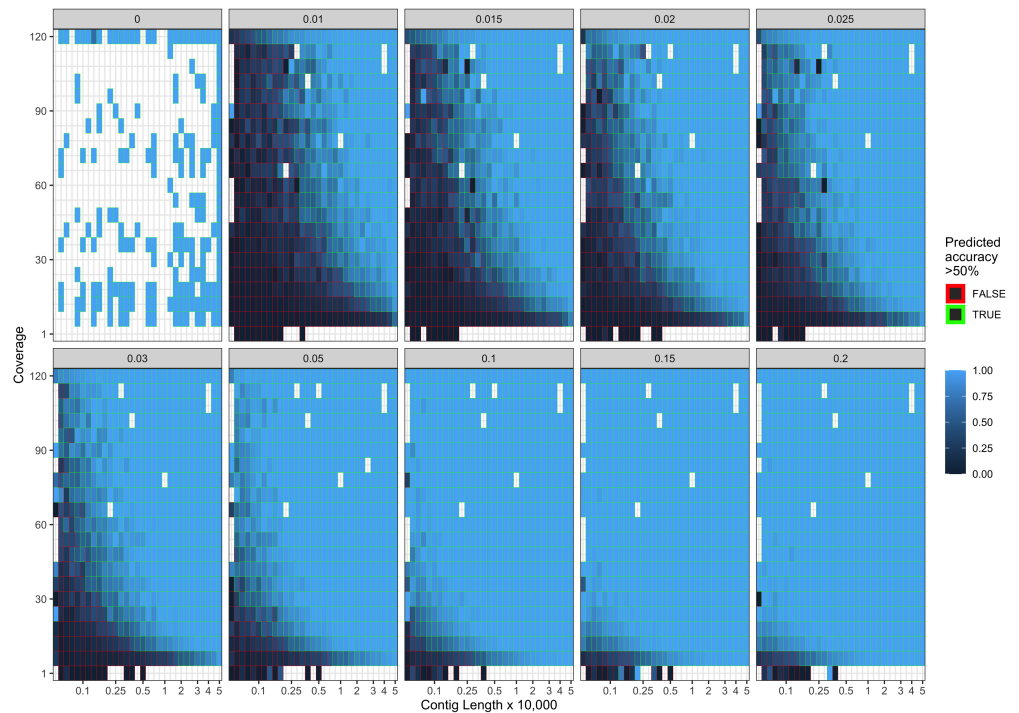


Figure 4 Observed model accuracy of simulated data. The grey title box above each panel is the simulated damage frequency on the 5' end. Light blue indicates improved model accuracy, with parameter combinations resulting in better than 50% accuracy are outlined with green lines. White tiles represent parameter combinations that were not sampled.

Full-size DOI: 10.7717/peerj.11845/fig-4

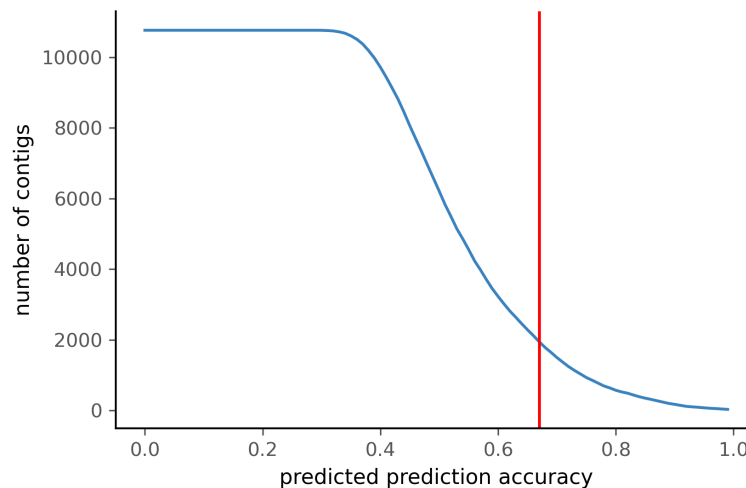


Figure 5 Number of ZSM028 contigs filtered by PyDamage with a q -value ≤ 0.05 as a function of the predicted prediction accuracy. In total, 12,271 of the 17,061 contigs were assigned q -value ≤ 0.05 . The red vertical line is the predicted accuracy threshold of 0.67.

Full-size DOI: 10.7717/peerj.11845/fig-5

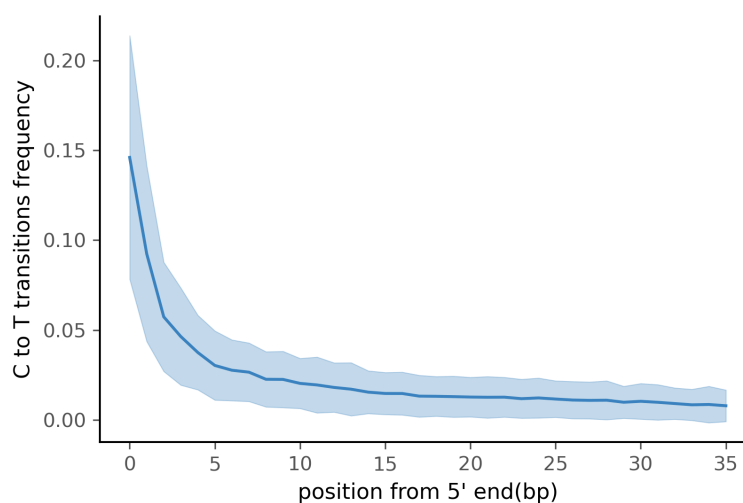


Figure 6 Damage profile of PyDamage filtered contigs of ZSM028. The center line is the mean, the shaded area is \pm one standard-deviation around the mean.

Full-size  DOI: [10.7717/peerj.11845/fig-6](https://doi.org/10.7717/peerj.11845/fig-6)

human gut microbiome, including *Prevotella* (239 contigs), *Treponema* (166 contigs), *Bacteroides* (103 contigs), *Lachnospiraceae* (119 contigs) *Blautia* (36 contigs), *Ruminococcus* (25 contigs), *Phocaeicola* (18 contigs) and *Romboutsia* (16 contigs) ([Schnorr et al., 2016](#); [Pasolli et al., 2019](#); [Singh et al., 2017](#)), as well as taxonomic groups known to be involved in initial decomposition, such as *Clostridium* (145 contigs) ([Hyde et al., 2017](#); [Harrison et al., 2020](#); [Dash & Das, 2020](#)). In addition, eukaryotic contigs were assigned to humans (18 contigs), and to the plant families Fabaceae (18 contigs) and Solanaceae (18 contigs), two families of economically important crops in the Americas that include beans, tomatoes, chile peppers, and tobacco. The remaining contigs were almost entirely assigned to higher taxonomic levels within the important gut microbiome phyla Bacteroidetes, Firmicutes, Proteobacteria, and Spirochaetes, as well as to the Streptophyta phylum of vascular plants. Collectively, these five phyla accounted for 1,283 of to 1,494 contigs that could be taxonomically assigned.

Functional annotation of the authenticated ancient contigs using Prokka was successful for 1,901 of 1,944 contigs. Among these, multiple genes of functional interest were identified, including contigs annotated as encoding the multidrug resistance proteins MdtA, MdtB, and MdtC, which convey, among other functions, bile salt resistance ([Nagakubo et al., 2002](#)) (Table 2). Kraken2 taxonomic profiling of these three contigs yields a taxonomic assignment to the gut spirochaete *Treponema succinifaciens*, a species absent in the gut microbiome of industrialized populations, but which is found globally in societies practicing traditional forms of subsistence ([Obregon-Tito et al., 2015](#); [Schnorr et al., 2014](#)). Other authenticated contigs contained genes associated with resistance to the natural antimicrobial compounds fosmidomycin, colistin, daunorubicin/doxorubicin, tetracycline, polymyxin, and linearmycin. A growing body of evidence supports an ancient

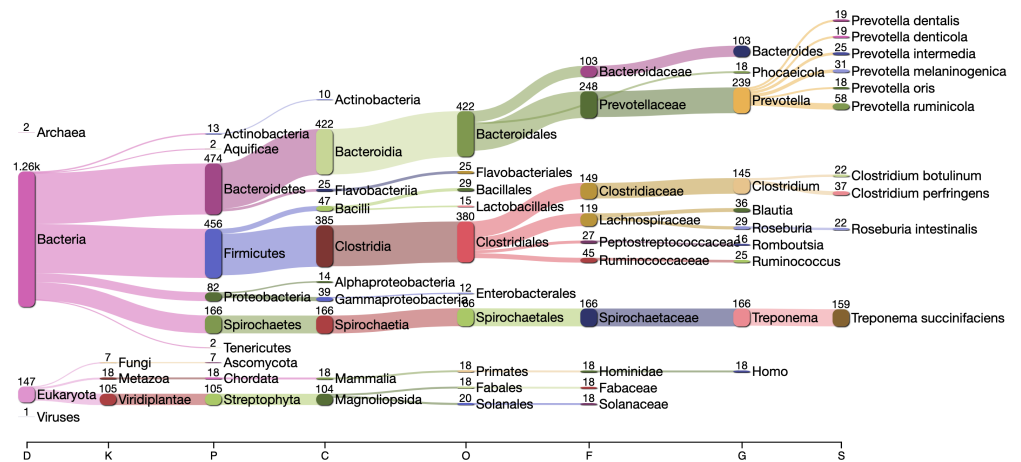


Figure 7 Taxonomic assignment by Kraken2 of the contigs filtered by PyDamage with q -value ≤ 0.05 , $p_d^j \leq 0.6$, and prediction accuracy ≥ 0.67 .

Full-size DOI: 10.7717/peerj.11845/fig-7

origin for resistance to most classes of natural antibiotics (D'Costa *et al.*, 2011; Warinner *et al.*, 2014; Christaki, Marcou & Tofarides, 2020; Wibowo *et al.*, 2021).

DISCUSSION

De novo sequence assembly is increasingly being applied to ancient metagenomic data in order to improve lower rank taxonomic assignment and to enable functional profiling of ancient bacterial communities. The ability to reconstruct reference-free ancient genes, gene complexes, or even genomes opens the door to exploring microbial evolutionary histories and past functional diversity that may be underrepresented or absent in present-day microbial communities. A critical step in reconstructing this past diversity, however, is being able to distinguish DNA of ancient and modern origin (Warinner *et al.*, 2017). Characteristic forms of damage that accumulate in DNA over time, such as DNA fragmentation and cytosine deamination, are widely used to authenticate aDNA (Orlando *et al.*, 2021) and have been important, for example, in enabling the reconstruction of the Neanderthal genome from skeletal remains contaminated with varying levels of modern human DNA (Briggs *et al.*, 2007; Bokelmann *et al.*, 2019; Peyr gne *et al.*, 2019).

Nevertheless, applying such an approach to complex ancient microbial communities, such as archaeological microbiome samples or sediments, is more challenging. Existing microbial reference sequences in databases such as NCBI RefSeq have been found to be insufficiently representative of modern microbial diversity (Pasolli *et al.*, 2019; Manara *et al.*, 2019), let alone ancient diversity, making reference-free *de novo* assembly particularly desirable for both modern and ancient microbial metagenomics. However, *de novo* assembly of aDNA has always been a challenge due to its highly fragmented nature. While tools have been designed to improve the assembly of ancient metagenomics data (Seitz & Nieselt, 2017), assessing the damage carried by the assembled contigs has remained an open problem.

Table 2 Contigs assembled by metaSPAdes, identified by PyDamage as carrying damage, and annotated as carrying resistance genes by Prokka.

Contig name	Contig length (bp)	Coverage	Product
NODE_2446	3232	64.3	Arsenical-resistance protein Acr3
NODE_45	28638	26.0	Bifunctional polymyxin resistance protein ArnA
NODE_832	6259	46.3	Cobalt-zinc-cadmium resistance protein CzcA
NODE_832	6259	46.3	Cobalt-zinc-cadmium resistance protein CzcB
NODE_2661	3058	91.5	Colistin resistance protein EmrA
NODE_2661	3058	91.5	Colistin resistance protein EmrA
NODE_215	13020	27.0	Daunorubicin/doxorubicin resistance ATP-binding protein DrrA
NODE_136	16294	26.0	Daunorubicin/doxorubicin resistance ATP-binding protein DrrA
NODE_1676	4090	81.3	Fosmidomycin resistance protein
NODE_8410	1542	77.3	Linearmycin resistance ATP-binding protein LnrL
NODE_29	35207	27.8	Multidrug resistance ABC transporter ATP-binding and permease protein
NODE_232	12485	31.9	Multidrug resistance protein MdtA
NODE_97	19553	27.4	Multidrug resistance protein MdtA
NODE_12	45672	45.6	Multidrug resistance protein MdtA
NODE_10	46280	59.8	Multidrug resistance protein MdtA
NODE_97	19553	27.4	Multidrug resistance protein MdtB
NODE_97	19553	27.4	Multidrug resistance protein MdtB
NODE_12	45672	45.6	Multidrug resistance protein MdtC
NODE_10	46280	59.8	Multidrug resistance protein MdtC
NODE_232	12485	31.9	Multidrug resistance protein MdtC
NODE_17	41269	29.9	Multidrug resistance protein MdtK
NODE_465	8695	37.5	Tetracycline resistance protein TetO
NODE_204	13262	44.9	Tetracycline resistance protein, class C

Existing tools such as HOPS (Hübler *et al.*, 2019) and mapDamage2 (Jónsson *et al.*, 2013) are readily available programs used to investigate ancient DNA deamination damage. However they perform a very different analysis compared to PyDamage, and their method, scope of application, and ability to scale are dissimilar to PyDamage. While HOPS applies additional authentication criteria beyond deamination (e.g., edit distance), when it comes to C → T substitutions, it uses a simpler heuristic to segregate damaged from non-damaged sequences, and does not provide any statistical testing of the damage. Furthermore, it is intended to be used in a targeted manner, with the user having to specify a list of known organisms of interest to look for, which is incompatible with *de novo* assembly that can potentially recover previously unknown species. Regarding mapDamage2, while it provides a statistical framework for aDNA damage modelling that can be used to rescale alignment quality score, it is not intended to be used in a metagenomics context. MapDamage2 does not provide a statistical test of the damage, and it is only designed to be used for single

genomes, which poses scalability issues when using it with thousands of references typically generated by metagenomics *de novo* assembly.

Here, we have presented PyDamage as a tool to rapidly assess aDNA damage patterns for numerous reference sequences in parallel, allowing fast damage profiling of metagenome assembled contigs. To evaluate the performance of PyDamage model fitting and statistical testing, we benchmarked the tool using simulated assembly data of known coverage, length, GC content, read length, and damage level. Because PyDamage predicts based on C → T transition frequency, we originally expected GC-content to impact the available number of possible C → T transitions, and hence influence the predictions of PyDamage. However we found that GC content and read length were not a major driver of the accuracy of PyDamage's predictions, but contig length, coverage, and damage level each played major roles. Taken together, this three parameter combination greatly influenced the ability of PyDamage to make accurate damage assessments for a given contig. Overall, PyDamage has highly reliable damage prediction accuracy for contigs with high coverage, long lengths, and high damage, but the tool's power to assess damage is reduced for lower coverage, shorter contigs length, and lower deamination damaged contigs. Although aDNA damage levels (cytosine deamination and fragmentation) are features of the DNA itself and out of the researcher's control, we show that researchers can generally improve model accuracy through deeper sequencing.

When comparing the parameter range of our simulated data to real world *de novo* assembly data, we find that some of PyDamage prediction accuracy limitations are mitigated by the assembly process itself: *de novo* assemblers usually need a minimum of approximately 5X coverage to assemble contigs (Fig. 8) (Wibowo *et al.*, 2021), and it is common practice to discard short contigs (<1000 bp) before further processing steps in a classical metagenomic *de novo* assembly analysis process. Nevertheless, low coverage, low damage, short contigs will remain a marginal challenge for damage characterization, even with further manual inspection. For example, for a 10,000 bp *de novo* assembled contig with 5% damage, PyDamage will only start to make reliable predictions once a coverage of 12X is reached (Fig. 3, interactive app available at <https://maxibor.shinyapps.io/pydamageglm>). For a similar contig with 10% damage, model accuracy is high even from 1X coverage. Overall, we find that PyDamage generally performs well on ancient metagenomic data with > 5% damage, but contig length and coverage are also essential factors in determining the model accuracy for a given contig.

Although we used the kneedle method (Satopaa *et al.*, 2011) to select the prediction accuracy threshold for paleofeces sample ZSM028, users can adjust the selected prediction accuracy threshold according to the needs of their research question. For example, for some research questions where high accuracy in verifying damage is paramount, more stringent thresholds can be applied to minimize false positives, even though this increases false negatives. For other questions and where additional authentication criteria are available (such as taxonomic information or metagenomic bins), lower thresholds may be applied to reduce the number of false negatives due to insufficient coverage or contig length.

PyDamage is designed to estimate accumulated DNA damage in *de novo* assembled metagenomic sequences. However, although DNA damage can be used to authenticate

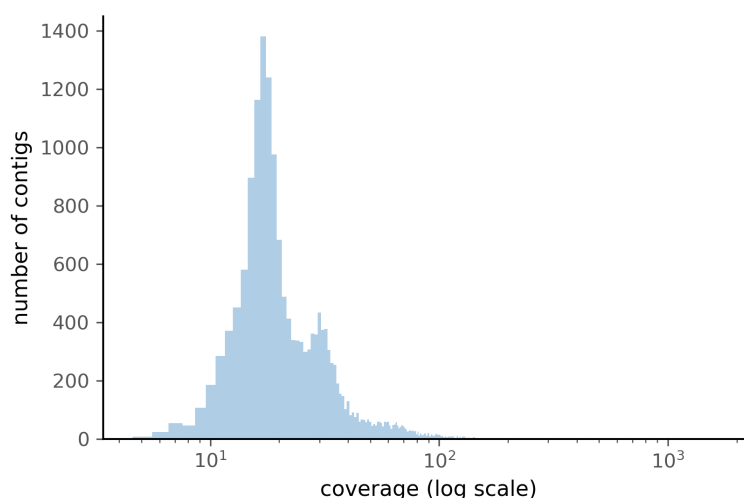


Figure 8 Distribution of the coverage for ZSM028 contigs > 1,000 bp assembled by metaSPAdes.

Full-size [DOI: 10.7717/peerj.11845/fig-8](https://doi.org/10.7717/peerj.11845/fig-8)

DNA as ancient, it is important to note that it is not necessarily an indicator of *intra vitam* endogeneity. DNA within ancient remains typically consists of both an endogenous fraction present during life and an exogenous fraction accumulated after death. For skeletal remains, the endogenous fraction typically consists of host DNA, as well as possibly pathogen DNA if the host was infected at the time of death. For paleofeces or dental calculus, the endogenous fraction typically consists of microbiome DNA, as well as trace amounts of host, parasite, and dietary DNA. In both cases, the endogenous fraction of DNA is expected to carry DNA damage accumulated since the death (skeletal remains, dental calculus) or defecation (paleofeces) of the individual. Within the exogenous fraction, however, the DNA may span a range of ages. Nearly all ancient remains undergo some degree of degradation and decomposition, during which either endogenous (thanatomicrobiome) or exogenous (necrobiome) bacteria invade the remains and grow (Hyde *et al.*, 2017; Harrison *et al.*, 2020; Dash & Das, 2020). DNA from bacteria that participated early in this process (shortly after death or defecation), will carry similar levels of damage as the endogenous DNA because they are of similar age. In contrast, more recent necrobiome activity will carry progressively less age-related damage, and very recent sources of contamination from excavation, storage, curation, and laboratory handling are expected to carry little to no DNA damage.

To demonstrate the utility of PyDamage on ancient metagenomic data, we applied PyDamage to paleofeces ZSM028, a ca. 1,300-year-old specimen of feces from a dry rockshelter site in Mexico that was previously shown to have excellent preservation of endogenous gut microbiome DNA and low levels of environmental contamination (Borrry *et al.*, 2020). Using PyDamage, we assessed the damage profiles of contigs with lengths >1,000 bp, and authenticated nearly 2,000 contigs as carrying damage patterns consistent with ancient DNA. The overwhelming majority of these contigs were consistent with bacterial members of the human gut microbiome, as well as expected host and dietary

components, but a small fraction of authenticated contigs were assigned to environmental bacteria and fungi, including the exogenous soil bacteria *Clostridium botulinum* (22 contigs) and *Clostridium perfringens* (38 contigs). These taxa are known to be important early decomposers in the necrobiome (Harrison *et al.*, 2020), and the damage they carry suggests that they likely participated in the early degradation of the paleofeces before decomposition was arrested by the extreme aridity of the rockshelter.

Among the PyDamage authenticated contigs assigned to gut-associated taxa, NODE_10, NODE_12, and NODE_97 are of particular interest. These contigs encode a multidrug resistant ABC (MdtABC) transporter associated with bile salt resistance in the bacterium *T. succinifaciens*. *T. succinifaciens* is a human-associated gut species that is today only found in the gut microbiomes of individuals engaging in traditional forms of dietary subsistence (Obregon-Tito *et al.*, 2015; Schnorr *et al.*, 2014; Angelakis *et al.*, 2019). It is not found in the gut microbiomes of members of industrialized societies, and is believed extinct in these groups (Schnorr *et al.*, 2016). Its identification within paleofeces provides insights into the evolutionary history of this enigmatic microorganism and its functional adaptation to the human gut (Schnorr *et al.*, 2019). The additional identification of other resistance genes among the authenticated contigs provides further evidence regarding the evolution of antimicrobial resistance in human-associated microbes.

CONCLUSION

As the fields of microbiology and evolutionary biology increasingly turn to the archaeological record to investigate the rich and dynamic evolutionary history of ancient microbial communities, it has become vital to develop tools for assembling and authenticating ancient metagenomic DNA. Coupled with aDNA *de novo* assembly, PyDamage opens up new doors to explore and understand the functional diversity of ancient metagenomes.

ACKNOWLEDGEMENTS

We thank Nigel Bean and Jonathon Tuke for extremely useful discussions.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Alexander Hübner was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2051 –Project-ID 390713860). Adam B Rohrlach was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under grant agreement no. 771234 –PALEORIDER. Maxime Borry and Christina Warinner were funded by the Werner Siemens Foundation ("Paleobiotechnology"). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

DFG, German Research Foundation: 390713860.

European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program: 771234 –PALEoRIDER.

Werner Siemens Foundation (“Paleobiotechnology”).

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Maxime Borry, Alexander Hübner and Adam B. Rohrlach conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Christina Warinner analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The genetic data for ZSM028 is available on the European Nucleotide Archive (ENA): [PRJEB33577](https://ena.ebi.ac.uk/ena/record/PRJEB33577).

The PyDamage Software and source code available from Github: <https://github.com/maxibor/pydamage>, license: GPLv3.

The code to replicate the simulation of reads and contigs, and the figures is available at GitHub: DOI: <https://doi.org/10.5281/zenodo.4981768>.

REFERENCES

- Angelakis E, Bachar D, Yasir M, Musso D, Djossou F, Gaborit B, Brah S, Diallo A, Ndombe GM, Mediannikov O, Robert C, Azhar EI, Bibi F, Nsana NS, Parra HJ, Akiana J, Sokhna C, Davoust B, Dutour A, Raoult D. 2019. Treponema species enrich the gut microbiota of traditional rural populations but are absent from urban individuals. *New Microbes and New Infections* 27:14–21 DOI [10.1016/j.nmni.2018.10.009](https://doi.org/10.1016/j.nmni.2018.10.009).
- Babjak MA. 2004. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine* 66(3):411–421 DOI [10.1097/01.psy.0000127692.23278.a9](https://doi.org/10.1097/01.psy.0000127692.23278.a9).
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1):289–300 DOI [10.1111/j.2517-6161.1995.tb02031.x](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x).
- Bokelmann L, Hajdinjak M, Peyrégne S, Brace S, Essel E, De Filippo C, Glocke I, Grote S, Mafessoni F, Nagel S, Kelso J, Prüfer K, Vernot B, Barnes I, Pääbo S, Meyer M, Stringer C. 2019. A genetic analysis of the Gibraltar Neanderthals. *Proceedings of the National Academy of Sciences of the United States of America* 116(31):15610–15615 DOI [10.1073/pnas.1903984116](https://doi.org/10.1073/pnas.1903984116).

- Borry M, Cordova B, Perri A, Wibowo M, Honap TP, Ko J, Yu J, Britton K, Girdland-Flink L, Power RC, Stuijts I, Salazar-García DC, Hofman C, Hagan R, Kagon TS, Meda N, Carabin H, Jacobson D, Reinhard K, Lewis C, Kostic A, Jeong C, Herbig A, Hübner A, Warinner C. 2020. CoproID predicts the source of coprolites and paleofeces using microbiome composition and host DNA content. *PeerJ* 8:e9001 DOI 10.7717/peerj.9001.
- Branch MA, Coleman TF, Li Y. 1999. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal on Scientific Computing* 21(1):1–23 DOI 10.1137/S1064827595289108.
- Brealey JC, Leito HG, Van der Valk T, Xu W, Bougiouri K, Daln L, Guschanski K. 2020. Dental Calculus as a Tool to Study the Evolution of the Mammalian Oral Microbiome. *Molecular Biology and Evolution* 37(10):3003–3022 DOI 10.1093/molbev/msaa135.
- Breitwieser FP, Salzberg SL. 2016. Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. *BioRxiv*. 084715 DOI 10.1101/084715.
- Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America* 104(37):14616–14621 DOI 10.1073/pnas.0704665104.
- Chan M. 2020. rwa: perform a relative weights analysis. R package version 0.0.3. Available at <https://CRAN.R-project.org/package=rwa>.
- Christaki E, Marcou M, Tofarides A. 2020. Antimicrobial resistance in bacteria: mechanisms, evolution, and persistence. *Journal of Molecular Evolution* 88(1):26–40 DOI 10.1007/s00239-019-09914-3.
- Compeau PE, Pevzner PA, Tesler G. 2011. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology* 29(11):987–991 DOI 10.1038/nbt.2023.
- Dabney J, Meyer M, Pääbo S. 2013. Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology* 5(7):a012567 DOI 10.1101/cshperspect.a012567.
- Dash HR, Das S. 2020. Thanatomicrobiome and epinecrotic community signatures for estimation of post-mortem time interval in human cadaver. *Applied Microbiology and Biotechnology* 104:9497–9512 DOI 10.1007/s00253-020-10922-3.
- D’Costa VM, King CE, Kalan L, Morar M, Sung WWL, Schwarz C, Froese D, Zazula G, Calmels F, Debruyne R, Golding GB, Poinar HN, Wright GD. 2011. Antibiotic resistance is ancient. *Nature* 477(7365):457–461 DOI 10.1038/nature10388.
- Ginolhac A, Rasmussen M, Gilbert MTP, Willerslev E, Orlando L. 2011. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics (Oxford, England)* 27(15):2153–2155 DOI 10.1093/bioinformatics/btr347.
- Hansen HB, Damgaard PB, Margaryan A, Stenderup J, Lynnerup N, Willerslev E, Allentoft ME. 2017. Comparing ancient DNA preservation in petrous bone and tooth cementum. *PLOS ONE* 12(1):e0170940 DOI 10.1371/journal.pone.0170940.
- Harrison L, Kooienga E, Speights C, Tomberlin J, Lashley M, Barton B, Jordan H. 2020. Microbial succession from a subsequent secondary death event following mass mortality. *BMC Microbiology* 20(1):1–11 DOI 10.1186/s12866-020-01969-3.

- Hofreiter M, Jaenicke V, Serre D, Haeseler AV, Pääbo S. 2001.** DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research* **29(23)**:4793–4799 DOI [10.1093/nar/29.23.4793](https://doi.org/10.1093/nar/29.23.4793).
- Huber PJ. 1992.** Robust estimation of a location parameter. In: *Breakthroughs in statistics*. New York: Springer, 492–518 DOI [10.1007/978-1-4612-4380-9_35](https://doi.org/10.1007/978-1-4612-4380-9_35).
- Hübler R, Key FM, Warinner C, Bos KI, Krause J, Herbig A. 2019.** HOPS: automated detection and authentication of pathogen DNA in archaeological remains. *Genome Biology* **20(1)**:1–13 DOI [10.1186/s13059-019-1903-0](https://doi.org/10.1186/s13059-019-1903-0).
- Hyde ER, Metcalf JL, Bucheli SR, Lynne AM, Knight R. 2017.** Microbial communities associated with decomposing corpses. In: *Forensic Microbiology, Wiley Online Books*. the Atrium, Southern Gate, Chichester, West Sussex, UK: John Wiley & Sons Ltd., 245–273 DOI [10.1002/9781119062585.ch10](https://doi.org/10.1002/9781119062585.ch10).
- Jónsson H, Ginolhac A, Schubert M, Johnson P. L. F., Orlando L. 2013.** mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics (Oxford, England)* **29(13)**:1682–1684 DOI [10.1093/bioinformatics/btt193](https://doi.org/10.1093/bioinformatics/btt193).
- Kang DD, Froula J, Egan R, Wang Z. 2015.** MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**:e1165 DOI [10.7717/peerj.1165](https://doi.org/10.7717/peerj.1165).
- Langmead B, Salzberg SL. 2012.** Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9(4)**:357–359 DOI [10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923).
- Li H, Durbin R. 2009.** Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25(14)**:1754–1760 DOI [10.1093/bioinformatics/btp324](https://doi.org/10.1093/bioinformatics/btp324).
- Manara S, Asnicar F, Beghini F, Bazzani D, Cumbo F, Zolfo M, Nigro E, Karcher N, Manghi P, Metzger MI, Pasolli E, Segata N. 2019.** Microbial genomes from non-human primate gut metagenomes expand the primate-associated bacterial tree of life with over 1000 novel species. *Genome Biology* **20(1)**:299 DOI [10.1186/s13059-019-1923-9](https://doi.org/10.1186/s13059-019-1923-9).
- Mann AE, Sabin S, Ziesemer K, Vagene AJ, Schroeder H, Ozga AT, Sankaranarayanan K, Hofman CA, Fellows Yates JA, Salazar-García DC, Frohlich B, Aldenderfer M, Hoogland M, Read C, Milner GR, Stone AC, Lewis CM, Krause J, Hofman C, Bos KI, Warinner C. 2018.** Differential preservation of endogenous human and microbial DNA in dental calculus and dentin. *Scientific Reports* **8(1)**:9822 DOI [10.1038/s41598-018-28091-9](https://doi.org/10.1038/s41598-018-28091-9).
- Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, Filippo C d, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andrs AM, Eichler EE, Slatkin M, Reich D, Kelso J, Pääbo S. 2012.** A high-coverage genome sequence from an archaic denisovan individual. *Science* **338(6104)**:222–226 DOI [10.1126/science.1224344](https://doi.org/10.1126/science.1224344).

- Nagakubo S, Nishino K, Hirata T, Yamaguchi A. 2002.** The putative response regulator BaeR stimulates multidrug resistance of *Escherichia coli* via a novel multidrug exporter system, MdtABC. *Journal of Bacteriology* **184**(15):4161–4167 DOI [10.1128/JB.184.15.4161-4167.2002](https://doi.org/10.1128/JB.184.15.4161-4167.2002).
- Neukamm J, Peltzer A, Nieselt K. 2020.** DamageProfiler: fast damage pattern calculation for ancient DNA. *BioRxiv*. DOI [10.1101/2020.10.01.322206](https://doi.org/10.1101/2020.10.01.322206).
- Neyman J, Pearson ES. 1933.** IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical Or Physical Character* **231**(694-706):289–337 DOI [10.1098/rsta.1933.0009](https://doi.org/10.1098/rsta.1933.0009).
- Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017.** metaSPAdes: a new versatile metagenomic assembler. *Genome Research* **27**(5):824–834 DOI [10.1101/gr.213959.116](https://doi.org/10.1101/gr.213959.116).
- Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, Zech Xu Z, Van Treuren W, Knight R, Gaffney PM, Spicer P, Lawson P, Marin-Reyes L, Trujillo-Villaruel O, Foster M, Guija-Poma E, Troncoso-Corzo L, Warinner C, Ozga AT, Lewis CM. 2015.** Subsistence strategies in traditional societies distinguish gut microbiomes. *Nature Communications* **6**(1):6505 DOI [10.1038/ncomms7505](https://doi.org/10.1038/ncomms7505).
- Orlando L, Allaby R, Skoglund P, Der Sarkissian C, Stockhammer PW, Ávila Arcos MC, Fu Q, Krause J, Willerslev E, Stone AC, Warinner C. 2021.** Ancient DNA analysis. *Nature Reviews Methods Primers* **1**(1):1–26 DOI [10.1038/s43586-020-00011-0](https://doi.org/10.1038/s43586-020-00011-0).
- Pasoli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. 2019.** Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**(3):649–662 DOI [10.1016/j.cell.2019.01.001](https://doi.org/10.1016/j.cell.2019.01.001).
- Peyrégne S, Slon V, Mafessoni F, Filippo CD, Hajdinjak M, Nagel S, Nickel B, Essel E, Cabec AL, Wehrberger K, Conard NJ, Kind CJ, Posth C, Krause J, Abrams G, Bonjean D, Modica KD, Toussaint M, Kelso J, Meyer M, Pääbo S, Prüfer K. 2019.** Nuclear DNA from two early Neandertals reveals 80,000 years of genetic continuity in Europe. *Science Advances* **5**(6):eaaw5873 DOI [10.1126/sciadv.aaw5873](https://doi.org/10.1126/sciadv.aaw5873).
- pysam developers. 2018.** Pysam: a python module for reading and manipulating files in the SAM/BAM format. Available at <https://github.com/pysam-developers/pysam> DOI [10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352).
- R Core Team. 2018.** R: A Language and Environment for Statistical Computing. Vienna, Austria: Available at <https://www.R-project.org/>.
- Renaud G, Hanghøj K, Willerslev E, Orlando L. 2017.** gargammel: a sequence simulator for ancient DNA. *Bioinformatics* **33**(4):577–579 DOI [10.1093/bioinformatics/btw670](https://doi.org/10.1093/bioinformatics/btw670).
- Rho M, Tang H, Ye Y. 2010.** FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research* **38**(20):e191–e191 DOI [10.1093/nar/gkq747](https://doi.org/10.1093/nar/gkq747).
- Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D. 2015.** Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**(1660):20130624 DOI [10.1098/rstb.2013.0624](https://doi.org/10.1098/rstb.2013.0624).

- Satopaa V, Albrecht J, Irwin D, Raghavan B. 2011. Finding a “Kneedle” in a Haystack: detecting knee points in system behavior. In: *2011 31st international conference on distributed computing systems workshops*. Minneapolis: IEEE, 166–171 DOI [10.1109/ICDCSW.2011.20978-1-4577-0384-3](https://doi.org/10.1109/ICDCSW.2011.20978-1-4577-0384-3).
- Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, Turroni S, Biagi E, Peano C, Severgnini M, Fiori J, Gotti R, De Bellis G, Luiselli D, Brigidi P, Mabulla A, Marlowe F, Henry AG, Crittenden AN. 2014. Gut microbiome of the Hadza hunter-gatherers. *Nature Communications* 5(1):3654 DOI [10.1038/ncomms4654](https://doi.org/10.1038/ncomms4654).
- Schnorr SL, Hofman CA, Netshifhefhe S. R., Duncan FD, Honap TP, Lesnik J., Lewis CM. 2019. Taxonomic features and comparisons of the gut microbiome from two edible fungus-farming termites (*Macrotermes falciger*; *M. natalensis*) harvested in the Vhembe district of Limpopo, South Africa. *BMC Microbiology* 19(1):1–22 DOI [10.1186/s12866-019-1540-5](https://doi.org/10.1186/s12866-019-1540-5).
- Schnorr SL, Sankaranarayanan K, Lewis Jr CM, Warinner C. 2016. Insights into human evolution from ancient and contemporary microbiome studies. *Current Opinion in Genetics & Development* 41:14–26 DOI [10.1016/j.gde.2016.07.003](https://doi.org/10.1016/j.gde.2016.07.003).
- Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes* 9:88 DOI [10.1186/s13104-016-1900-2](https://doi.org/10.1186/s13104-016-1900-2).
- Seabold S, Perktold J. 2010. Statsmodels: Econometric and statistical modeling with python. In: *9th Python in science conference*. DOI [10.25080/Majora-92bf1922-011](https://doi.org/10.25080/Majora-92bf1922-011).
- Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)* 30(14):2068–2069 DOI [10.1093/bioinformatics/btu153](https://doi.org/10.1093/bioinformatics/btu153).
- Seitz A, Nieselt K. 2017. Improving ancient DNA genome assembly. *PeerJ* 5:e3126 DOI [10.7717/peerj.3126](https://doi.org/10.7717/peerj.3126).
- Singh RK, Chang H-W, Yan D, Lee KM, Ucmak D, Wong K, Abrouk M, Farahnik B, Nakamura M, Zhu TH, Bhutani T, Liao W. 2017. Influence of diet on the gut microbiome and implications for human health. *Journal of Translational Medicine* 15(1):73 DOI [10.1186/s12967-017-1175-y](https://doi.org/10.1186/s12967-017-1175-y).
- Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Pääbo S, Krause J, Jakobsson M. 2014. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences of the United States of America* 111(6):2229–2234 DOI [10.1073/pnas.1318934111](https://doi.org/10.1073/pnas.1318934111).
- Tett A, Huang KD, Asnicar F, Fehlner-Peach H, Pasolli E, Karcher N, Armanini F, Manghi P, Bonham K, Zolfo M, Filippis FD, Magnabosco C, Bonneau R, Lusingu J, Amuasi J, Reinhard K, Rattei T, Boulund F, Engstrand L, Zink A, Collado MC, Littman DR, Eibach D, Ercolini D, Rota-Stabelli O, Huttenhower C, Maixner F., Segata N. 2019. The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. *Cell Host & Microbe* 26(5):666–679.e7 DOI [10.1016/j.chom.2019.08.018](https://doi.org/10.1016/j.chom.2019.08.018).
- Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson

- J, Jarrod Millman K, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey C, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, Contributors S. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* 17:261–272 DOI [10.1038/s41592-019-0686-2](https://doi.org/10.1038/s41592-019-0686-2).
- Warinner C, Herbig A, Mann A, Fellows Yates JA, Weiß CL, Burbano HA, Orlando L, Krause J. 2017. A robust framework for microbial archaeology. *Annual Review of Genomics and Human Genetics* 18:321–356 DOI [10.1146/annurev-genom-091416-035526](https://doi.org/10.1146/annurev-genom-091416-035526).
- Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, Grossmann J, Radini A, Hancock Y, Tito RY, Fiddyment S, Speller C, Hendy J, Charlton S, Luder HU, Salazar-García DC, Eppler E, Seiler R, Hansen LH, Castruita JAS, Barkow-Oesterreicher S, Teoh KY, Kelstrup CD, Olsen JV, Nanni P, Kawai T, Willerslev E, Von Mering C, Lewis CM, Collins MJ, Gilbert MTP, Rühli F, Cappellini E. 2014. Pathogens and host immunity in the ancient human oral cavity. *Nature Genetics* 46(4):336–344 DOI [10.1038/ng.2906](https://doi.org/10.1038/ng.2906).
- Wibowo MC, Yang Z, Borry M, Hübner A, Huang KD, Tierney BT, Zimmerman S, Barajas-Olmos F, Contreras-Cubas C, García-Ortiz H, Martínez-Hernández A, Lubber JM, Kirstahler P, Blohm T, Smiley FE, Arnold R, Ballall SA, Pamp SJ, Russ J, Maixner F, Rota-Stabelli O, Segata N, Reinhard K, Orozco L, Warinner C, Snow M, LeBlanc S, Kostic AD. 2021. Reconstruction of ancient microbial genomes from the human gut. *Nature* 594:234–239 DOI [10.1038/s41586-021-03532-0](https://doi.org/10.1038/s41586-021-03532-0).
- Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biology* 20(1):1–13 DOI [10.1186/s13059-019-1891-0](https://doi.org/10.1186/s13059-019-1891-0).

Application to the fermentation microbiome

Manuscript E: *Fermentation microbiome analysis of biblical king Herod's wine*

Maxime Borry, Tziona Ben Gedalya, Alexander Herbig, and Christina Warinner
Draft article

In Manuscript E, we apply the tools developed in chapter 2 for the analysis of ancient wine fermentation samples. Since the first ancient metagenomic studies, a variety of sample types has been explored, mostly limited to human associated microbiomes, such as the dental calculus, teeth, bones, or paleofeces, and different type of sediments (Fig 3.1).

However, one category of artefacts commonly found in archaeological records that remained so far mostly unexplored by metagenomic approaches, is ancient fermentation vessels (Drieu et al., 2020).

In this manuscript, we analyzed ancient wine fermentation vessels, which once belonged to the winery of king Herod "the great" of Judea, known through the Bible. After identifying which plants were present in the metagenomic records with sam2lca, we then developed a differential abundance analysis to identify which fermentation microbes still remained in the fermentation vessels. After having enriched some of these bacteria by applying in-solution capture assays to the sequencing libraries, we *de novo* assembled and binned them, checked the aDNA deamination with pyDamage, and proceeded with a functional and phylogenetic analyses.

Our findings open up a new door for the study of ancient fermentation vessels, and confirm the potential instability of the roman wine fermentation process due to the presence of spoilage microbes.

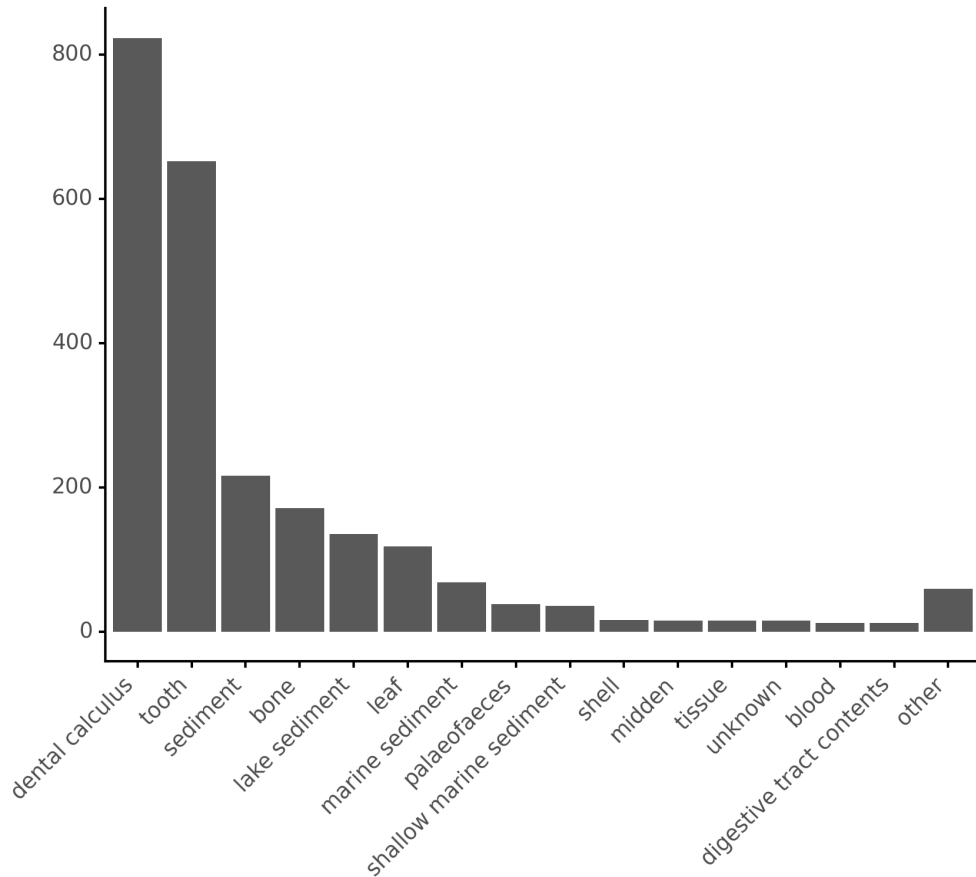


Figure 3.1: **Occurrence of different sample material types in metagenomic studies.** The number of sample for each material type was computed from the *ancientmetagenome-environmental*, *ancientmetagenome-hostassociated*, and *ancientsinglegenome-hostassociated* sample tables retrieved from the AncientMetagenomeDir, commit 079be27 (Fellows Yates et al., 2021)

Authors contributions

- **The candidate is:** first author.
- **Status:** in preparation

Authors' contributions (in %) to the given categories of the publication

Author	Conceptual	Data analysis	Experimental	Writing the manuscript	Provision of material
Maxime Borry	50	80	-	95	-
Tziona Ben Gedalya	-	10	20	-	100
Alexander Herbig	25	-	20	-	-
Christina Warinner	25	10	-	5	-
Total:	100%	100%	100%	100%	100%

Fermentation microbiome analysis of biblical king Herod's wine

Maxime Borry¹, Tziona Ben Gedalya², Alexander Herbig¹, and Christina Warinner^{1,3,4}

¹Microbiome Sciences Group, Department of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

²Eastern R&D Center, Ariel University, Ariel, Israel

³Faculty of Biological Sciences, Friedrich-Schiller University, Jena, Germany

⁴Department of Anthropology, Harvard University, Cambridge, MA, USA

ABSTRACT

The fortress of the Herodium built in 15 BCE by the Herod "the great", king of Judea, was a testimony of the expansion of the Roman culture in the Middle-East. Among the different roman influences identified, a winery was unearthed, containing clay fermentation vessels, known as dolia. The metagenomics shotgun sequencing of these dolia revealed a preservation of the fermentation microbes, from which genomes were reconstructed and phylogenetically and functionally described. The functional characterization of these bacterial genomes sheds a new light on the roman practice of wine fermentation.

Keywords: wine, roman, fermentation, microbiome, metagenomics

INTRODUCTION

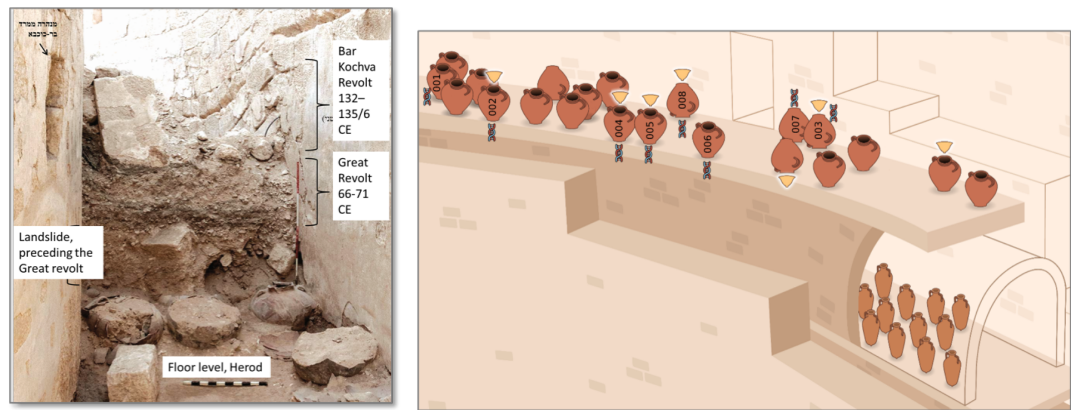
Located 12 km south of Jerusalem, at the edge of the Judean desert, the site of the Herodium, settled during the early Roman period, is a double-walled palace-fortress, in a circular shape, flanked by four towers embedded within the compound's walls, built on top of a natural hill. Ordered in 22 BCE by Herod "the Great", the construction of the palace took until 15 BCE. Herod "the Great", or Herod I, also known from the Bible, was the client king of Judea between 37 and 4 BCE, under Roman overlordship, appointed by the Roman Senate. This made Judea part of the Roman empire, which opened the doors to an array of cultural practices, which greatly influenced the style of the Herodium.

The palace occupied both a defensive and residential role, with guard towers and defense walls, but also a bathhouse, a theatre, a synagogue, mosaic floors and frescoes. It is believed that the Herodium is *Herod's* final resting place (Netzer et al., 2013). However, the body of the king has yet to be found. The palace served mostly during his reign, and the one of his son, *Herod Archelaus*, the governor of Samaria, Judea, and Idumea. It slowly fell into disuse thereafter, and was eventually partly destroyed by an earthquake, occupied by the rebels during the First Jewish–Roman War in 66 AD, and eventually sacked by the Romans in 71 AD.

While the first archeological excavations on the hilltop were conducted in the late 1960s (Corbo, 1972), it is only during the 1970's that the underground system beneath the palace was unearthed (Netzer, 1988; Netzer and Arazi, 1985). In 2017, during a new archeological campaign between the two walls of the northern wing, we discovered an early Roman period winery beneath 3 layers of later historical occupation of the site (Figure 1A), on the ground floor of the Herodium (Porat et al., 2018).

Among the various artifacts discovered in this layer, we unearthed 22 dolia, 1-meter-high large earthenware containers, which were typically used in the Roman empire as wine fermentation vessels. We found biogenic *pomace* (what is left of the pulp after a fruit has been pressed into juice) remains at the bottom of 8 of these dolia.

In the Roman empire, dolia were typically partly embedded in an earthen floor and served as receptacle for the different steps of wine making such as fermentations, racking, ageing, filtration and clarification, as it is reported by ancient roman writers such as Pliny, in his *Naturalis historia* (Brun, 2004).



(a) The structure and components of the Herodian layers (b) Reconstitution of the Herodium winery layout

Figure 1. The Herodium wine archeological excavation. (A) The layer 3, approximately up to 1 meter above the earth floor, with some areas penetrating lower, contained earth spills and collapsed structures: ceilings, roofs and walls of the perimeter corridors, as well as artifacts. This collapse layer lay on top of dozens of dolia, densely placed along the northern wing corridor. The dolia which had broken in situ, were found embedded in the floor of the original structure. The collapse layer penetrated into the top levels of several of the dolia. Beneath this collapsed layer, an intermediate level, about half a meter above the floor, includes an accumulation of sediments in which daily utensils and food scraps were found. The lower level of layer 4 begins approximately 20 cm above floor level and is situated upon compacted earth flooring (typical of most of the Herodian spaces in the fortress palace) which covered the ground level of the perimeter corridor. This lower level contained biogenic remains layered upon the inner base of the dolia, the so-called "pomace", and artifacts laying on the floor in between the dolia. Underneath the ground floor, two stories of barrel-vaulted cellars are situated, built upon the natural bedrock. (B) Dolia sampled for metagenomic analysis are annotated with a DNA pictogram, and their corresponding sample ID.

Even though the ecological understanding of the wine microbial fermentation was only made possible with the invention of the microscope in 17th century, some 1700 years later, the romans already had acquired a broad understanding of wine fermentation. They already had already established and codified a set of different practices and recommendations for wine production (Columella, 1745). For example, a great emphasis was given to the necessary cleanliness of the dolia, with a regular re-pitching of their inner surface with pine-resin, and a late harvesting of the wine grapes to maximize the sugar content, and kick-start the yeast alcoholic fermentation.

Nowadays, we have a much better understanding of the microorganisms involved in the wine fermentation process, which can broadly be divided into two different categories. The first category contains the micro-organisms responsible for the alcoholic (also known as primary) fermentation, often yeasts, among which *Saccharomyces cerevisiae* is the main actor. The second category regroups the microorganisms responsible for malolactic (or secondary) fermentation. This group is mainly composed of bacteria part of the Lactobacillaceae family, part of the Firmicutes phyla. On top of malolactic fermentation, which turns the tart-tasting malic acid into the softer tasting lactic acid, these bacteria, known as Lactic acid bacteria (LAB), are also responsible for the production of taste altering compounds, some of which are sought after, some others considered as spoilage metabolites (Vinderola et al., 2019).

While aDNA metagenomics has already been applied to a variety of material types, from human dental calculus (Warinner et al., 2014; Adler et al., 2013), to coprolites (Tito et al., 2008; Bon et al., 2012; Borry et al., 2020; Wibowo et al., 2021), ancient chewing-gum (Jensen et al., 2019), and even ancient wine grapes (Bouby et al., 2020), our understanding of the genetics of ancient wine microbes (Ramos-Madriral et al., 2019) and their role in wine fermentation remained so far very limited (Drieu et al., 2020).

Here we present the first metagenomic analysis of ancient fermentation vessels, from which we sequenced DNA obtained from pomace samples. We investigated the microbial community diversity, reconstructed genomes with *de novo* assemblies, and performed functional and phylogenetic analyses. Altogether, these findings shed a light on what could roman wine have tasted like, and how the concerns of the bible's winemakers remain the same for their modern colleagues.

MATERIALS AND METHODS

Samples

We collected pomace from 8 dolia (HEO001-008) (Fig1B) and 4 control soil samples from around the Herodium: soil from the organic rich inner garden of the Herodium palace (HEO009), foreign soil used to cover the Herodium when it fell in disuse after the death of Herod (HEO010), soil from the second archeological layer, dating to 66-70 CE (HEO011), and soil collected in the Herodian layer, between dolia (HEO012) (table S1). From each sample, we subsampled ≈ 50 mg of material before proceeding to the DNA extraction. At every step, we handled and processed the samples in dedicated ancient DNA laboratory facilities.

DNA extraction and library generation

We used a DNA extraction protocol optimised for the recovery of ultra-short DNA molecules, adapted from the Dabney extraction protocol (Dabney et al., 2013), described in details by Aron et al. (2020a) and Mann et al. (2018). We constructed double-stranded DNA sequencing libraries after UDG-half treatment (Rohland et al., 2015) following a protocol adapted from Meyer and Kircher (2010) described by Aron et al. (2020b). We double indexed the sequencing libraries following Stahl et al. (2021), and amplified them following Aron and Brandt (2020).

Targeted capture enrichment

Because the DNA originally present in archaeological samples, the so-called endogenous DNA, often only amounts to a small fraction of all recovered DNA molecules at the time of sampling, targeted capture enrichment approaches tailored for aDNA have been developed (Enk et al., 2014; Carpenter et al., 2013). Thanks to these methods, where custom designed probes will bind pre-selected sequences from genomes of interest, the sequenced fraction of endogenous DNA is greatly increased. This amounts to selectively "fishing" for the genome of taxons of interest in the ancient DNA sequencing libraries.

Based on the first estimate of the taxonomic composition of the dolia HEO001-003, and on typical

expected microbes associated with wine-fermentation (Barata et al., 2012), we designed multi-species capture arrays, for two species of yeasts and 7 lactic acid bacteria (Tab S2).

Due to differences in genome size, two different captures were designed: one for the two yeast species, and another for the lactic acid bacteria. For both captures DNA probes were designed with a length of 52 bp with an additional 8bp linker sequence (CACTGCGG) as described in Fu et al. (2013). Duplicated probes and probes with low sequence complexity were removed. For the yeast capture, probes were designed with 9 bp tiling resulting in 2,600,999 unique probe sequences. For the bacteria capture, probes were designed with 7 bp tiling resulting in 2,590,619 unique probe sequences. Each probeset was distributed over three Agilent one-million feature SureSelect DNA Capture Arrays, which were turned into in-solution DNA capture libraries as described elsewhere (Fu et al., 2013).

DNA sequencing and data processing

After an initial shallow sequencing screen was performed, we first deeper shotgun sequenced all libraries, and finally sequenced the captured libraries. The DNA libraries were sequenced on a Illumina NextSeq 500 platform for the HEO001-003 shotgun screening libraries, and on a HiSeq 4000 for all other libraries. All libraries were sequenced using a 2×75 bp chemistry.

Microbial taxonomy and community analyses

We used the nf-core/eager v2.4.4 (Ewels et al., 2020; Yates et al., 2021) pipeline to construct the metagenomic profiles. nf-core eager first aligns the raw DNA sequencing reads in FASTQ format to the GRCh38 human genome after poly-G removal with fastp (Chen et al., 2018). Unmapped reads are then taxonomically profiled by nf-core/eager with Kraken2 v2.1.2 (Wood et al., 2019) using a custom database composed of all the RefSeq reference representative genomes or sequences from archaea, bacteria, plasmids, bacteria, plasmids, viruses, human, fungi, plant, univec artificial sequences, and vertebrates, built in 03/2022.

We then processed the Kraken2 taxonomic profiles using libraries of the Python ecosystem, with Pandas v1.4.2 (McKinney et al., 2011), and taxopy v0.10.2 (Camargo and Borry, 2022) for dealing with the taxonomic information. To account for the false positive rate of Kraken2, we created a score S for each taxon, using the duplication rate d , and number of reads r , provided by Kraken2.

$$S = d \times \text{scaling}(r) + 0.01 \quad (1)$$

with

$$\text{scaling}(r) = \begin{cases} 0.01 & \text{if } r > 1000 \\ \text{rescale}(r)^4 & \text{if } r \leq 1000 \end{cases} \quad (2)$$

and

$$\text{rescale}(r) = \frac{r}{1000} \times 1.5 - 1.5 \quad (3)$$

A higher score S corresponds to taxons that are likely to be false positive taxonomic assignments (FigS1)

We discarded all taxons at the species level with a score $S > 5$, which corresponds to species with a low read count and a high duplication rate, and we kept only species present in at least 30% of the samples. To account for the compositional nature of the sequencing data, we transformed the Kraken2 species taxonomic profile with a Centered Log-Ratio transformation (CLR) (Calle, 2019) with scikit-bio v0.5.7 (scikit-bio development team, 2022). We performed a Principal Component Analysis (PCA) at the species level on the CLR transformed data with scikit-learn v1.0.2 (Pedregosa et al., 2011). For this PCA analysis, we also included modern wine (Sternes et al., 2017) and modern human skin samples (Oh et al., 2014; Chng et al., 2016; Human Microbiome Project Consortium, 2012) selected using curatedMetagenomicsData (Pasolli et al., 2017), and processed exactly as the Herodium samples. Using Statsmodels v0.13.2 (Seabold and Perktold, 2010), we then carried out a differential abundance analysis at

the species level on the CLR transformed count of the *Herodium dolia* and soil samples, for each bacterial species b , with a Linear Mixed Model (LMM) of the following design:

$$CLR(C_b) \sim \beta_t X_{b,t} + \beta_e X_{b,e} + \beta_l X_{b,l} + \beta_q X_{b,q} + (1|\alpha_s)$$

The fixed effects, β , are the different samples types $t = \{Dolia, Control\}$, the different extraction batches $e = \{1, 2\}$, the different library types $l = \{Shotgun, Capture\}$, the different Sequencers $q = \{NextSeq, HiSeq\}$, while the random effect α_s is to account for each individual sample s , because of the inclusion of more than one sequencing library per sample. Finally, we adjusted the p-values for multiple testing with the Benjamini Hochberg procedure.

Microbial genome reconstruction

We used the nextflow (Di Tommaso et al., 2017) nf-core/mag (Krakau et al., 2022) (commit 00ccccfe on the dev branch) pipeline for performing *de novo* assembly of the reads that did not map to the human genome using MEGAHIT (Li et al., 2015) after merging the libraries per sample. Assembled contigs are then checked for ancient DNA damage using PyDamage (Borry et al., 2021), and binned into Metagenome Assembled Genomes (MAGs) with MetaBAT2 (Kang et al., 2019) and MaxBin2 (Wu et al., 2016). Bins are quality assessed with Busco (Manni et al., 2021), refined with DAS Tool (Sieber et al., 2018) and taxonomically annotated using GTDB-TK (Chaumeil et al., 2020). Following the MIMAG standards (Bowers et al., 2017), we then categorized the MAGs into high (HQ), medium (MQ), and low quality (LQ), drafts according to the MIMAG reporting standards (Bowers et al., 2017).

Microbial functional and phylogenetic analysis

We used the corephylo (Borry, 2023) nextflow pipeline to conduct the functional annotation of the HQ MAGs using Bakta (Schwengers et al., 2021). From these functional annotated MAGs, corephylo uses Panaroo (Tonkin-Hill et al., 2020) to extract a core-genome alignment, which is then cleaned of recombination bearing regions with ClonalFrameML (Didelot and Wilson, 2015) (FigS4). corephylo then uses this multi core-genome non recombinant alignment to compute a maximum likelihood phylogeny using IQTree (Minh et al., 2020), with ultra-fast Bootstrapping (Hoang et al., 2018), or classic bootstrap when computationally tractable.

Plant DNA identification

We used sam2lca (Borry et al., 2022) v1.1.2 with its plant marker database, which combines ITS markers from the PLANiTS database (Banchi et al., 2020), 18s markers from the SILVA (Quast et al., 2012) and PR2 database (Guillou et al., 2012), rbcL markers from (Bell et al., 2017), and 353 flowering plant markers from the Kew Tree of Life database (Johnson et al., 2019; Baker et al., 2022). We then combined the sam2lca results for taxons having been aligned to at least 2 different references, with the Viridiplantae taxon identified by Kraken2, with a score $S < 5$.

RESULTS

Microbial ecology

We sequenced a total of 1.31 billion reads of paired-end shotgun and targeted enrichment captured DNA reads. We analyzed the Kraken2 metagenomics profiles in lower dimensions using a PCA. In this analysis, the *Herodium* samples display a clear separation from the modern wine and human skin. Within the *Herodium* samples cluster, *dolia* and soil control samples are also separated one from another (Fig2).

We then looked at the differential abundance analysis results, and the most impactful predictor variable turned out to be the sample type t : *dolia* or soil (FigS2). After regressing out the other variables thanks to the LMM, we analysed the differential abundance of taxons between *dolia* and soil (Fig3).

Among the bacterial species enriched in the *dolia* samples, we identified an elevated amount of LAB. To check if the *Herodium dolia* were especially enriched in these LAB, we performed a fisher exact test. This test indicates a significant enrichment in wine LAB, identified by Barata et al. (2012), while the *Herodium* soil control samples showed no significant enrichment (Fig S3).

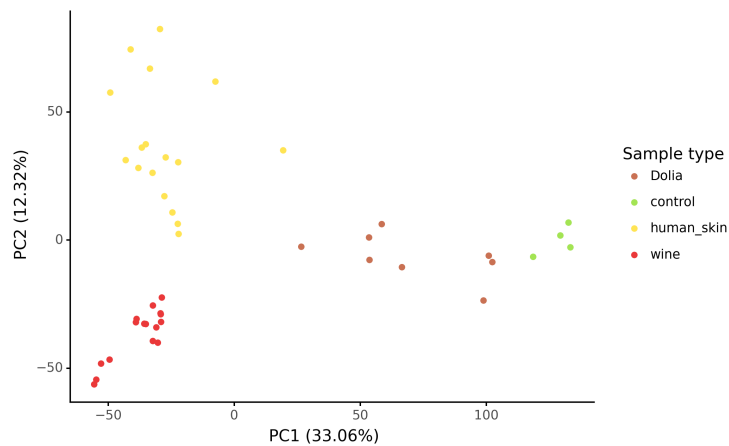


Figure 2. PCA analysis at the species level of the CLR transformed Kraken2 metagenomic profiles, of the *Herodium dolia*, and soil (control) samples, modern human skin, and modern wine samples.

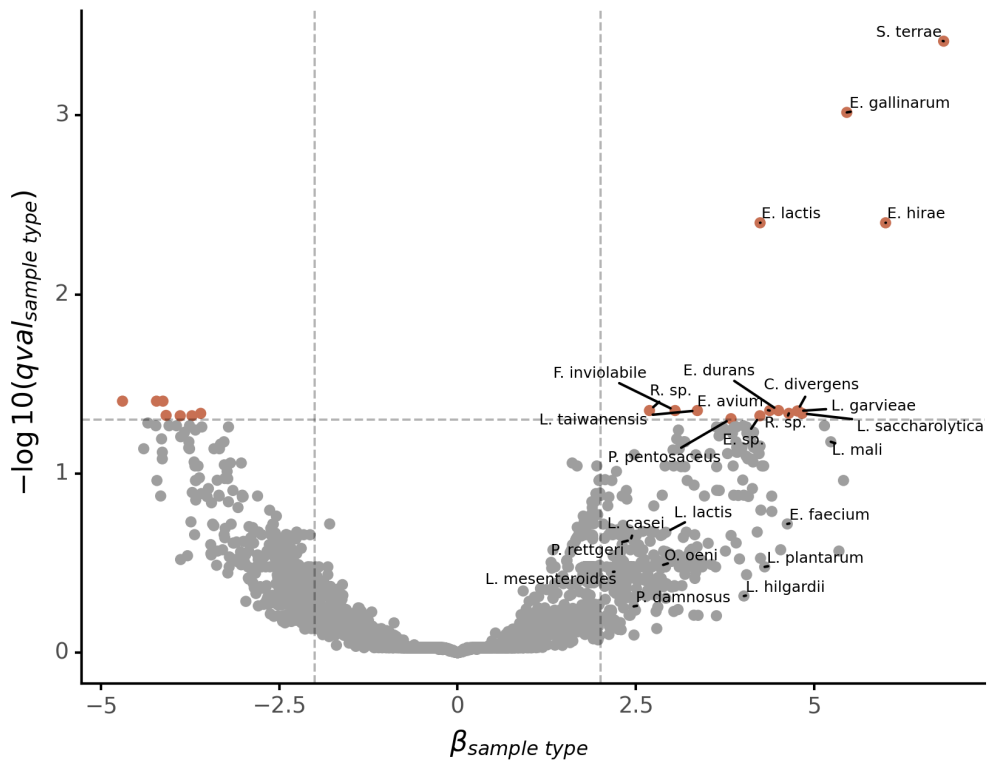


Figure 3. Volcano plot of the differential abundance analysis of *Herodium dolia* versus soil samples, with other variables regressed out by the LMM. The vertical dashed lines are the fold change (FC) thresholds of -2 , 2 , while the horizontal dashed line is the corrected p -value threshold of $-\log_{10}(0.05)$. Bacterial species passing both log fold change and corrected p -value thresholds are colored in terracotta orange. Wine LAB are annotated with their species name. The species enriched in the *dolia* samples are found on the right-most side, while the species enriched in the *Herodium* soil samples are on the left-most side.

De novo genome reconstruction

We managed to assemble between 35.2 and 336 Mbp per sample with MEGAHIT, for a N50 of between 1.1 to 3.1 Kbp per sample (Tab S3). From these contigs, we reconstructed 230 MAGs with MetaBAT2 and

MaxBin2, of which 37 were HQ, and 84 were MQ MAGs. After dereplication and bin refinement with Busco, 88 MAGs remain. Among these dereplicated refined HQ mags, 3 of them were taxonomically annotated to species that belong to the group of significantly enriched taxa in wine (S3). These HQ wine LAB MAGs (Tab 1) all display typical ancient DNA damage patterns for UDG-half libraries (Fig S5).

MAG Name	Dolia	Species	Completion (%)	Contamination (%)	N50	Number of contigs	MAG length	Coverage	GC (%)
MEGAHIT-MaxBin2Refined-HEO002-005	HEO002	<i>Lentilactobacillus hilgardii</i>	98.3	0.2	14719	299	2675945	157.6	40.1
MEGAHIT-MaxBin2Refined-HEO008-010	HEO008	<i>Pediococcus parvulus</i>	95.8	0.0	7018	330	1674845	62.4	38.93
MEGAHIT-MaxBin2Refined-HEO008.014	HEO008	<i>Lactiplantibacillus plantarum</i>	96.8	0.5	7091	532	2774156	49.5	45.51

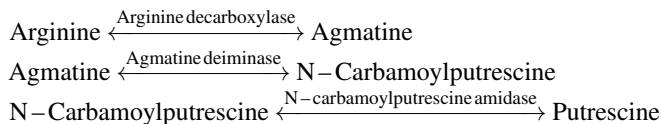
Table 1. High Quality MAGs summary statistics. The completion is Busco reported percentage of complete and single-copy specific genes, while the contamination is the Busco reported percentage of complete and duplicated specific genes.

Phylogenetic and functional analysis

We used the corephilo pipeline to functionally annotate the HQ MAGs and generate a maximum-likelihood non-recombinant core-genome phylogeny. As *L. hilgardii* (Douglas and Cruess, 1936) and *P. parvulus* (Pérez-Ramos et al., 2016, 2018) are known to be often associated with wine spoilage (Landete et al., 2005; Miranda-Castilleja et al., 2016; Inês and Falco, 2018; Barbieri et al., 2019; Vinderola et al., 2019), we looked for genes encoding for enzymes catalyzing the production wine spoilage compounds, such as biogenic amines like putrescine, and tyramine, or other wine spoilage metabolites such as β glucan polysaccharides, or ethyl carbamate. Additionally, we also looked for the malolactic enzyme.

Putrescine production pathway

The biogenic amine putrescine has a rather unpleasant odor, sometimes referred as the "smell of death" (Wisman and Shrir, 2015), and is often found in alcoholic beverages. Its production pathway, already identified in *L. hilgardii* (Arena et al., 2008), transforms the arginine amino acid into the putrescine biogenic amine. Another starting point can be the agmatine, also naturally present in wine.

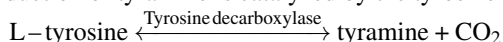


There are three enzymes, encoded by three different genes in this pathway:

- Arginine Decarboxylase, EC: 4.1.1.19
- Agmatine deiminase, EC:3.5.3.12
- N-carbamoylputrescine amidase, EC: 3.5.1.53

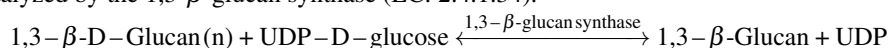
Tyramine production pathway

Tyramine is another type of biogenic amine, which in elevated concentration, can be associated with migraines (Moffett et al., 1972), increased heart rate, and blood pressure (Scriven et al., 1984). The production of tyramine is catalyzed by the tyrosine decarboxylase enzyme (EC: 4.1.1.25)



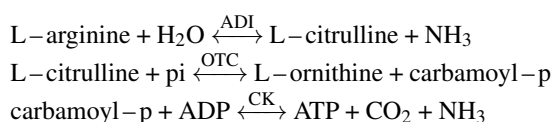
β glucan polysaccharides production pathway

β glucan polysaccharides are a class of exopolysaccharides that are responsible for giving the wine a ropy, or viscous texture. Of these, the most common is the 1,3- β -Glucan, formed by the following reaction, catalyzed by the 1,3- β -glucan synthase (EC: 2.4.1.34).



Ethyl carbamate production pathway

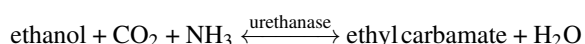
The ethyl carbamate, also know as urethane, is a carcigenenic and genotoxic compound, often found in alcoholic beverages (on the Evaluation of Carcinogenic Risks to Humans et al., 2010). It is formed by the reaction of alcohol with urea. Urea (NH₃) itself is formed by the following reactions



These reactions are catalyzed by the following enzymes:

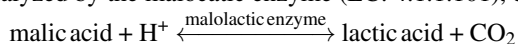
- Arginine deiminase (ADI), EC:3.5.3.6
- Ornithine carbamoyltransferase (OTC), EC:2.1.3.3
- Carbamate kinase (CK), EC: 2.7.2.2

Ethyl carbamate formation is in turn catalyzed by the enzyme urethanase (EC: 3.5.1.75)



Lactic acid production pathway

The production of lactic acid in wine is the typical thought after fermentation, known as the malolactic, or secondary fermentation. It turns the tart-tasting malic acid into the softer tasting lactic acid, a reaction catalyzed by the malolactic enzyme (EC: 4.1.1.101), encoded by the *mle* gene.



Lentilactobacillus hilgardii

The MAG *MEGAHIT-MaxBin2Refined-HEO002-005* was taxonomically assigned to the species *Lentilactobacillus hilgardii* by GTDB-Tk. We therefore constructed the phylogeny using all 13 available *L. hilgardii* reference genome assemblies on NCBI Genbank, and rooted the phylogeny with *L. farraginis* (GCA_001435875) as outgroup. The *MEGAHIT-MaxBin2Refined-HEO002-005* MAG has the shortest root to tip branch length of all *L. hilgardii* strains (Fig 4).

In *L. hilgardii*, the arginine deiminase, and *N*-carbamoylputrescine amidase enzymes, involved in the putrescine production pathway, are respectively encoded by the *aguA* and *aguB* genes. Both these genes are present in most *L. hilgardii* strains, including the HEO002 MAG (Fig 4).

Conversely, the tyrosine decarboxylase enzyme is encoded by the *tdc* gene, but only identified in two *L. hilgardii* strains (Fig 4).

Regarding the ethyl carbamate formation pathway in *L. hilgardii*, the ADI, OTC, and CK enzymes are respectively encoded by the *arcA*, *argF*, and *arcC* genes, all present in all *L. hilgardii* strains, including the HEO002 MAG (Fig 4).

The urethanase enzyme was not directly annotated by Bakta, however, by means of sequence homology in the Uniprot database (UniProt Consortium, 2019), the *gatA* gene (UniProt ID: A0A544U8T5) was identified as belonging to the same Uniref90 cluster as the urethanase (UniProt ID: A0A4Y5NHK8). While the *gatA* gene was annotated by Bakta in all *L. hilgardii* strains (locus tag *FIANLH_04270*), we also confirmed the urethanase function of the *gatA* gene annotated in the HEO002 MAG by predicting its structure from its sequence using AlphaFold2 (AF2) (Cramer, 2021) through the ColabFold interface (v1.3.0) (Mirdita et al., 2022). We then aligned it, using PyMol, to the AF2 predicted structure of A0A4Y5NHK8. The root mean square deviation (RMSD) of this structural alignment is 0.760Å (Fig S6), which is a lower than the typical resolution of crystallography resolved protein structure.

Similarly, we identified the *mle* gene encoding the malolactic fermentation enzyme by sequence homology of the *FIANLH_12060* locus tag annotated by Bakta with the Uniref90 C0XI94 cluster. This cluster also contains the A0A6G9Q9M6 protein, annotated as the malolactic enzyme. We further confirmed the annotation by following the same prediction and structure alignment, of *FIANLH_12060* to A0A6G9Q9M6, as described above, with a RMSD of 0.468Å (Fig S7).

Pediococcus parvulus

The *MEGAHIT-MaxBin2Refined-HEO008-010* MAG was taxonomically assigned by GTDB-Tk to the *Pediococcus parvulus* species. We used all 7 available reference genome assemblies for this species to construct the maximum likelihood phylogeny tree. The phylogeny was outgroup rooted with the *P. damnosus* (GCF_001611155). The *P. parvulus* MAG reconstructed from the HEO008 sample falls on the branch with the shortest root to tip distance (Fig 5).

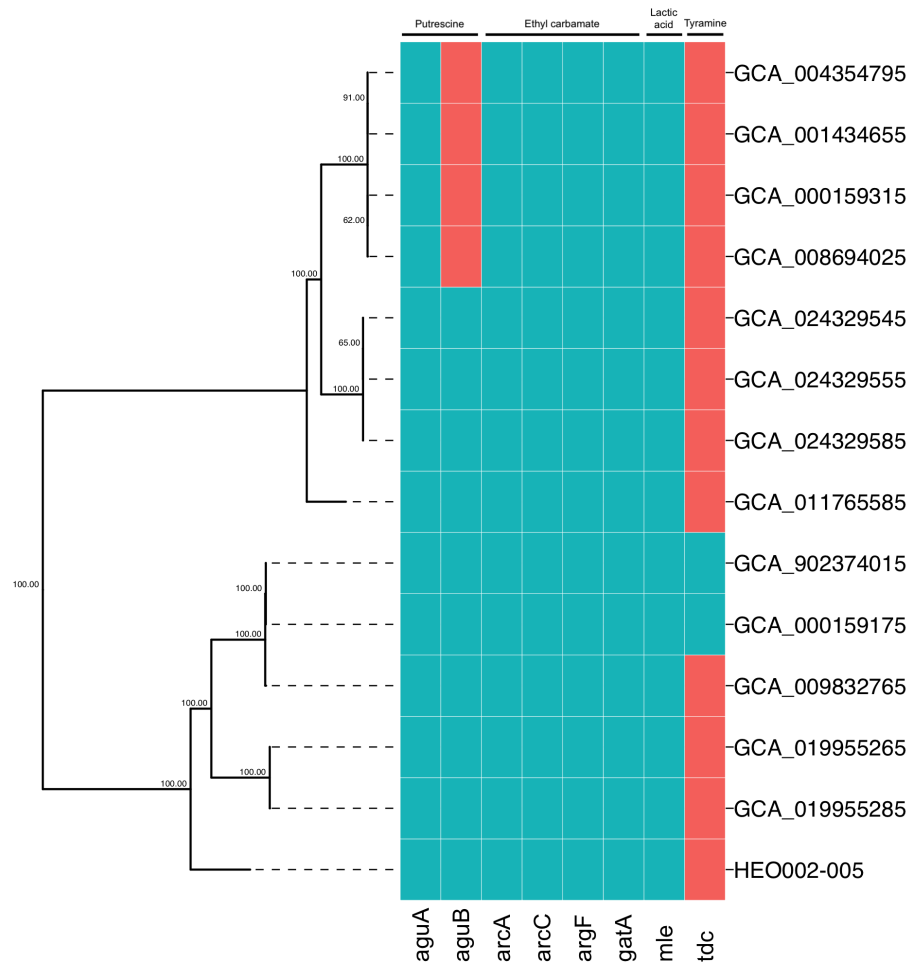


Figure 4. Outgroup rooted maximum likelihood phylogeny of the *Lentilactobacillus hilgardii* species with available reference genomes and the reconstructed MEGAHIT-MaxBin2Refined-HEO002-005 MAG. Values at the nodes indicated the UFbootstrap values. In the heatmap, a blue box indicates a gene presence, while a red box indicates its absence in the corresponding strain assembly. The product of each enzymatic pathway/reaction is indicated on the top of the heatmap.

All strains of *P. parvulus* are capable of performing malolactic fermentation thanks to the *mle* gene (Fig 5). The gene encoding the 1,3- β -glucan synthase in *P. parvulus* is *gtf*, also called *bcsA*. It is only found in the strain assembly GCF_0016440785 (Fig 5).

Plant DNA identification

Three plants were reliably identified by both sam2lca (with alignments to at least 2 references) and Kraken2 (with a $S < 5$), in 2 different samples (Tab S4).

DISCUSSION

With the excavation of the northern wing, we re-discovered the original layer of the herodian winery, buried under 5 meters, and 3 layers of later historical periods. Among the artefacts we excavated, the dolia turned out to be promising candidates for biomolecular analyses, thanks to the pomace remaining at their bottom. After the ^{14}C isotopic dating further confirmed the Herodium contextual dates, we proceeded with the metagenomic analysis of the pomace.

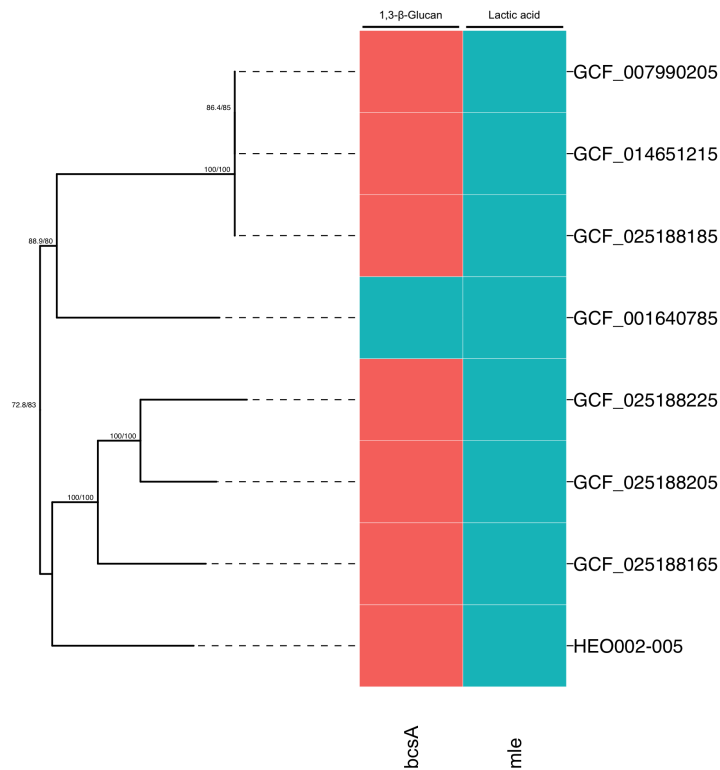


Figure 5. Outgroup rooted maximum likelihood phylogeny of the *Pediococcus parvulus* species with available reference genomes and the reconstructed MEGAHIT-MaxBin2Refined-HEO008-010 MAG. Values at the nodes indicated the Bootstrap/SH-aLRT values. In the heatmap, a blue box indicates a gene presence, while a red box indicates its absence in the corresponding strain assembly. The product of the enzymatic reaction is indicated on the top of the heatmap.

After a DNA extraction tailored for the recovery of ancient degraded DNA molecules from metagenomic sample, we submitted the the sequencing libraries to shotgun and targeted enrichment sequencing. To overcome the false negative detection potential of Kraken2, we devised the custom *S* score which proved to be effective at segregating true detected taxa from false positive artifacts: taxa only identified from a few reads were only retained if they had a low enough duplication rate (S1).

All three plants reliably identified in the pomace by both sam2lca and Kraken2, namely olives, barley, and chickpeas, were part of the common diet in judea at the time of Herod (MacDonald, 2008). However, because these plants were both identified in the pomace and in the soil control samples, they might also come from contamination from a later period, after the floor collapsed on the dolia, or crushed them.

The overall dolia bacterial communities appear to bear little resemblance with modern wine microbiomes. However, they do not display signs of complete decomposition either, as they appear separated from the Herodium soil samples. Furthermore, a modern contamination due to the handling of the samples by modern human skin microbes can be excluded as they fall in a completely different cluster (Fig 2).

Thanks to inclusion of soil samples from selected locations around the Herodium (Tab S1), we were able to conduct a differential abundance analysis to identify which factor contributed the most to explain the difference between the dolia and the soil samples bacterial communities. After having confirmed that the technical artefacts only played a minor role (Fig S2), we regressed them out to solely focused on the bacterial differential abundance explained by the difference between dolia and soil samples.

Among the bacteria most enriched in the dolia samples, we identified an elevated amount of LAB that are typically known to be associated with wine fermentation (3), which we further tested for statistical

significance (Fig S3).

After a targeted enrichment of a selection of these LAB enriched in dolia samples, we reconstructed 88 MAGs using *de novo* (Tab S3) assembly and binning of the shotgun and captured libraries. Among these, we identified HQ MAGs of the wine LAB species *Lentilactobacillus hilgardii*, and *Pediococcus parvulus* (Tab 1), which we further subjected to functional and phylogenetic analyses.

The functional analysis of *L. hilgardii* and *Pediococcus parvulus* revealed a variety of genes encoding enzymes associated with the production of wine spoilage compounds, such as the biogenic amines putrescine, responsible for foul odors, and tyramine, contributing to migraines, together with the carcinogenic compound ethyl carbamate, and the β -glucan responsible for the ropy aspect of wines. Despite being able to produce these wine spoilage metabolites, both *L. hilgardii* and *P. parvulus* also possess a gene encoding the malolactic enzyme, responsible for the malolactic fermentation, a wine quality enhancing metabolic reaction (Fig 4,5).

From a phylogenetic standpoint, both *L. hilgardii* and *P. parvulus* are falling on the branch with the shortest root to tip distance, indicating less genetic changes, and therefore suggesting a more ancient origin of these MAGs compared to reference strains (Fig 4,5).

CONCLUSION

While wine has been a staple of human beverages since 6000 BC, our knowledge of the genetic bases of its ancient fermentation was so far very limited. With this study, we brought a better understanding of some of the key actors in the process of wine microbial fermentation. Our findings shed a new light on the common roman habit of wine taste alteration with varied spices and aromas: the discovery of a variety of spoilage metabolites producing genes leads us to believe that the taste of wine produced in the Herodium wineries might have required these post-fermentation flavour enhancing practices.

DATA AVAILABILITY

The sequencing data have been deposited on ENA under the accession XXXXXXXX, and the code and scripts to reproduce the analysis of this project is available at github.com/maxibor/herodium-metagenomics

ACKNOWLEDGMENTS

Additional information can be given in the template, such as to not include funder information in the acknowledgments section.

REFERENCES

- Christina J Adler, Keith Dobney, Laura S Weyrich, John Kaidonis, Alan W Walker, Wolfgang Haak, Corey JA Bradshaw, Grant Townsend, Arkadiusz Sołtysiak, Kurt W Alt, et al. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the neolithic and industrial revolutions. *Nature genetics*, 45(4):450–455, 2013.
- M.e. Arena, J.m. Landete, M.c. Manca de Nadra, I. Pardo, and S. Ferrer. Factors affecting the production of putrescine from agmatine by *Lactobacillus hilgardii* X1B isolated from wine. *Journal of Applied Microbiology*, 105(1):158–165, 2008. ISSN 1365-2672. doi: 10.1111/j.1365-2672.2008.03725.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2672.2008.03725.x>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2672.2008.03725.x>.
- Franziska Aron and Guido Brandt. Amplification and pooling. *protocols.io*, 12 2020. <https://dx.doi.org/10.17504/protocols.io.beqkjdju>.
- Franziska Aron, Courtney Hofman, Zandra Fagernäs, Irina Velsko, Eirini Skourtanioti, Guido Brandt, and Christina Warinner. Ancient dna extraction from dental calculus. *protocols.io*, 12 2020a. <https://dx.doi.org/10.17504/protocols.io.bidyka7w>.
- Franziska Aron, Gunnar Neumann, and Guido Brandt. Half-udg treated double-stranded ancient dna library preparation for illumina sequencing. *protocols.io*, 12 2020b. <https://dx.doi.org/10.17504/protocols.io.bmh6k39e>.
- William J Baker, Paul Bailey, Vanessa Barber, Abigail Barker, Sidonie Bellot, David Bishop, Laura R Botigué, Grace Brewer, Tom Carruthers, James J Clarkson, et al. A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Systematic biology*, 71(2):301–319, 2022.
- Elisa Banchi, Claudio G Ametrano, Samuele Greco, David Stanković, Lucia Muggia, and Alberto Pallavicini. Planits: a curated sequence reference dataset for plant its dna metabarcoding. *Database*, 2020, 2020.
- A Barata, Manuel Malfeito-Ferreira, and Virgilio Loureiro. The microbial ecology of wine grape berries. *International journal of food microbiology*, 153(3):243–259, 2012.
- Federica Barbieri, Chiara Montanari, Fausto Gardini, and Giulia Tabanelli. Biogenic Amine Production by Lactic Acid Bacteria: A Review. *Foods*, 8(1):17, January 2019. ISSN 2304-8158. doi: 10.3390/foods8010017. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351943/>.
- Karen L Bell, Virginia M Loeffler, and Berry J Brosi. An rbcl reference library to aid in the identification of plant species mixtures by dna metabarcoding. *Applications in plant sciences*, 5(3):1600110, 2017.
- Céline Bon, Véronique Berthonaud, Frédéric Maksud, Karine Labadie, Julie Poulain, François Artiguenave, Patrick Wincker, Jean-Marc Aury, and Jean-Marc Elalouf. Coprolites as a source of information on the genome and diet of the cave hyena. *Proceedings of the Royal Society B: Biological Sciences*, 279(1739):2825–2830, 2012.
- Maxime Borry. maxibor/corephylo: corephylo v1.0.1, January 2023. URL <https://doi.org/10.5281/zenodo.7509937>.
- Maxime Borry, Bryan Cordova, Angela Perri, Marsha Wibowo, Tanvi Prasad Honap, Jada Ko, Jie Yu, Kate Britton, Linus Girdland-Flink, Robert C Power, et al. Coproind predicts the source of coprolites and paleofeces using microbiome composition and host dna content. *PeerJ*, 8:e9001, 2020.
- Maxime Borry, Alexander Hübner, Adam B Rohrlach, and Christina Warinner. Pydamage: automated ancient damage identification and estimation for contigs in ancient dna de novo assembly. *PeerJ*, 9:e11845, 2021.
- Maxime Borry, Alexander Hübner, and Christina Warinner. sam2lca: Lowest common ancestor for sam/bam/cram alignment files. *Journal of Open Source Software*, 7(74):4360, 2022.
- Laurent Bouby, Nathan Wales, Mindia Jalabadze, Nana Rusishvili, Vincent Bonhomme, Jazmín Ramos-Madrugal, Allowen Evin, Sarah Ivorra, Thierry Lacombe, Clémence Pagnoux, Elisabetta Boaretto, M. Thomas P. Gilbert, Roberto Bacillieri, David Lordkipanidze, and David Maghradze. Tracking the history of grapevine cultivation in Georgia by combining geometric morphometrics and ancient DNA. *Vegetation History and Archaeobotany*, October 2020. ISSN 1617-6278. doi: 10.1007/s00334-020-00803-0.
- Robert M Bowers, Nikos C Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, TBK Reddy, Frederik Schulz, Jessica Jarett, Adam R Rivers, Emiley A Eloë-Fadrosh, et al. Minimum information about a single amplified genome (misag) and a metagenome-assembled genome (mimag) of bacteria and archaea. *Nature biotechnology*, 35(8):725–731, 2017.

- Jean-Pierre Brun. *Archeologie Du Vin et de l'huile Dans l'Empire Romain*. Collection Des Hesperides. 2004. ISBN 978-2-87772-293-3.
- M Luz Calle. Statistical analysis of metagenomics data. *Genomics & informatics*, 17(1), 2019.
- Antônio Camargo and Maxime Borry. apcamargo/taxopy: v0.10.2, August 2022. URL <https://doi.org/10.5281/zenodo.7010602>.
- Meredith L Carpenter, Jason D Buenrostro, Cristina Valdiosera, Hannes Schroeder, Morten E Allentoft, Martin Sikora, Morten Rasmussen, Simon Gravel, Sonia Guillén, Georgi Nekhrizov, et al. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient dna sequencing libraries. *The American Journal of Human Genetics*, 93(5):852–864, 2013.
- Pierre-Alain Chaumeil, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. Gtdb-tk: a toolkit to classify genomes with the genome taxonomy database, 2020.
- Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018.
- Kern Rei Chng, Angeline Su Ling Tay, Chenhao Li, Amanda Hui Qi Ng, Jingjing Wang, Bani Kaur Suri, Sri Anusha Matta, Naomi McGovern, Baptiste Janela, Xuan Fei Colin C Wong, et al. Whole metagenome profiling reveals skin microbiome-dependent susceptibility to atopic dermatitis flare. *Nature microbiology*, 1(9):1–10, 2016.
- Lucius Junius Moderatus Columella. *L. Junius Moderatus Columella Of Husbandry: In Twelve Books: and His Book Concerning Trees*. A. Millar, 1745.
- Virgilio C Corbo. *Herodion / I Gli Edifici Della Reggia-Fortezza*. Studium Biblicum Franciscanum Collectio Maior. Franciscan Print. Press, 1972.
- Patrick Cramer. Alphafold2 and the future of structural biology. *Nature Structural & Molecular Biology*, 28(9):704–705, 2021.
- Jesse Dabney, Michael Knapp, Isabelle Glocke, Marie-Theres Gansauge, Antje Weihmann, Birgit Nickel, Cristina Valdiosera, Nuria García, Svante Pääbo, Juan-Luis Arsuaga, et al. Complete mitochondrial genome sequence of a middle pleistocene cave bear reconstructed from ultrashort dna fragments. *Proceedings of the National Academy of Sciences*, 110(39):15758–15763, 2013.
- Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319, 2017.
- Xavier Didelot and Daniel J Wilson. Clonalframeml: efficient inference of recombination in whole bacterial genomes. *PLoS computational biology*, 11(2):e1004041, 2015.
- H. C. Douglas and W. V. Cruess. A LACTOBACILUUS FROM CALIFORNIA WINE: LACTOBACILLUS HILGARDII. *Journal of Food Science*, 1(2):113–119, March 1936. ISSN 0022-1147, 1750-3841. doi: 10.1111/j.1365-2621.1936.tb17774.x. URL <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2621.1936.tb17774.x>.
- Léa Drieu, Maxime Rageot, Nathan Wales, Ben Stern, Jasmine Lundy, Maximilian Zerrer, Isabella Gaffney, Manon Bondetti, Cynthia Spiteri, Jane Thomas-Oates, et al. Is it possible to identify ancient wine production using biomolecular approaches? *STAR: Science & Technology of Archaeological Research*, 6(1):16–29, 2020.
- Jacob M Enk, Alison M Devault, Melanie Kuch, Yusuf E Murgha, Jean-Marie Rouillard, and Hendrik N Poinar. Ancient whole genome enrichment using baits built from modern dna. *Molecular biology and evolution*, 31(5):1292–1294, 2014.
- Philip A Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. The nf-core framework for community-curated bioinformatics pipelines. *Nature biotechnology*, 38(3):276–278, 2020.
- Qiaomei Fu, Matthias Meyer, Xing Gao, Udo Stenzel, Hernán A Burbano, Janet Kelso, and Svante Pääbo. Dna analysis of an early modern human from tianyuan cave, china. *Proceedings of the National Academy of Sciences*, 110(6):2223–2227, 2013.
- Laure Guillou, Dipankar Bachar, Stéphane Audic, David Bass, Cédric Berney, Lucie Bittner, Christophe Boutte, Gaétan Burgaud, Colomban de Vargas, Johan Decelle, et al. The protist ribosomal reference database (pr2): a catalog of unicellular eukaryote small sub-unit rna sequences with curated taxonomy. *Nucleic acids research*, 41(D1):D597–D604, 2012.
- Diep Thi Hoang, Olga Chernomor, Arndt Von Haeseler, Bui Quang Minh, and Le Sy Vinh. Ufboot2: improving the ultrafast bootstrap approximation. *Molecular biology and evolution*, 35(2):518–522,

- 2018.
- Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012. ISSN 1476-4687. doi: 10.1038/nature11234. URL <https://www.nature.com/articles/nature11234>.
- António Inês and Virgílio Falco. *Lactic Acid Bacteria Contribution to Wine Quality and Safety*. IntechOpen, November 2018. ISBN 978-1-78984-453-5. doi: 10.5772/intechopen.81168. URL <https://www.intechopen.com/chapters/undefined/state.item.id>. Publication Title: Generation of Aromas and Flavours.
- Theis ZT Jensen, Jonas Niemann, Katrine Højholt Iversen, Anna K Fotakis, Shyam Gopalakrishnan, Åshild J Vågene, Mikkel Winther Pedersen, Mikkel-Holger S Sinding, Martin R Ellegaard, Morten E Allentoft, et al. A 5700 year-old human genome and oral microbiome from chewed birch pitch. *Nature Communications*, 10(1):1–10, 2019.
- Matthew G Johnson, Lisa Pokorny, Steven Dodsworth, Laura R Botigue, Robyn S Cowan, Alison Devault, Wolf L Eiserhardt, Niroshini Epitawalage, Félix Forest, Jan T Kim, et al. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic biology*, 68(4):594–606, 2019.
- Dongwan D Kang, Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359, 2019.
- Sabrina Krakau, Daniel Straub, Hadrien Gourel, Gisela Gabernet, and Sven Nahnsen. nf-core/mag: a best-practice pipeline for metagenome hybrid assembly and binning. *NAR Genomics and Bioinformatics*, 4(1):lqac007, 2022.
- Jason Kwong. Maskrc-svg, February 2023. URL <https://github.com/kwongj/maskrc-svg>.
- J.M. Landete, S. Ferrer, and I. Pardo. Which lactic acid bacteria are responsible for histamine production in wine? *Journal of Applied Microbiology*, 99(3):580–586, September 2005. ISSN 1364-5072, 1365-2672. doi: 10.1111/j.1365-2672.2005.02633.x. URL <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2672.2005.02633.x>.
- Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676, 2015.
- Nathan MacDonald. *What Did the Ancient Israelites Eat?: Diet in Biblical Times*. Wm. B. Eerdmans Publishing, 2008.
- Allison E Mann, Susanna Sabin, Kirsten Ziesemer, Åshild J Vågene, Hannes Schroeder, Andrew T Ozga, Krithivasan Sankaranarayanan, Courtney A Hofman, James A Fellows Yates, Domingo C Salazar-García, et al. Differential preservation of endogenous human and microbial dna in dental calculus and dentin. *Scientific reports*, 8(1):1–15, 2018.
- Mosè Manni, Matthew R Berkeley, Mathieu Seppey, Felipe A Simão, and Evgeny M Zdobnov. Busco update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution*, 38(10):4647–4654, 2021.
- Wes McKinney et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9, 2011.
- Matthias Meyer and Martin Kircher. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, 2010(6):pdb-prot5448, 2010.
- Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt Von Haeseler, and Robert Lanfear. Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, 37(5):1530–1534, 2020.
- Dalia E. Miranda-Castilleja, Ramon Alvar Martínez-Peniche, J. A. Aldrete-Tapia, Lourdes Soto-Muñoz, Montserrat H. Iturriaga, J. R. Pacheco-Aguilar, and Sofía M. Arvizu-Medrano. Distribution of Native Lactic Acid Bacteria in Wineries of Queretaro, Mexico and Their Resistance to Wine-Like Conditions. *Frontiers in Microbiology*, 7, 2016. ISSN 1664-302X. URL <https://www.frontiersin.org/articles/10.3389/fmicb.2016.01769>.
- Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature Methods*, pages 1–4, 2022.
- Adrienne Moffett, Michael Swash, and DF Scott. Effect of tyramine in migraine: a double-blind study.

- Journal of Neurology, Neurosurgery & Psychiatry*, 35(4):496–499, 1972.
- Ehud 1934-2010 Netser. Jewish Rebels Dig Strategic Tunnel System. *Biblical archaeology review*, 14(4): 18, 1988. ISSN 0098-9444.
- E. Netzer and S. Arazi. The Tunnels of Herodium. *Qadmoniot: A Journal for the Antiquities of Eretz-Israel and Bible Lands*, (1/2 (69/70)):33–38, 1985. ISSN 0033-4839.
- Ehud Netzer, Roi Porat, Yakov Kalman, and Rachel Chachy. Herodium. *Herod the Great, The King's Final Journey, Jerusalem*, pages 126–161, 2013.
- Julia Oh, Allyson L Byrd, Clay Deming, Sean Conlan, Heidi H Kong, and Julia A Segre. Biogeography and individuality shape function in the human skin metagenome. *Nature*, 514(7520):59–64, 2014.
- IARC Working Group on the Evaluation of Carcinogenic Risks to Humans et al. Alcohol consumption and ethyl carbamate. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, 96:3, 2010.
- Edoardo Pasolli, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, Faizan Malik, Marcel Ramos, Jennifer B Dowd, et al. Accessible, curated metagenomic data through experimenthub. *Nature methods*, 14(11):1023–1024, 2017.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python, 2011.
- Roi Porat, Yakov Kalman, Rachel Chachy, shulamit terem, Rachel Bar-Natan, Avner Ecker, Tziona Ben-Gedalya, Elyashiv Drori, and Ehud Weiss. Herod's Royal Winery and Wine Storage Facility in the Outer Structure of the Mountain Palace-Fortress at Herodium (Hebrew- Qadmoniot. 156:106–114, January 2018.
- Adrián Pérez-Ramos, M. Luz Mohedano, Ana Puertas, Antonella Lamontanara, Luigi Orru, Giuseppe Spano, Vittorio Capozzi, M. Teresa Dueñas, and Paloma López. Draft Genome Sequence of *Pediococcus parvulus* 2.6, a Probiotic beta-Glucan Producer Strain. *Genome Announcements*, 4(6):e01381–16, December 2016. doi: 10.1128/genomeA.01381-16. URL <https://journals.asm.org/doi/full/10.1128/genomeA.01381-16>. Publisher: American Society for Microbiology.
- Adrián Pérez-Ramos, Maria L. Mohedano, Miguel A. Pardo, and Paloma López. Beta-Glucan-Producing *Pediococcus parvulus* 2.6: Test of Probiotic and Immunomodulatory Properties in Zebrafish Models. *Frontiers in Microbiology*, 9, 2018. ISSN 1664-302X. URL <https://www.frontiersin.org/articles/10.3389/fmicb.2018.01684>.
- Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(D1):D590–D596, 2012.
- Jazmín Ramos-Madrigal, Anne Kathrine Wiborg Runge, Laurent Bouby, Thierry Lacombe, José Alfredo Samaniego Castruita, Anne-Françoise Adam-Blondon, Isabel Figueiral, Charlotte Hallavant, José M Martínez-Zapater, Caroline Schaal, et al. Palaeogenomic insights into the origins of french grapevine diversity. *Nature plants*, 5(6):595–603, 2019.
- Nadin Rohland, Eadaoin Harney, Swapan Mallick, Susanne Nordenfelt, and David Reich. Partial uracil-dna-glycosylase treatment for screening of ancient dna. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1660):20130624, 2015.
- Oliver Schwengers, Lukas Jelonek, Marius Alfred Dieckmann, Sebastian Beyvers, Jochen Blom, and Alexander Goesmann. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial genomics*, 7(11), 2021.
- The scikit-bio development team. scikit-bio: A bioinformatics library for data scientists, students, and developers, 2022. URL <http://scikit-bio.org>.
- AJ Scriven, Morris J Brown, Michael B Murphy, and Colin T Dollery. Changes in blood pressure and plasma catecholamines caused by tyramine and cold exposure. *Journal of cardiovascular pharmacology*, 6(5):954–960, 1984.
- Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, pages 10–25080. Austin, TX, 2010.
- Christian MK Sieber, Alexander J Probst, Allison Sharrar, Brian C Thomas, Matthias Hess, Susannah G Tringe, and Jillian F Banfield. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature microbiology*, 3(7):836–843, 2018.
- Raphaela Stahl, Christina Warinner, Irina Velsko, Eleftheria Orfanou, Franziska Aron, and Guido

- Brandt. Illumina double-stranded dna dual indexing for ancient dna. *protocols.io*, 06 2021. <https://dx.doi.org/10.17504/protocols.io.bvt8n6rw>.
- Peter R Sternes, Danna Lee, Dariusz R Kutyna, and Anthony R Borneman. A combined meta-barcoding and shotgun metagenomic analysis of spontaneous wine fermentation. *Gigascience*, 6(7):gix040, 2017.
- Raúl Y Tito, Simone Macmil, Graham Wiley, Fares Najar, Lauren Cleeland, Chunmei Qu, Ping Wang, Frederic Romagne, Sylvain Leonard, Agustín Jiménez Ruiz, et al. Phylotyping and functional analysis of two ancient human microbiomes. *PLoS One*, 3(11):e3703, 2008.
- Gerry Tonkin-Hill, Neil MacAlasdair, Christopher Ruis, Aaron Weimann, Gal Horesh, John A Lees, Rebecca A Gladstone, Stephanie Lo, Christopher Beaudoin, R Andres Floto, et al. Producing polished prokaryotic pangenomes with the panaroo pipeline. *Genome biology*, 21(1):1–21, 2020.
- UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1): D506–D515, 2019.
- Gabriel Vinderola, Arthur Ouwehand, Seppo Salminen, and Atte von Wright. *Lactic acid bacteria: microbiological and functional aspects*. Crc Press, 2019.
- Christina Warinner, João F Matias Rodrigues, Rounak Vyas, Christian Trachsel, Natallia Shved, Jonas Grossmann, Anita Radini, Y Hancock, Raul Y Tito, Sarah Fiddyment, et al. Pathogens and host immunity in the ancient human oral cavity. *Nature genetics*, 46(4):336–344, 2014.
- Marsha C Wibowo, Zhen Yang, Maxime Borry, Alexander Hübner, Kun D Huang, Braden T Tierney, Samuel Zimmerman, Francisco Barajas-Olmos, Cecilia Contreras-Cubas, Humberto García-Ortiz, et al. Reconstruction of ancient microbial genomes from the human gut. *Nature*, 594(7862):234–239, 2021.
- Arnaud Wisman and Ilan Shrira. The smell of death: evidence that putrescine elicits threat management mechanisms. *Frontiers in Psychology*, 6:1274, 2015.
- Derrick E Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with kraken 2. *Genome biology*, 20(1):1–13, 2019.
- Yu-Wei Wu, Blake A Simmons, and Steven W Singer. Maxbin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4):605–607, 2016.
- James A Fellows Yates, Theseas C Lamnidis, Maxime Borry, Aida Andrades Valtueña, Zandra Fagernäs, Stephen Clayton, Maxime U Garcia, Judith Neukamm, and Alexander Peltzer. Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. *PeerJ*, 9:e10947, 2021.

SUPPLEMENTARY MATERIAL

Archeological ID	Sample ID	Sampled weight (mg)	# Input reads	Extraction Batch
P2 HR LH3540	HEO001	52,9	92 789 988	1
P5 HR L3546	HEO002	52,3	152 855 912	1
P20 HR L3563	HEO003	54,6	89 174 981	1
P10 L3559	HEO004	52,6	101 915 757	2
P12 L3570	HEO005	52,7	99 596 344	2
P13 L3558	HEO006	48,2	117 660 992	2
P14 L3564	HEO007	51,0	92 036 489	2
P23 L3578	HEO008	49,2	140 225 340	2
Garden soil L.H.3588	HEO009	52,7	109 905 794	2
Neutral earth control	HEO010	48,5	115 919 526	2
Floor remains	HEO011	51,8	108 591 098	2
Earth control	HEO012	51,5	93 041 849	2

Table S1. Description of the samples

Species	Type	NCBI genome accession
<i>Lentilactobacillus hilgardii</i>	bacteria	GCF_011765585.1
<i>Lactiplantibacillus plantarum</i>	bacteria	GCF_003269405.1
<i>Gluconobacter albidus</i>	bacteria	GCF_002005485.1
<i>Gluconobacter cerinus</i>	bacteria	GCF_002723935.1
<i>Pediococcus parvulus</i>	bacteria	GCF_007990205.1
<i>Pediococcus damnosus</i>	bacteria	GCF_001611155.1
<i>Oenococcus oeni</i>	bacteria	GCF_002966535.1
<i>Saccharomyces cerevisiae</i>	yeast	GCF_000146045.2
<i>Brettanomyces bruxellensis</i>	yeast	GCA_900496985.1

Table S2. Species used for the design of the targeted capture enrichment

Sample Name	N50 (Kbp)	N75 (Kbp)	L50 (K)	L75 (K)	Largest contig (Kbp)	Length (Mbp)
HEO001	2.6Kbp	1.1Kbp	10.0K	30.2K	222.6Kbp	134.2Mbp
HEO002	2.3Kbp	1.0Kbp	25.1K	84.3K	604.4Kbp	336.0Mbp
HEO003	1.9Kbp	0.9Kbp	13.8K	46.5K	521.3Kbp	160.5Mbp
HEO004	3.1Kbp	1.1Kbp	4.6K	15.8K	172.4Kbp	79.3Mbp
HEO005	2.4Kbp	0.9Kbp	2.8K	9.4K	85.0Kbp	35.2Mbp
HEO006	2.4Kbp	0.9Kbp	7.9K	29.5K	258.1Kbp	119.2Mbp
HEO007	1.2Kbp	0.7Kbp	11.8K	33.4K	81.9Kbp	76.3Mbp
HEO008	2.4Kbp	0.9Kbp	15.6K	53.7K	348.9Kbp	214.0Mbp
HEO009	1.1Kbp	0.7Kbp	22.5K	49.7K	86.8Kbp	92.0Mbp
HEO010	1.4Kbp	0.8Kbp	27.0K	68.0K	116.3Kbp	161.6Mbp
HEO011	1.2Kbp	0.7Kbp	43.4K	102.7K	120.9Kbp	213.1Mbp
HEO012	1.1Kbp	0.7Kbp	34.4K	78.2K	135.1Kbp	152.9Mbp

Table S3. *de novo* assembly summary statistics

Sample	Library	TAXID	Species	Kraken read count	Kraken S score	sam2lca read count	sam2lca reference count
HEO011	HEO011.A0101.SG1	4146	<i>Olea europaea</i>	3699	0.02	7	4
HEO011	HEO011.A0101.FB1	4146	<i>Olea europaea</i>	24022	0.02	7	4
HEO011	HEO011.A0101.SG1	4513	<i>Hordeum vulgare</i>	3550	0.02	12	6
HEO011	HEO011.A0101.FB1	4513	<i>Hordeum vulgare</i>	25774	0.02	12	6
HEO008	HEO008.A0101.FB1	3827	<i>Cicer arietinum</i>	189	4.69	1	8
HEO008	HEO008.A0101.SG1	3827	<i>Cicer arietinum</i>	169	4.21	1	8

Table S4. Plant species identified reliably by both sam2lca and Kraken2

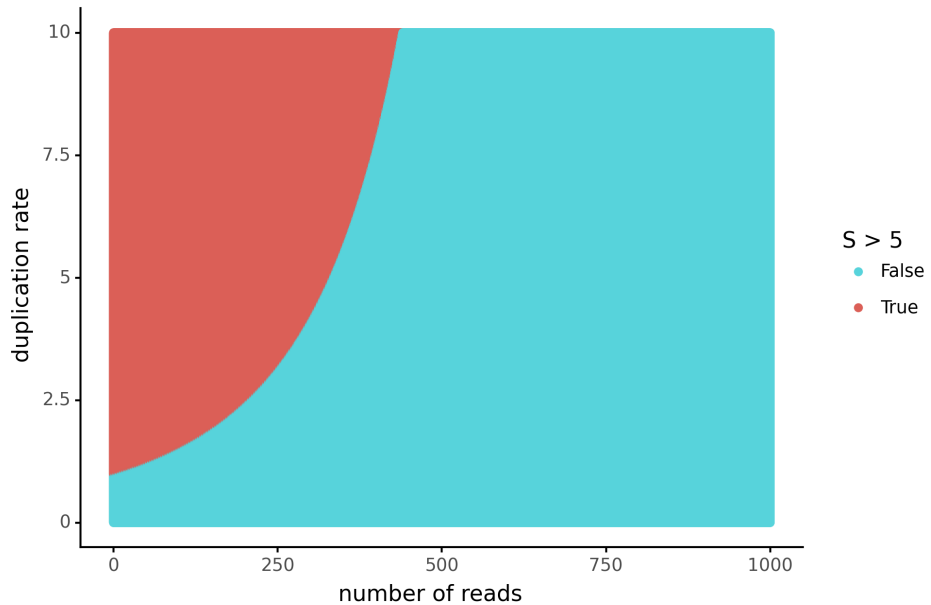


Figure S1. Kraken2 Score S computed from the number of reads and Kraken2 duplication rate. A higher S corresponds to an increased likelihood of a false positive taxonomic assignment.

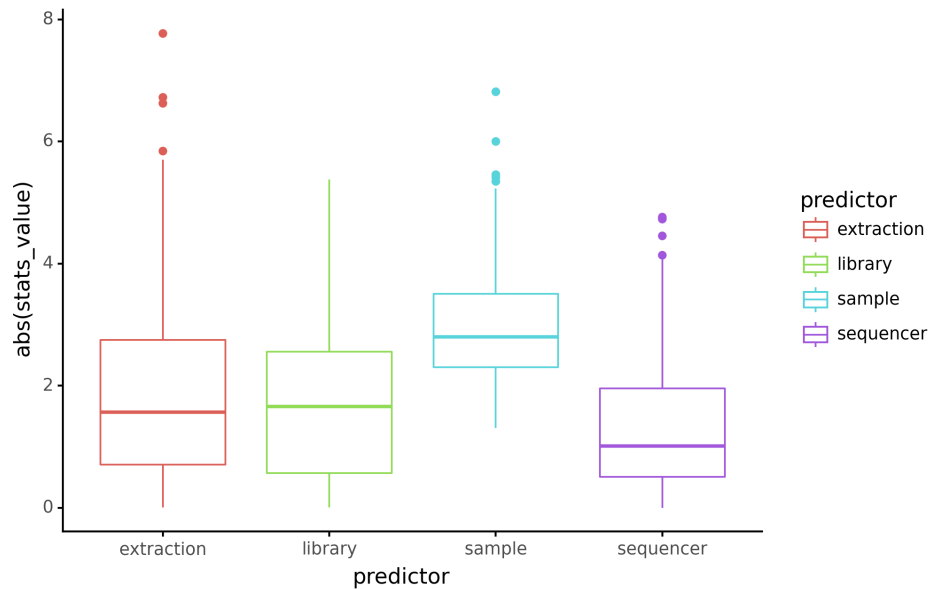


Figure S2. Boxplot of the coefficients for the different predictor variables of the Generalized Linear Mixed model. *Extraction* corresponds to the different DNA extraction batches, *library* corresponds to the different library processing methods (shotgun or targeted enrichment), *sequencer* corresponds to the different sequencers, and *sample* corresponds to the categories of the samples (dolia or soil).

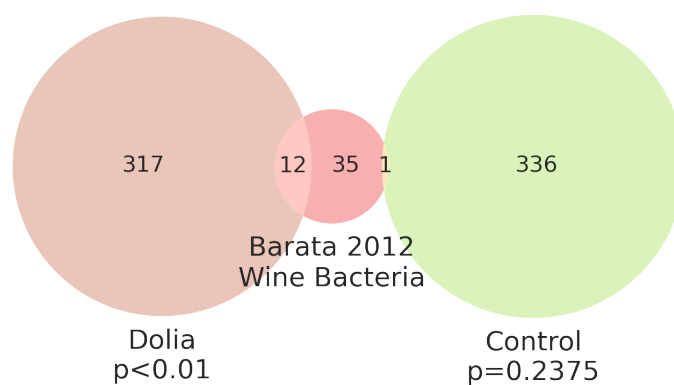
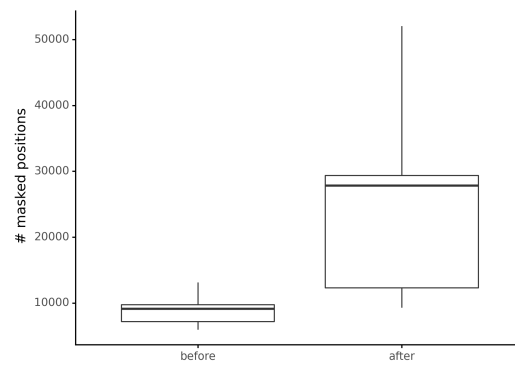
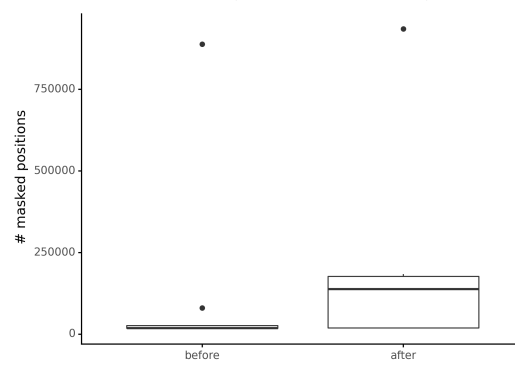


Figure S3. Venn diagram of the intersection of bacteria enriched in dolia ($FC > 2$), enriched in Herodium soils ($FC > 2$), and wine bacteria identified in Barata et al. (2012). p-value of the fisher exact test



(a) Number of masked positions in *L. hilgardii* genomes before and after checking for recombinant regions



(b) Number of masked positions in *P. parvulus* genomes before and after checking for recombinant regions

Figure S4. Recombinant regions identified with ClonalFrameML were then masked with maskrc-svg (Kwong, 2023)

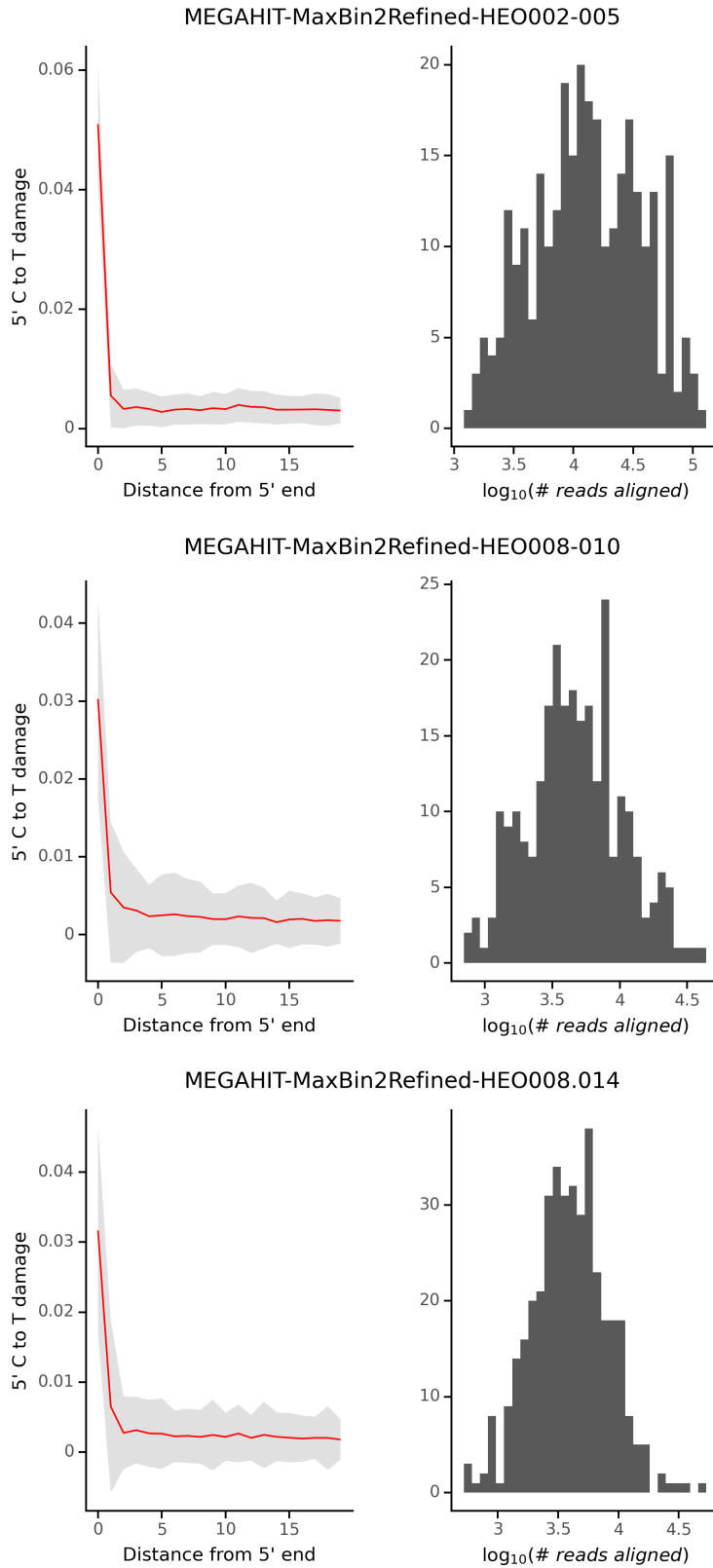


Figure S5. Ancient DNA C to T transitions from the 5' end damage plot (left), and distribution of the number of reads aligned per contig (right) for each HQ MAG. On the damage plot, the damage is averaged over all contigs of the MAG (red line), and the standard deviation is represented by the shaded area.



Figure S6. PyMol alignment of the AF2 predicted structure of the *gatA* gene in the *MEGAHIT-MaxBin2Refined-HEO002-005* MAG in pink, to the AF2 predicted structure of the Urethanase (uniprot ID: A0A4Y5NHK8) in grey. *RMSD* = 0.760Å

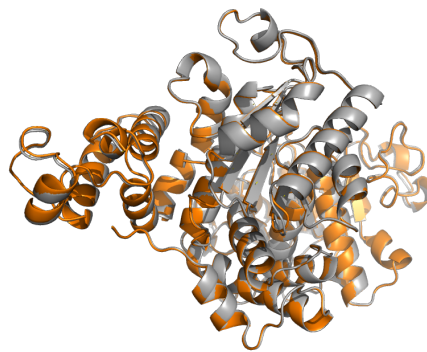


Figure S7. PyMol alignment of the AF2 predicted structure of the *FIANLH_12060* locus tag in the *MEGAHIT-MaxBin2Refined-HEO008-010* MAG in orange, to the AF2 predicted structure of the malolactic enzyme (uniprot ID: A0A6G9Q9M6) in grey. *RMSD* = 0.468Å

Discussion

An additional line of evidence for paleofeces identification

Out of the different possible archaeological materials available for the study of human microbiomes, paleofeces is the microbiome source that remains the most understudied (Fig 3.1) compared to its modern counterpart, arguably the most studied human microbiome: the gut microbiome. One of the reasons for the lack of ancient DNA studies on paleofeces is the archaeological context itself. While microbiome material sources like dental calculus, teeth, or bones are sampled directly from an archaeological source morphologically identifiable as human remains, paleofeces in archaeological context are mostly found exogenously, usually outside of human burial sites. Furthermore, the macro and micromorphological examination of paleofeces is often unable to pinpoint their source of origin. In addition, paleofeces typically need very special environmental conditions to preserve and not immediately decompose. A dry, relatively cold and stable environment is required, such as caves, or mines. The combination of all of these factors makes the study of paleofeces more challenging than the more abundant other sources of ancient microbiomes. It was therefore even more necessary to make sure that the few available paleofeces samples were correctly identified as humans. While this challenge had previously been acknowledged, the techniques used to address it, namely morphology, rehydration, macro and micro inclusions with the help of scanning electron microscopy (Reinhard et al., 2019), and parasite composition remained sometimes insufficient. The challenge of coprolite identification proved to be especially tough for distinguishing dog from human paleofeces (Poinar et al., 2009b), owing the shared habitat of humans and their four legged companions since the domestication of

dogs more than 12,000 years ago (Frantz et al., 2016). In order to provide an additional line of evidence to identify paleofeces, we leveraged the information potential provided by shotgun metagenomics sequencing, which is composed of both host, and microbial DNA. As humans and dogs have a very distinct gut microbiome composition, we were able to use this information, in addition to the amount of endogenous DNA in their feces, to predict the origin of their paleofeces. This approach allowed us to validate the human origin of the paleofeces included in our study. It serves as a stepping stone for future studies on human paleofeces, which will help to decipher the unexplored diversity of ancient human gut microbiomes.

Machine learning methods to predict the source of microbiome samples

Estimating the contamination of a microbiome sample, an approach known as source tracking, has been a recurring challenge since the adoption of shotgun metagenomics methods. While different solutions have been proposed, relying on Naive Bayes (Greenberg et al., 2010), and Random Forest (Smith et al., 2010) algorithms, it is only with the appearance of SourceTracker (Knights et al., 2011) that a source tracking method became widely adopted by the community. SourceTracker relies on the latent dirichlet allocation (LDA) algorithm, a generative statistical model first introduced to infer genetic admixture (Pritchard et al., 2000), and for topic modelling (Blei et al., 2003). SourceTracker uses Gibbs sampling, a MCMC approach, to infer the Dirichlet posterior probability distribution. While it proved to be an effective method for this challenge, it also turned out to be relatively hard to scale because of the long time needed for convergence of the MCMC chains when using many sources. With the development of new faster data embedding techniques, such as t-SNE (van der Maaten, 2014), and UMAP (Lozupone et al., 2011), and new ecological distance metrics such as weighted unifrac (Lozupone et al., 2011), I seized the opportunity to develop a new faster source tracking discriminative method relying on a machine learning source mixture prediction in a lower dimensional space. Due to the marginally different objective of estimating the source of paleofeces, I also implemented the slightly different task of source prediction, akin to a task of classification. Sourcepredict, depending of the embedding method uses, can be either used for source tracking, using a linear dimension reduction method, such as principal coordinate analysis, or for source prediction, using a non-linear dimension reduction method, such as t-SNE or UMAP. While much faster than SourceTracker, it is expected that SourceTracker will remain relevant when using vastly different sources

because of its generative model. However, while the scaling of LDA based methods has been limited by the computational complexity of MCMC approaches, there have been new developments to speed up the posterior distribution sampling with the use of new neural network based variational inference methods such as prodLDA (Srivastava and Sutton, 2017), which hold great promises for the scalability of generative source tracking. In addition, whilst source tracking methods have so far relied on the abundance of taxons in a sample, new methods such as DECOM (González et al., 2023) using k-mer abundance could provide an alternative approach that might prove very relevant when dealing with samples of less studied sources.

The required scalability of metagenomics methods

With the ever decreasing price of sequencing thanks to the development of shotgun WGS, the amount of sequencing data has grown up massively, faster than the available computing capacity (Fig 0.2). Computational tools that were once used manually now have to be integrated in workflow managers to deal with the processing of many samples in parallel. Furthermore, in the field of metagenomics, the complexity can be quadratically increased. An example is taxonomic classification, where each sample needs to be compared to every available reference genome. To circumvent this exploding complexity, approaches applied to modern metagenomics samples have been relying on simplifying heuristics, such as the alignment free approaches of Kraken (Wood and Salzberg, 2014) for instance. However, in aDNA metagenomics, where an alignment step is still often required to compute a damage profile, these simplifying heuristics are not always applicable. Furthermore, even if the comparison step of query and reference sequences becomes computationally tractable thanks to these heuristics, the size of reference database and their in-memory indexing remains a challenge. With sam2lca, we introduced an approach based on a divide and conquer strategy in combination with a workflow manager like nextflow enabling a parallelization of the computation on all the computing nodes of a cluster, or a cloud. In the divide step, one can align each sample to each reference separately with a fast short read aligner, hence solving the issue of in-memory loading of whole reference genomes database encountered by tools such as MALT (Herbig et al., 2016). The conquer step is then performed downstream by sam2lca after all alignment files have been gathered. While the sam2lca approach solves the issue of in-memory loading of reference database, an alternative approach will be eventually needed in order to reduce the redundancy of the reference genomes, while preserving the sensitivity.

Approaches such as SPARSE (Zhou et al., 2018) have relied on using a representative genome after genome clustering to reduce the size of the reference database, however, with the multiplicity of genome rearrangements in bacteria, even genome clustering is not a viable enough long term solution. Fortunately, the progress in the field of sequence graph representation applied to reference sequence database have allowed a reduction of the index size by factors of up to 1000X, with very fast query times (Karasikov et al., 2020). Furthermore, sequence graph based approaches are much better equipped to deal with reference bias, as they encode all sequence variations in a single reference graph (Martiniano et al., 2020).

The ever decreasing cost of sequencing data also enabled the field of aDNA metagenomics to sequence samples ever deeper. And with deeper sequencing, aDNA metagenomics *de novo* became a reality (Wibowo et al., 2021), and tools to assess the amount of DNA damage were suddenly facing a scalability issue as well. MapDamage, as well as DamageProfiler were designed to assess the damage of reads aligned against a single reference, either with the appreciation of the damage left to interpretation of the user with a smiley plot, or with a bayesian MCMC, but slower, statistical model. Regardless of the method, these approaches both suffered from a scalability issue that either lied on the human side, or on the computational side. MapDamage and DamageProfiler quickly became inadapated to assess the damage of the many thousands of contigs typically generated by *de novo* assembly. Therefore, to develop a scalable approach, we decided to rely on a faster, albeit simpler heuristic. With a model having fewer parameters, and a parameter estimation relying on a convex optimization, pyDamage was now able to scale for the damage assessment of the many thousands of contigs generated by metagenomics *de novo* assembly, while maintaining a good damage prediction accuracy.

After contigs are assembled, the necessary next step to reconstruct genome from metagenomics *de novo* assembly is to perform binning, a form of clustering of the contigs into artificial genomes, the MAGs. Most binning methods cluster the MAGs based on short k-mer frequencies similarity, and evenness of coverage between contigs of the same bins. While these two metrics already create coherent MAGs in modern metagenomics *de novo* assemblies, aDNA samples could use the additional damage information to provide another dimension to help with the clustering of contigs into MAGs.

Finally, provided that there is enough coverage, pyDamage could be integrated with sam2lca to automatically provide a damage assessment at each taxon LCA.

Ancient microbiomes need not to be only human

Like many scientific fields, before gaining insights, and later forming theories, aDNA proceeds with a trial and error methodology. The study of the different aDNA material samples is a good illustration of this process. For example, there has been a long standing quest for the ideal human bone, the one preserving the most endogenous human DNA, but its only after many trial and error that the petrous bone was identified as one of the best candidate (Parker et al., 2020; Pinhasi et al., 2015). In parallel, for aDNA metagenomics, many different material types have been explored (Fig 3.1), to gain a general understanding of ancient microbiomes, but also to answer different research questions.

But one type of archeological artifact that remained so far unsuccessfully exploited by a DNA metagenomics are fermentation vessels. Because humans have been producing wine, and alcoholic beverages for at least 8000 years (Harutyunyan and Malfeito-Ferreira, 2022), there is a plethora of fermentation vessels among the different excavated archeological artefacts. Some of these vessels, like amphoras of the roman period, have been extensively studies for an epigraphic standpoint (Lorenzo et al., 2021), but until now, never using aDNA metagenomics methods. As very little was known about the potential microbial composition of these samples, we couldn't rely on source tracking approaches, with no available comparative samples being available. We therefore had to rely on other approaches, such as the differential abundance analysis to assess the preservation of our sample. Combining the tools developed earlier in this thesis, as well as established metagenomics approaches allowed us to reconstruct the genome of wine fermentation bacteria, and study them from a functional and phylogenetic standpoint. Our findings highlighted the potential presence of spoilage metabolites in Herodian wines, which further explains the Roman habit to better the flavour of their wine with a diversity of spices and aromas (Dodd, 2022). Since our approach relied on targeted capture enrichment to retrieve enough fermentation bacteria DNA to study their genome, we had to restrict ourselves to a limited list of known wine fermentation microbes. While this approach limits the discovery of entirely new microbes, it is however very efficient at recovering the genome of low abundance ancient microbes diluted among the more abundant DNA of contaminating modern microbes. Furthermore, in combination with already established methods for the study of ancient microbes and their products, such as proteomics, and metabolomics, there is a lot of potential to expand our knowledge on the ancient practices of alcoholic fermentation. Furthermore, this approach is not limited to alcoholic fermentation, but can also be applied to other types of non-alcoholic fermentation in food processing, that have also been a key cultural aspect

of ancient diets, which we nowadays benefit from on a daily basis (Steinkraus, 1997).

Conclusion

In this work, I developed a variety of new bioinformatics methods applied to ancient DNA metagenomics. With the current scale of data production, and what can be anticipated to be produced in the coming years, one of the main challenge that I tried to address was the scalability of these methods. I have demonstrated that some of the methods developed during this thesis can be applied for the identification of human paleofeces, and the reconstruction of ancient fermentation microbiome and their associated bacterial genomes from metagenomics WGS data.

Not only do the methods developed in this thesis allow for a better scalability, but the recent introduction of technological solutions such as more efficient workflow manager also allow for an easier scaling of already existing methods, and will allow to deal with the current amount of produced data in the near future.

Furthermore, a new generation of algorithms is coming to the field of metagenomics, for example with graph based representation of databases, and deep learning approaches showing even greater promises to deal with the ever growing influx of new aDNA data on a longer term.

In the four years of this thesis, both the field of aDNA metagenomics and its community saw a great expansion. Nevertheless this expansion hasn't been detrimental to the openness of the community. On the contrary, its blooming community sees flourishing exchanges between its members which, together with the technical developments, hold great promises for the future of the field.

References

- Aleberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11):1144–1146.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.
- Blanco-Miguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., Manghi, P., Dubois, L., Huang, K. D., Thomas, A. M., et al. (2022). Extending and improving metagenomic taxonomic profiling with uncharacterized species with MetaPhlan 4. *bioRxiv : the preprint server for biology*, pages 2022–08.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.
- Borry, M. (2019a). Sourcepredict example 1: Gut host species prediction — sourcepredict 0.3.2 documentation.
- Borry, M. (2019b). Sourcepredict example2: Estimating source proportions — sourcepredict 0.3.2 documentation.
- Borry, M. (2023). maxibor/adnamap: adnamap v1.0dev.
- Borry, M., Hübner, A., and Warinner, C. (2022). Sam2lca: Lowest Common Ancestor for SAM/BAM/CRAM alignment files. *Journal of Open Source Software*, 7(74):4360.
- Bos, K. I., Schuenemann, V. J., Golding, G. B., Burbano, H. A., Waglechner, N., Coombes, B. K., McPhee, J. B., DeWitte, S. N., Meyer, M., Schmedes, S., Wood, J., Earn, D. J. D., Herring, D. A., Bauer, P., Poinar, H. N., and Krause, J. (2011). A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature*, 478(7370):506–510.
- Breitwieser, F. P., Baker, D. N., and Salzberg, S. L. (2018). KrakenUniq: Confident and fast metagenomics classification using unique k-mer counts. *Genome Biology*, 19(1):198.
- Breitwieser, F. P., Lu, J., and Salzberg, S. L. (2017). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*.

- Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M. T., Lachmann, M., and Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences*, 104(37):14616–14621.
- Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10(1):421.
- Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2020). GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, 36(6):1925–1927.
- Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890.
- Compeau, P. E. C., Pevzner, P. A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29(11):987–991.
- Dodd, E. (2022). The Archaeology of Wine Production in Roman and Pre-Roman Italy. *American Journal of Archaeology*, 126(3):443–480.
- Drieu, L., Rageot, M., Wales, N., Stern, B., Lundy, J., Zerrer, M., Gaffney, I., Bondetti, M., Spiteri, C., Thomas-Oates, J., and Craig, O. E. (2020). Is it possible to identify ancient wine production using biomolecular approaches? *STAR: Science & Technology of Archaeological Research*, 6(1):16–29.
- Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLOS Computational Biology*, 7(10):e1002195.
- Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., and Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, 38(3):276–278.
- Fellows Yates, J. A., Andrades Valtueña, A., Vågane, Å. J., Cribdon, B., Velsko, I. M., Borry, M., Bravo-Lopez, M. J., Fernandez-Guerra, A., Green, E. J., Ramachandran, S. L., Heintzman, P. D., Spyrou, M. A., Hübner, A., Gancz, A. S., Hider, J., Allshouse, A. F., Zaro, V., and Warinner, C. (2021). Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir. *Scientific Data*, 8(1):31.
- Frantz, L. A. F., Mullin, V. E., Pionnier-Capitan, M., Lebrasseur, O., Ollivier, M., Perri, A., Linderholm, A., Mattiangeli, V., Teasdale, M. D., Dimopoulos, E. A., Tresset, A., Duffraisse, M., McCormick, F., Bartosiewicz, L., Gál, E., Nyerges, É. A., Sablin, M. V., Bréhard, S., Mashkour, M., Bălăşescu, A., Gillet, B., Hughes, S., Chassaing, O., Hitte, C., Vigne, J.-D., Dobney, K., Hänni, C., Bradley, D. G., and Larson, G. (2016). Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science*, 352(6290):1228–1231.
- Ginolhac, A., Rasmussen, M., Gilbert, M. T. P., Willerslev, E., and Orlando, L. (2011). mapDamage: Testing for damage patterns in ancient DNA sequences. *Bioinformatics (Oxford, England)*, 27(15):2153–2155.
- González, C. D., Vicedomini, R., Lemane, T., Rascovan, N., Richard, H., and Chikhi, R. (2023). decOM: Similarity-based microbial source tracking of ancient oral samples using k-mer-based methods.

- Granehäll, L., Huang, K. D., Tett, A., Manghi, P., Paladin, A., O'Sullivan, N., Rota-Stabelli, O., Segata, N., Zink, A., and Maixner, F. (2021). Metagenomic analysis of ancient dental calculus reveals unexplored diversity of oral archaeal *Methanobrevibacter*. *Microbiome*, 9(1):197.
- Greenberg, J., Price, B., and Ware, A. (2010). Alternative estimate of source distribution in microbial source tracking using posterior probabilities. *Water Research*, 44(8):2629–2637.
- Harutyunyan, M. and Malfeito-Ferreira, M. (2022). The Rise of Wine among Ancient Civilizations across the Mediterranean Basin. *Heritage*, 5(2):788–812.
- Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8.
- Herbig, A., Maixner, F., Bos, K. I., Zink, A., Krause, J., and Huson, D. H. (2016). MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman.
- Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A., and Wilson, A. C. (1984). DNA sequences from the quagga, an extinct member of the horse family. *Nature*, 312(5991):282–284.
- Huber, J. A., Mark Welch, D. B., Morrison, H. G., Huse, S. M., Neal, P. R., Butterfield, D. A., and Sogin, M. L. (2007). Microbial Population Structures in the Deep Marine Biosphere. *Science*, 318:97–100.
- Hübner, R., Key, F. M., Warinner, C., Bos, K. I., Krause, J., and Herbig, A. (2019). HOPS: Automated detection and authentication of pathogen DNA in archaeological remains. *Genome Biology*, 20(1):280.
- Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214.
- Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, 17(3):377–386.
- Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119.
- Jensen, T. Z. T., Niemann, J., Iversen, K. H., Fotakis, A. K., Gopalakrishnan, S., Vågene, Å. J., Pedersen, M. W., Sinding, M.-H. S., Ellegaard, M. R., Allentoft, M. E., Lanigan, L. T., Taurozzi, A. J., Nielsen, S. H., Dee, M. W., Mortensen, M. N., Christensen, M. C., Sørensen, S. A., Collins, M. J., Gilbert, M. T. P., Sikora, M., Rasmussen, S., and Schroeder, H. (2019). A 5700 year-old human genome and oral microbiome from chewed birch pitch. *Nature Communications*, 10(1):5520.
- Jiao, J.-Y., Liu, L., Hua, Z.-S., Fang, B.-Z., Zhou, E.-M., Salam, N., Hedlund, B. P., and Li, W.-J. (2021). Microbial dark matter coming to light: challenges and opportunities. *National Science Review*, 8(3):nwaa280.
- 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 (Oxford, England)*, 29(13):1682–1684.
- Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359.

- Karasikov, M., Mustafa, H., Danciu, D., Zimmermann, M., Barber, C., Räsch, G., and Kahles, A. (2020). MetaGraph: Indexing and Analysing Nucleotide Archives at Petabase-scale.
- Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., Bushman, F. D., Knight, R., and Kelley, S. T. (2011). Bayesian community-wide culture-independent microbial source tracking. *Nature Methods*, 8(9):761–763.
- Lederberg, J. and McCray, A. T. (2001). Ome sweetomics—a genealogical treasury of words. *The scientist*, 15(7):8–8.
- Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.
- Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, 362(6422):709–715.
- Lorenzo, B., Manuel, J., González, J. P., Rull, G., and Martín, A. A. (2021). Roman open data: a data visualization & exploratory interface for the academic training of university students. In *CINAIC*.
- Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., and Knight, R. (2011). UniFrac: An effective distance metric for microbial community comparison. *The ISME Journal*, 5(2):169–172.
- Marchesi, J. R. and Ravel, J. (2015). The vocabulary of microbiome research: A proposal. *Microbiome*, 3(1):31.
- Martiniano, R., Garrison, E., Jones, E. R., Manica, A., and Durbin, R. (2020). Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biology*, 21(1):250.
- McCliment, E. A., Voglesonger, K. M., O’Day, P. A., Dunn, E. E., Holloway, J. R., and Cary, S. C. (2006). Colonization of nascent, deep-sea hydrothermal vents by a novel Archaeal and Nanoarchaeal assemblage. *Environmental Microbiology*, 8(1):114–125.
- Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, 7(1):11257.
- Meyer, F., Fritz, A., Deng, Z.-L., Koslicki, D., Lesker, T. R., Gurevich, A., Robertson, G., Alser, M., Antipov, D., Beghini, F., Bertrand, D., Brito, J. J., Brown, C. T., Buchmann, J., Buluç, A., Chen, B., Chikhi, R., Clausen, P. T. L. C., Cristian, A., Dabrowski, P. W., Darling, A. E., Egan, R., Eskin, E., Georganas, E., Goltsman, E., Gray, M. A., Hansen, L. H., Hofmeyr, S., Huang, P., Irber, L., Jia, H., Jørgensen, T. S., Kieser, S. D., Klemetsen, T., Kola, A., Kolmogorov, M., Korobeynikov, A., Kwan, J., LaPierre, N., Lemaitre, C., Li, C., Limasset, A., Malcher-Miranda, F., Mangul, S., Marcelino, V. R., Marchet, C., Marijon, P., Meleshko, D., Mende, D. R., Milanese, A., Nagarajan, N., Nissen, J., Nurk, S., Olike, L., Paoli, L., Peterlongo, P., Piro, V. C., Porter, J. S., Rasmussen, S., Rees, E. R., Reinert, K., Renard, B., Robertsen, E. M., Rosen, G. L., Ruscheweyh, H.-J., Sarwal, V., Segata, N., Seiler, E., Shi, L., Sun, F., Sunagawa, S., Sørensen, S. J., Thomas, A., Tong, C., Trajkovski, M., Tremblay, J., Uritskiy, G., Vicedomini, R., Wang, Z., Wang, Z., Wang, Z.,

- Warren, A., Willassen, N. P., Yelick, K., You, R., Zeller, G., Zhao, Z., Zhu, S., Zhu, J., Garrido-Oter, R., Gastmeier, P., Hacquard, S., Häußler, S., Khaledi, A., Maechler, F., Mesny, F., Radutoiu, S., Schulze-Lefert, P., Smit, N., Strowig, T., Bremges, A., Sczyrba, A., and McHardy, A. C. (2022). Critical Assessment of Metagenome Interpretation: The second round of challenges. *Nature Methods*, 19(4):429–440.
- Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1986). Specific amplification of dna in vitro: the polymerase chain reaction. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 51, page 263.
- Neukamm, J., Peltzer, A., and Nieselt, K. (2021a). DamageProfiler: Fast damage pattern calculation for ancient DNA. *Bioinformatics*, 37(20):3652–3653.
- Neukamm, J., Peltzer, A., and Nieselt, K. (2021b). DamageProfiler: Fast damage pattern calculation for ancient DNA. *Bioinformatics*, 37(20):3652–3653.
- NIH (2022). DNA Sequencing Costs: Data.
- Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes: A new versatile metagenomic assembler. *Genome Research*, 27(5):824–834.
- Obregon-Tito, A. J., Tito, R. Y., Metcalf, J., Sankaranarayanan, K., Clemente, J. C., Ursell, L. K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P. M., Spicer, P., Lawson, P., Marin-Reyes, L., Trujillo-Villarreal, O., Foster, M., Gujja-Poma, E., Troncoso-Corzo, L., Warinner, C., Ozga, A. T., and Lewis, C. M. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. *Nature Communications*, 6(1):6505.
- Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., Stone, A. C., and Warinner, C. (2021). Ancient DNA analysis. *Nature Reviews Methods Primers*, 1(1):14.
- Ounit, R. and Lonardi, S. (2016). Higher classification sensitivity of short metagenomic reads with CLARK-S. *Bioinformatics*, 32(24):3823–3825.
- Pace, N. R. (1997). A Molecular View of Microbial Diversity and the Biosphere. *Science*, 276(5313):734–740.
- Parker, C., Rohrlach, A. B., Friederich, S., Nagel, S., Meyer, M., Krause, J., Bos, K. I., and Haak, W. (2020). A systematic investigation of human DNA preservation in medieval skeletons. *Scientific Reports*, 10(1):18225.
- Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., and Hugenholtz, P. (2022). GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794.
- Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M. C., Rice, B. L., DuLong, C., Morgan, X. C., Golden, C. D., Quince, C., Huttenhower, C., and Segata, N. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, 176(3):649–662.e20.
- Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428.

- Pinhasi, R., Fernandes, D., Sirak, K., Novak, M., Connell, S., Alpaslan-Roodenberg, S., Gerritsen, F., Moiseyev, V., Gromov, A., Raczky, P., Anders, A., Pietruszewski, M., Rollefson, G., Jovanovic, M., Trinhhoang, H., Bar-Oz, G., Oxenham, M., Matsumura, H., and Hofreiter, M. (2015). Optimal Ancient DNA Yields from the Inner Ear Part of the Human Petrous Bone. *PLOS ONE*, 10(6):e0129102.
- Poinar, H., Fiedel, S., King, C. E., Devault, A. M., Bos, K., Kuch, M., and Debruyne, R. (2009a). Comment on “DNA from Pre-Clovis Human Coprolites in Oregon, North America”. *Science*, 325(5937):148–148.
- Poinar, H., Fiedel, S., King, C. E., Devault, A. M., Bos, K., Kuch, M., and Debruyne, R. (2009b). Comment on “DNA from Pre-Clovis Human Coprolites in Oregon, North America”. *Science*, 325(5937):148–148.
- Poinar, H. N., Schwarz, C., Qi, J., Shapiro, B., MacPhee, R. D. E., Buigues, B., Tikhonov, A., Huson, D. H., Tomsho, L. P., Auch, A., Rampp, M., Miller, W., and Schuster, S. C. (2006). Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA. *Science*, 311(5759):392–394.
- Prescott, S. L. (2017). History of medicine: Origin of the term microbiome and why it matters. *Human Microbiome Journal*, 4:24–25.
- Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959.
- Proctor, L. M., Creasy, H. H., Fettweis, J. M., Lloyd-Price, J., Mahurkar, A., Zhou, W., Buck, G. A., Snyder, M. P., Strauss, J. F., Weinstock, G. M., White, O., Huttenhower, C., and The Integrative HMP (iHMP) Research Network Consortium (2019). The Integrative Human Microbiome Project. *Nature*, 569(7758):641–648.
- Reinhard, K., Camacho, M., Geyer, B., Hayek, S., Horn, C., Otterson, K., and Russ, J. (2019). Imaging coprolite taphonomy and preservation. *Archaeological and Anthropological Sciences*, 11(11):6017–6035.
- Renaud, G., Stenzel, U., and Kelso, J. (2014). leeHom: Adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Research*, 42(18):e141.
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467.
- Schnorr, S. L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., Turrone, S., Biagi, E., Peano, C., Severgnini, M., Fiori, J., Gotti, R., De Bellis, G., Luiselli, D., Brigidi, P., Mabulla, A., Marlowe, F., Henry, A. G., and Crittenden, A. N. (2014). Gut microbiome of the Hadza hunter-gatherers. *Nature Communications*, 5(1):3654.
- Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes*, 9:88.
- Schwengers, O., Jelonek, L., Dieckmann, M. A., Beyvers, S., Blom, J., and Goesmann, A. (2021). Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial Genomics*, 7(11):000685.
- Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., DeMaere, M. Z., Chikhi, R., Nagarajan, N., Quince, C., Meyer, F., Balvočiūtė, M.,

- Hansen, L. H., Sørensen, S. J., Chia, B. K. H., Denis, B., Froula, J. L., Wang, Z., Egan, R., Don Kang, D., Cook, J. J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y.-W., Singer, S. W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M. D., Lingner, T., Lin, H.-H., Liao, Y.-C., Silva, G. G. Z., Cuevas, D. A., Edwards, R. A., Saha, S., Piro, V. C., Renard, B. Y., Pop, M., Klenk, H.-P., Göker, M., Kyrpides, N. C., Woyke, T., Vorholt, J. A., Schulze-Lefert, P., Rubin, E. M., Darling, A. E., Rattei, T., and McHardy, A. C. (2017). Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nature Methods*, 14(11):1063–1071.
- Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)*, 30(14):2068–2069.
- Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8):811–814.
- Seitz, A. and Nieselt, K. (2017). Improving ancient DNA genome assembly. *PeerJ*, 5:e3126.
- Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, 5.
- Shenhav, L., Thompson, M., Joseph, T. A., Briscoe, L., Furman, O., Bogumil, D., Mizrahi, I., Pe’er, I., and Halperin, E. (2019). FEAST: Fast expectation-maximization for microbial source tracking. *Nature Methods*, 16(7):627–632.
- Skoglund, P., Northoff, B. H., Shunkov, M. V., Derevianko, A. P., Pääbo, S., Krause, J., and Jakobsson, M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences*, 111(6):2229–2234.
- Smith, A., Sterba-Boatwright, B., and Mott, J. (2010). Novel application of a statistical technique, Random Forests, in a bacterial source tracking study. *Water Research*, 44(14):4067–4076.
- Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., and Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proceedings of the National Academy of Sciences*, 103(32):12115–12120.
- Soo, R. M., Wood, S. A., Grzymalski, J. J., McDonald, I. R., and Cary, S. C. (2009). Microbial biodiversity of thermophilic communities in hot mineral soils of Tramway Ridge, Mount Erebus, Antarctica. *Environmental Microbiology*, 11(3):715–728.
- Srivastava, A. and Sutton, C. (2017). Autoencoding Variational Inference For Topic Models.
- Steinegger, M. and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028.
- Steinkraus, K. H. (1997). Classification of fermented foods: Worldwide review of household fermentation techniques. *Food Control*, 8(5):311–317.
- Stewart, R. D., Auffret, M. D., Warr, A., Walker, A. W., Roehe, R., and Watson, M. (2019). Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nature Biotechnology*, 37(8):953–961.

- Tito, R. Y., Knights, D., Metcalf, J., Obregon-Tito, A. J., Cleeland, L., Najar, F., Roe, B., Reinhard, K., Sobolik, K., Belknap, S., Foster, M., Spicer, P., Knight, R., and Lewis, C. M. (2012). Insights from Characterizing Extinct Human Gut Microbiomes. *PLoS ONE*, 7(12):e51146.
- Tito, R. Y., Macmil, S., Wiley, G., Najar, F., Cleeland, L., Qu, C., Wang, P., Romagne, F., Leonard, S., Ruiz, A. J., Reinhard, K., Roe, B. A., and Jr, C. M. L. (2008). Phylotyping and Functional Analysis of Two Ancient Human Microbiomes. *PLOS ONE*, 3(11):e3703.
- Tommaso, P. D., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows.
- Tringe, S. G. and Hugenholtz, P. (2008). A renaissance for the pioneering 16S rRNA gene. *Current Opinion in Microbiology*, 11(5):442–446.
- Tringe, S. G. and Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics*, 6(11):805–814.
- Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nature methods*, 12(10):902–903.
- van der Maaten, L. (2014). Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*, 15(93):3221–3245.
- Warinner, C., Rodrigues, J. F. M., Vyas, R., Trachsel, C., Shved, N., Grossmann, J., Radini, A., Hancock, Y., Tito, R. Y., Fiddyment, S., Speller, C., Hendy, J., Charlton, S., Luder, H. U., Salazar-García, D. C., Eppler, E., Seiler, R., Hansen, L. H., Castruita, J. A. S., Barkow-Oesterreicher, S., Teoh, K. Y., Kelstrup, C. D., Olsen, J. V., Nanni, P., Kawai, T., Willerslev, E., von Mering, C., Lewis, C. M., Collins, M. J., Gilbert, M. T. P., Rühli, F., and Cappellini, E. (2014). Pathogens and host immunity in the ancient human oral cavity. *Nature Genetics*, 46(4):336–344.
- Whipps, J. M., Lewis, K., and Cooke, R. (1988). Mycoparasitism and plant disease control.
- Wibowo, M. C., Yang, Z., Borry, M., Hübner, A., Huang, K. D., Tierney, B. T., Zimmerman, S., Barajas-Olmos, F., Contreras-Cubas, C., García-Ortiz, H., Martínez-Hernández, A., Lubner, J. M., Kirstahler, P., Blohm, T., Smiley, F. E., Arnold, R., Ballal, S. A., Pamp, S. J., Russ, J., Maixner, F., Rota-Stabelli, O., Segata, N., Reinhard, K., Orozco, L., Warinner, C., Snow, M., LeBlanc, S., and Kostic, A. D. (2021). Reconstruction of ancient microbial genomes from the human gut. *Nature*, pages 1–6.
- Willerslev, E., Hansen, A. J., Binladen, J., Brand, T. B., Gilbert, M. T. P., Shapiro, B., Bunce, M., Wiuf, C., Gilichinsky, D. A., and Cooper, A. (2003). Diverse Plant and Animal Genetic Records from Holocene and Pleistocene Sediments. *Science*, 300(5620):791–795.
- Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews*, 51(2):221–271.
- Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1):257.
- Wood, D. E. and Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46.

- Wu, Y.-W., Simmons, B. A., and Singer, S. W. (2016). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4):605–607.
- Yatsunenکو, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., Knight, R., and Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227.
- 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 Raphael, B. J., editor, *Research in Computational Molecular Biology*, Lecture Notes in Computer Science, pages 225–240, Cham. Springer International Publishing.

Abbreviations

- aDNA:** Ancient DNA
- BER:** Base Excision Repair
- CDS:** CoDing Sequence
- GTDB:** Genome Taxonomy DataBase
- HMM:** Hidden Markov Model
- LCA:** Lowest Common Ancestor
- LRT:** Likelihood Ratio Test
- MAG:** Metagenome Assembled Genome
- MCMC:** Markov Chain Monte Carlo
- PCR:** Polymerase Chain Reaction
- rRNA:** Ribosomal RNA
- TAXID:** TAXonomic IDentifier
- WGS:** Whole Genome Shotgun

Eidesstattliche Erklärung

Maxime Borry

Max Planck Institute for Evolutionary Anthropology, und Friedrich-Schiller-Universität
Jena

Hiermit erkläre ich,

- (a) dass mir die geltende Promotionsordnung bekannt ist,
- (b) dass ich die Dissertation selbst angefertigt habe, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen in der Arbeit angegeben habe,
- (c) dass ich alle Personen, die mich bei der Auswahl und Auswertung sowie bei der Herstellung des Manuskriptes unterstützt haben, in der Autorenliste der Manuskripte und den entsprechenden Danksagungen namentlich erwähnt habe,
- (d) dass ich nicht die Hilfe einer kommerziellen Promotionsvermittlung in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen,
- (e) dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe

Unterschrift:

Maxime Borry

Detailed contributions

Manuscript B**Short reference:** Borry et al. (2020), PeerJ**Contribution of the doctoral candidate**

Contribution of the doctoral candidate to figures reflecting experimental data (only for original articles):

Figure(s) # 1*	<input type="checkbox"/> 100% (the data presented in this figure come entirely from experimental work carried out by the candidate) <input checked="" type="checkbox"/> 0% (the data presented in this figure are based exclusively on the work of other co-authors) <input type="checkbox"/> Approximate contribution of the doctoral candidate to the figure: _____% Brief description of the contribution: (e.g. "Figure parts a, d and f" or "Evaluation of the data" etc.)
* Can refer to more than one fig. if the answer is the same	

Figure(s) # 2-8*	<input checked="" type="checkbox"/> 100% (the data presented in this figure come entirely from experimental work carried out by the candidate) <input type="checkbox"/> 0% (the data presented in this figure are based exclusively on the work of other co-authors) <input type="checkbox"/> Approximate contribution of the doctoral candidate to the figure: _____% Brief description of the contribution: (e.g. "Figure parts a, d and f" or "Evaluation of the data" etc.)
* Can refer to more than one fig. if the answer is the same	

(Add more table boxes depending on the number of figures)

Signature candidate_____
Signature supervisor (member of the Faculty)

² The signatures must be original only in the completed form to be submitted separately to the Dean's Office. The signatures and signature fields are not necessarily required in the version included in the dissertation.

Manuscript C**Short reference** Borry et al. (2020), JOSS**Contribution of the doctoral candidate**

Contribution of the doctoral candidate to figures reflecting experimental data (only for original articles):

Figure(s) # 1	<input checked="" type="checkbox"/> 100% (the data presented in this figure come entirely from experimental work carried out by the candidate)
	<input type="checkbox"/> 0% (the data presented in this figure are based exclusively on the work of other co-authors)
	<input type="checkbox"/> Approximate contribution of the doctoral candidate to the figure: _____% Brief description of the contribution: <i>(e.g. "Figure parts a, d and f" or "Evaluation of the data" etc.)</i>

** Can refer to more than one fig. if the answer is the same*

(Add more table boxes depending on the number of figures)

Signature candidate_____
Signature supervisor (member of the Faculty)

Manuscript D**Short reference** Borry et al. (2021), PeerJ**Contribution of the doctoral candidate**

Contribution of the doctoral candidate to figures reflecting experimental data (only for original articles):

Figure(s) # 1-2	<input type="checkbox"/> 100% (the data presented in this figure come entirely from experimental work carried out by the candidate) <input checked="" type="checkbox"/> 0% (the data presented in this figure are based exclusively on the work of other co-authors) <input type="checkbox"/> Approximate contribution of the doctoral candidate to the figure: _____% Brief description of the contribution: <i>(e.g. "Figure parts a, d and f" or "Evaluation of the data" etc.)</i>
<p><i>* Can refer to more than one fig. if the answer is the same</i></p>	

Figure(s) # 3-4	<input type="checkbox"/> 100% (the data presented in this figure come entirely from experimental work carried out by the candidate) <input type="checkbox"/> 0% (the data presented in this figure are based exclusively on the work of other co-authors) <input checked="" type="checkbox"/> Approximate contribution of the doctoral candidate to the figure: 30% Brief description of the contribution: <i>Implementation of the software</i>
<p><i>* Can refer to more than one fig. if the answer is the same</i></p>	

Figure(s) # 5-8	<input checked="" type="checkbox"/> 100% (the data presented in this figure come entirely from experimental work carried out by the candidate) <input type="checkbox"/> 0% (the data presented in this figure are based exclusively on the work of other co-authors) <input type="checkbox"/> Approximate contribution of the doctoral candidate to the figure: 30% Brief description of the contribution:
------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

² The signatures must be original only in the completed form to be submitted separately to the Dean's Office. The signatures and signature fields are not necessarily required in the version included in the dissertation.

** Can refer to more than
one fig. if the answer is
the same*

(Add more table boxes depending on the number of figures)

Signature candidate

Signature supervisor (member of the Faculty)

Manuscript E**Short reference** Borry (2023), Unpublished**Contribution of the doctoral candidate**

Contribution of the doctoral candidate to figures reflecting experimental data (only for original articles):

Figure(s) # 1	<input type="checkbox"/> 100% (the data presented in this figure come entirely from experimental work carried out by the candidate) <input checked="" type="checkbox"/> 0% (the data presented in this figure are based exclusively on the work of other co-authors) <input type="checkbox"/> Approximate contribution of the doctoral candidate to the figure: _____% Brief description of the contribution: <i>(e.g. "Figure parts a, d and f" or "Evaluation of the data" etc.)</i>
<p><i>* Can refer to more than one fig. if the answer is the same</i></p>	

Figure(s) # 2-5	<input checked="" type="checkbox"/> 100% (the data presented in this figure come entirely from experimental work carried out by the candidate) <input type="checkbox"/> 0% (the data presented in this figure are based exclusively on the work of other co-authors) <input type="checkbox"/> Approximate contribution of the doctoral candidate to the figure: _____% Brief description of the contribution: <i>(e.g. "Figure parts a, d and f" or "Evaluation of the data" etc.)</i>
<p><i>* Can refer to more than one fig. if the answer is the same</i></p>	

(Add more table boxes depending on the number of figures)

Signature candidate_____
Signature supervisor (member of the Faculty)