HAYSTAC: A Bayesian framework for robust and rapid species identification in high-throughput sequencing data

Evangelos A. Dimopoulos[1,*,✉], **Alberto Carmagnini**[2,*], **Irina M. Velsko**[3], **Christina Warinner**[3,4], **Greger Larson**[1], **Laurent A. F. Frantz**[2,5,#,✉], **and Evan K. Irving-Pease**[1,6,#,✉]

[1]The Palaeogenomics and Bio-archaeology Research Network, Research Laboratory for Archaeology and History of Art, University of Oxford, Oxford, UK.
[2]School of Biological and Chemical Sciences, Queen Mary University of London, London, UK.
[3]Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany.
[4]Department of Anthropology, Harvard University, Cambridge, USA.
[5]Palaeogenomics Group, Department of Veterinary Sciences, Ludwig Maximilian University, Munich, Germany.
[6]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark
[*]Contributed equally
[#]Co-supervised this work

Identification of specific species in metagenomic samples is critical for several key applications, yet many tools available require large computational power and are often prone to false positive identifications. Here we describe High-AccuracY and Scalable Taxonomic Assignment of MetagenomiC data (HAYSTAC), which can estimate the probability that a specific taxon is present in a metagenome. HAYSTAC provides a user-friendly tool to construct databases, based on publicly available genomes, that are used for competitive reads mapping. It then uses a novel Bayesian framework to infer the abundance and statistical support for each species identification, and provide per-read species classification. Unlike other methods, HAYSTAC is specifically designed to efficiently handle both ancient and modern DNA data, as well as incomplete reference databases, making it possible to run highly accurate hypothesis-driven analyses (*i.e.*, assessing the presence of a specific species) on variably sized reference databases while dramatically improving processing speeds. We tested the performance and accuracy of HAYSTAC using simulated Illumina libraries, both with and without ancient DNA damage, and compared the results to other currently available methods (*i.e.*, Kraken2/Braken, MALT/HOPS, and Sigma). HAYSTAC identified fewer false positives than both Kraken2/Braken and MALT/HOPS in all simulations, and fewer than Sigma in simulations of ancient data. It runs faster than Sigma for large databases, uses less memory than Kraken2/Braken, and is both faster and uses less memory than MALT/HOPS. Lastly, we used HAYSTAC to search for specific pathogens in two published ancient metagenomic datasets, demonstrating how it can be applied to empirical datasets. HAYSTAC is available from `https://github.com/antonisdim/HAYSTAC`

metagenomics | taxonomic classification | pathogens | ancient DNA

Correspondence: ***evangelos-antonios.dimopoulos@arch.ox.ac.uk, laurent.frantz@lmu.de, evan.irvingpease@gmail.com***

Introduction

Metagenomics allows high-throughput sequencing of complex microbial communities that may not be possible to grow in a laboratory. Accurate identification of specific species within a complex community is critical for metagenomic applications in a wide variety of fields, including medicine, microbiome sciences, environmental sciences, biosurveillance, and biomolecular archaeology. Confidently assessing the presence of individual species is also crucial for taxonomic profiling (Gonzalez et al., 2016), inferring genomic evolution and phylogenetics (Tett et al., 2019), tracing disease outbreaks (Ahn et al., 2015), generating diagnostics from biofluids (Wilson et al., 2019) and studying ancient pathogens (Spyrou et al., 2019).

Recent studies of ancient pathogens based on high-throughput sequencing technology have provided unique insights into the evolutionary history of major human pathogens. These include the etiologic agents of both the Justinian and Black Death plagues (*Yersinia pestis*; (Bos et al., 2011; Schuenemann et al., 2011; Spyrou et al., 2016; Rasmussen et al., 2015)), tuberculosis (the *Mycobacterium tuberculosis* complex; (Pósa et al., 2015; Müller et al., 2014)), syphilis (*Treponema pallidum*; (Schuenemann et al., 2018; Arora et al., 2016)) and smallpox (*Variola major* and *Variola minor*; (Mühlemann et al., 2020)).

Accurately identifying specific microbial species from ancient metagenomes, however, can be challenging for numerous reasons. Firstly, short read sequences originating from closely related species may be difficult to distinguish from each other. In addition, the species of interest may have low abundance in the sample, or the genomes in the reference database may not be representative of the strain present in a sample (Warinner et al., 2017). Contamination from the laboratory environment or reagents may also obscure identification (Davis et al., 2018). Identifying specific species within ancient metagenomes is further complicated by issues inherent to ancient DNA, including short molecules (typically between 30–60 bp; (Kistler et al., 2017)), post-mortem miscoding lesions (e.g., cytosine deamination), contamination from soil and sediment bacteria in the burial environment, sample handling, and storage (Mann et al., 2018).

DRAFT

When present in ancient metagenomes, pathogen DNA is usually found in low abundance (Warinner et al., 2017), and researchers often initially screen many samples before deciding which libraries should be more deeply sequenced in order to obtain whole genome data, or be further processed using targeted capture techniques (Borówka et al., 2019; Key et al., 2017; Spyrou et al., 2019). Screening for pathogens is typically performed by shallow shotgun sequencing followed by *in silico* analyses of short read data using taxonomic classifiers. Following screening, libraries with a positive identification for a species of interest are then more deeply sequenced and investigated, with or without targeted capture (Spyrou et al., 2019). Confident species identification from metagenomic screening data is therefore critical for studies of ancient pathogens.

Tools for detecting microbial species from metagenomes include metagenomic classifiers, that report taxonomic relative abundances (Ye et al., 2019), such as Kraken2 (Wood and Salzberg, 2014; Wood et al., 2019) and the Megan Alignment Tool (MALT) (Vågene et al., 2018), as well as species identification pipelines, that classify individual reads, such as Sigma (Ahn et al., 2015), Strain Prediction and Analysis using Representative Sequences (SPARSE) (Zhou et al., 2018), and Heuristic Operations for Pathogen Screening (HOPS, specifically designed for ancient DNA) (Hübler et al., 2019). The metagenomic classifiers Kraken2 and MALT provide a good starting point for exploratory studies, especially when it is not known which pathogens may be present (Naccache et al., 2014). These tools, however, can be slow, require a large amount of RAM (Hübler et al., 2019; Vågene et al., 2018), and are prone to false positive identifications of low abundance species (Velsko et al., 2018). The species identification pipelines Sigma (Ahn et al., 2015) and SPARSE (Zhou et al., 2018) are based on more sophisticated statistical models but they require alignment to a large database, which can be slow (Zhou et al., 2018), and they do not provide statistical support for individual species identifications.

To address these issues, we developed High-AccuracY and Scalable Taxonomic Assignment of MetagenomiC data (HAYSTAC - `https://github.com/antonisdim/HAYSTAC`), a light-weight, fast, and user-friendly species identification tool. HAYSTAC evaluates the presence of a particular species of interest in a metagenomic sample, and provides statistical support for the species assignment. Our method is designed to estimate the probability that a specific taxon is present in a metagenomic sample given a set of sequencing reads and a database of reference genomes. We first derive the probability that the set of reads has originated from a single species (single source identification). This is useful, for example, for taxonomic identification of unknown source material (e.g. (Prendergast et al., 2017)). We then expand this method to identify multiple species in a metagenomic sample (multiple source identification).

To apply HAYSTAC, the user begins by constructing a database from publicly available genomes via a user-friendly automated process. Sample reads are then mapped to all reference genomes in the database using a Bowtie2 wrapper (Langmead and Salzberg, 2012). HAYSTAC then applies a novel Bayesian framework to infer abundance and statistical support for each species identification, and provides a per-read classification, allowing for both hypothesis-driven and exploratory analyses. HAYSTAC can build and make use of arbitrarily sized databases, with minimal effects on sensitivity and specificity. This means that HAYSTAC can provide reliable species identifications by aligning reads to a handful of reference genomes (instead of thousands), without misidentifying taxa because of the restricted database size. This feature dramatically improves its performance relative to other methods. Altogether, HAYSTAC provides a robust method to assess the presence of a specific species in a metagenomic sample in a single step without the need to run multiple programs to assess assignment accuracy, such as is required by other methods (*i.e.*, Bracken for Kraken2 (Lu et al., 2017) and HOPS for MALT).

To evaluate HAYSTAC's performance, we applied it to both simulated microbiome datasets and to empirical datasets that were generated from DNA extracts derived from archaeological human bone. We further used HAYSTAC to assess the presence of putative respiratory pathogens in oral microbiome-derived metagenomes from ancient dental calculus. While dental plaque is known to harbor respiratory pathogens (Sumi et al., 2007), few studies to date have reported these species in its mineralized analogue, dental calculus (Warinner et al., 2014; Eerkens et al., 2018), a discrepancy that may reflect identification biases in studies of the latter. Improved identification of ancient respiratory pathobionts (host-associated microbial taxa that may cause disease under specific conditions) has the potential to add to our understanding of pathogen evolutionary dynamics and even aid vaccine development.

DRAFT

Results

Method Overview

HAYSTAC is a metagenomic species identifier that works with arbitrary user defined databases. The database can be built from any combination of an NCBI search query, a user specified list of NCBI accession codes, the RefSeq representative database of prokaryotic species, or a list of user provided reference assemblies. To construct the database, HAYSTAC uses a heuristic method to select a single representative genome per taxon from those available in the input data.

A database-wide index is then built, followed by a competitive alignment (very fast local alignment mode in Bowtie2, with default parameters) of each sample against the full database. This step ensures that reads with no matches in the database are excluded in all subsequent analyses ("dark matter" category; Figure 1).

Reads passing this initial inclusion filter are then aligned against every individual genome in the database (with

Figure 1. HAYSTAC's workflow. HAYSTAC consists of three main modules: (i) DATABASE, which builds a database of reference genomes from various user input sources; (ii) SAMPLE, which handles downloading sequencing files from the SRA and pre-processing of samples prior to analysis; and (iii) ANALYSE, which performs an analysis of a sample against a database by applying the mathematical model (see methods) for taxonomic abundance estimation.

Bowtie2 in end-to-end mode). For each alignment, we count the number of transitions (T_s) and transversions (T_v), as each type of mismatch contributes differently to our statistical model, in order to handle ancient DNA damage. We then compute a posterior probability for each read/taxon alignment pair, using both mismatch count and expected mismatch probability (σ, default of 0.05). Accuracy can be improved by adjusting the mismatch probability (σ) using prior information on intra- and interspecific variability in the specific set of species analysed.

Reads that possess a posterior probability exceeding a user-defined threshold (default of 0.75, Supplemental Figure 5 and 6) are then assigned to a single taxon. Reads with a lower posterior probability are assigned to the "grey matter" category. For each taxon, the count of reads passing the posterior probability threshold are used to compute the mean posterior abundance (and 95% confidence interval); *i.e.*, the relative abundance of a taxon in a sample, using a Dirichlet distribution of the sample input data as a whole. A final validation step, for low depth sequencing data, using breadth of coverage is then used to assess if the assigned reads represent a random genome sample or cluster around specific genome areas.

Computational Performance

We assessed the computational performance of HAYSTAC relative to Sigma, Kraken2/Bracken and MALT/HOPS by measuring peak memory usage (maximum resident set size) and elapsed execution time (wall clock) for each software package on a machine with 16 CPU cores and 64 GB of RAM under controlled conditions. To control for variability in each measurement, we ran five independent replicates of each job, and performed a linear regression across the observed values.

We first assessed performance when building databases containing 10, 100 and 500 species. For all tested sizes, HAYSTAC ran faster and used less memory than either MALT/HOPS or Kracken2/Bracken, but was slower and used more memory than Sigma, which does not require large indices for its alignments (Figure 2; Supplemental Table S9).

We then assessed the performance of these methods when analysing a one million read dataset against the same three database sizes. HAYSTAC runs faster than Sigma for large databases ($\geq$ 500 species), uses less memory than Kraken2/Braken, and is both faster and uses less memory than MALT/HOPS for all database sizes (Figure 2; Supplemental Table S10).

Lastly, we assessed the performance of these methods when analysing datasets of 10 thousand, 100 thousand and 1 million reads, against a database of 500 species (Supplemental Figure 1; Supplemental Table S10). Memory performance was unchanged from the earlier benchmarks, showing that for all methods, peak memory usage scales with database size rather than sample size. With the exception of Kraken2/Bracken, all other methods saw only a modest increase in runtime with larger input sizes, showing that database size also has a substantial impact on overall runtime.

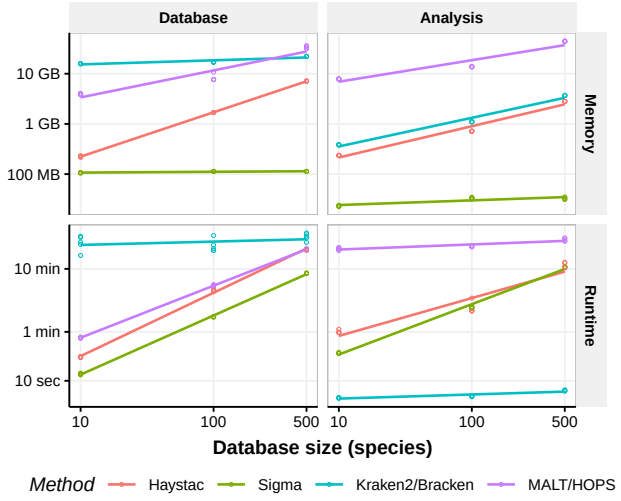

Figure 2. Computational performance of HAYSTAC and other methods. Linear regression of the elapsed time (wall clock) and peak memory usage (maximum resident set size) for a sample of size 1 million reads and reference databases containing either 10, 100 or 500 genomes, each with 5 replicates. When constructing the database, HAYSTAC uses substantially less memory and runs faster than either Kraken2/Bracken or MALT/HOPS. When performing analyses, HAYSTAC uses less memory than either Kraken2/Bracken or MALT/HOPS and runs faster than MALT/HOPS. For databases of 500 species or larger, HAYSTAC also runs faster than Sigma.

Analyses of simulated data sets using RefSeq prokaryotic database

To characterise the sensitivity and specificity of HAYSTAC, we compared its performance to that of three other widely used metagenomic profiling tools: Sigma, Kraken2/Bracken and MALT/HOPS. We simulated four datasets for these comparisons (Table 1), and then profiled each dataset with all four tools. Each simulated dataset was designed to test a different aspect of performance. For each test, we computed: (i) false positive count (total number of species that were reported but not present in the simulated sample); (ii) false negative count (total number of species present in the simulation that were not identified); and (iii) true positive count (total number of species correctly identified).

We first generated a dataset of samples comprising 10 species to quantify the ability of each program to accurately detect species in a simple metagenomic community (hereafter "Simple Microbiome" dataset). We generated two additional sets of profiles for this dataset: one that included modern human DNA ("contaminated") and one that did not ("non-contaminated"). This dataset was created using gargammel (Renaud et al., 2017) by randomly sampling 10 species from a modified RefSeq prokaryote representative genome database (see Methods) and building simulated libraries of one million single-end reads with an equal number of reads per species (100K), both with and without contamination from the human reference genome (25% of the total simulated reads).

All tools correctly identified the 10 species in each simulated sample set (both with and without human contamination; Figure 3). Only two programs reported false positives in the non-contaminated sample set: Kraken2/Bracken (average of three species) and MALT/HOPS (average of 0.5 species) (see methods for the definition of false positive identifications). False positive rates, however, dramatically increased in the sample set with human contamination. HAYSTAC reported the lowest false positive count (5) relative to Sigma (6), Kraken2/Bracken (12.5), and MALT/HOPS (11.5) (Figure 3). This suggests that HAYSTAC is less prone to false positive identifications caused by human contamination in bacterial reference genomes found in the RefSeq representative prokaryotic database. Across all tested programs for the contaminated simulation set, each false positive identification was to a species whose RefSeq genomes are known to contain human DNA contamination (Breitwieser et al., 2019) (Supplemental Table S7). This result strongly supports the prefiltering and removal of all human reads prior to microbial metagenomic profiling.

We then tested HAYSTAC's performance when profiling samples with damage patterns characteristic of aDNA (*i.e.*, sequence fragmentation and cytosine deamination). To do so, we generated eight simulated samples consisting of 10 mil-

DRAFT

Table 1. Characteristics of the six simulated samples sets of the Simple, Random and Oral microbiome datasets.

Simulation set	Dataset	Number of replicates	aDNA damage	Mean number of species
Simple with human DNA	Simple Microbiome	2	No	10
Simple without human DNA	Simple Microbiome	2	No	10
Microbiome 100 species ancient	General Microbiome	2	Yes	100
Microbiome 100 species modern	General Microbiome	2	No	100
Microbiome 500 species ancient	General Microbiome	2	Yes	500
Microbiome 500 species modern	General Microbiome	2	No	500
Oral microbiome aDNA damage	Oral Microbiome	6	Yes	178
Oral microbiome modern	Oral Microbiome	6	No	178

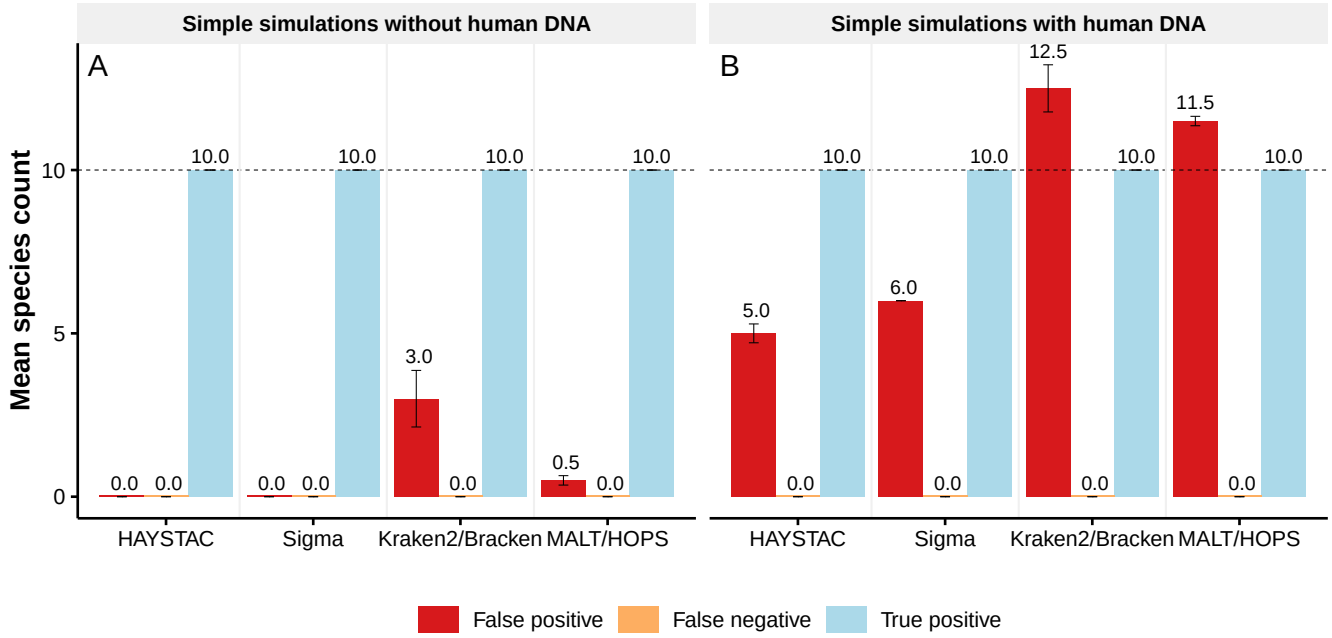

Figure 3. Accuracy of HAYSTAC and other methods for a simple simulation. Bar plot showing the mean count of false positives (red), false negatives (orange), and true detected species (blue) for two versions of the simple simulations dataset, each with two replicates: (A) without human DNA contamination ($n = 2$); and (B) with human DNA contamination ($n = 2$). The dotted line shows the number of simulated species in each set of samples (*i.e.*, the maximum true positive; $n = 10$), and numbers above the error bars indicate the mean species count in each category. For the simulation without human contamination, HAYSTAC outperforms both Kraken2/Bracken and MALT/HOPS, and performs equally with Sigma (*i.e.*, no false positives or false negatives). For the simulation with human contamination, HAYSTAC outperforms all three other methods. All of the 5 false positive species identified by HAYSTAC are known to contain human sequences in their reference genomes (Breitwieser et al., 2019), confounding any analyses which do not explicitly filter human contamination.

lion reads each, both with and without aDNA damage (hereafter "General Microbiome" dataset). Each sample contained simulated reads from either 100 species (four simulated samples; two random sets of species all with equal abundance) or 500 species (four samples; two random sets of species all with equal abundance) that were randomly sampled from the RefSeq representative prokaryotic database (Table1).

Overall, the incidence of false negatives was low. HAYSTAC reported 2.5 false negative species on average in the 100 species sample set, and 15.0 in the 500 species sample set. Sigma reported 1.5 and 14.0, Kraken2/Bracken 2.3 and 22.0 and MALT/HOPS 7.8 and 38.0, respectively (Supplemental Figure 2 and Supplemental Figure 3 A&B). False positives, in contrast, were variable between methods. HAYSTAC reported an average of only 0.5 false positives in the 100 species sample set, and an average of 1.0 false positives in the 500 species sample set. Sigma performed similarly, with an average of 0.5 and 1.0 false positives, while Kraken2/Bracken reported 11.5 and 17.0 false positives, and MALT/HOPS reported 27.5 and 25.5 false positives in the 100 and 500 species sample sets, respectively (Supplemental Figure 3 A&B and Supplemental Table S3).

We next performed more realistic simulations based on species compositions that mimic an oral microbial community from a previously analysed subgingival plaque sample ((Velsko et al., 2018); hereafter "Oral Microbiome" dataset). These simulations included 12 simulated samples of 5 million paired-end reads each, with and without aDNA damage, and with 176–180 species in varying or constant species abundance (Table1). Only 115 of the 196 total Oral Microbiome species had taxonomic representation in the RefSeq representative prokaryotic database, and only 4 species in the Oral Microbiome dataset had genomes in the modified representative prokaryotic RefSeq database we used (Supplemen-

Table 2. False negative and false positive species counts in the ancient and modern Oral Microbiome datasets.

	Oral microbiome aDNA damage		Oral microbiome modern	
	False negative	**False positive**	**False negative**	**False positive**
HAYSTAC	71.8	30.7	70.5	33.5
Sigma	67.5	44.2	67.5	31.5
Kraken2/Bracken	67.3	174.2	67.2	227.8
MALT/HOPS	69.0	115.3	68.3	325.8

tal Table S6).

As expected, given the incompleteness of the database, false negatives were elevated overall (Table 2). HAYSTAC failed to identify on average 71.8 species in the ancient (damaged) Oral Microbiome dataset, and 70.5 species in the modern Oral Microbiome dataset. Sigma failed to identify on average 67.5 and 67.5 species, Kraken2/Bracken failed to identify 67.3 and 67.2 species, and MALT/HOPS failed to identify 69.0 and 68.3 species in the ancient and modern Oral Microbiome datasets, respectively (Figure 4 A&B and Supplemental Figure 4). False positives, however, were highly variable between methods. HAYSTAC reported an average of 30.7 false positives in the ancient Oral Microbiome dataset and 33.5 false positives in the modern Oral Microbiome dataset. Sigma reported 44.2 and 31.5 false positives, Kraken2/Bracken reported 174.2 and 227.8 false positives, and MALT/HOPS reported 115.3 and 325.8 false positives in the ancient and modern datasets, respectively (Figure 4 A&B and Supplemental Table S4). Overall HAYSTAC outperformed the other methods in the ancient Oral Microbiome sample set with the lowest number of false positive identifications.

Analyses of simulated data sets using genus-level database

We next assessed how accurately HAYSTAC identified specific candidate species using smaller databases. To do so, we re-analysed 20 simulated samples included in the General and Oral Microbiome datasets (General Microbiome 100 species, General Microbiome 500 species, Oral Microbiome aDNA damage, and Oral Microbiome modern datasets) (Table1) using nine different genus-specific databases (see Methods). For each randomly selected genus, we constructed a database that included all genomes found in the RefSeq representative prokaryotic database. The nine genera that were selected for this analysis were: *Bacteroides*, *Burkholderia*, *Campylobacter*, *Clostridium*, *Corynebacterium*, *Desulfitobacterium*, *Mycobacterium*, *Solimonas*, and *Streptococcus*. Each database consisted of all RefSeq genomes from all species belonging to the genus instead of the full RefSeq representative prokaryotic database. This resulted in databases with between 3–136 species each, as opposed to $5,636$ species when using the full RefSeq representative prokaryotic database. For each analysis, we computed false positive, false negative, and true positive species counts as before.

Overall, false negatives were low. HAYSTAC reported on average 0.2 species missing in the samples of the General and Oral Microbiome datasets, while Sigma reported 0.1, Kraken2/Bracken 0.1 and MALT/HOPS 0.1 species missing (Figure 5, Supplemental Table S5). False positives, however, were variable between methods. HAYSTAC reported an average of 0.8 false positives, averaged across all samples in the General and Oral Microbiome datasets, while Sigma reported 1.1, Kraken2/Bracken 33.1, and MALT/HOPS 3.6 (Figure 5). We also used HAYSTAC to compare the mean posterior abundance computed using the full RefSeq database with abundance computed using the genus-specific databases (Figure 6). We found that while genus-specific abundances were slightly higher than those computed with the full RefSeq database, there was a very strong correlation between the two ($R^2 = 0.985$; slope$= 1.1$; p-value $< 10^{-6}$; GLM: $y = 1.1x + 8.2 \cdot 10^{-6}$).

Case study 1: Ancient *Yersinia pestis*

We then tested HAYSTAC's performance to detect specific pathogens in real archaeological samples. We first analysed 130 publicly available plague-positive ancient human libraries (Rasmussen et al., 2015), obtained from seven bone and tooth samples (RISE00, RISE139, RISE386, RISE397, RISE505, RISE509 and RISE511).

HAYSTAC was run against a database consisting of the longest complete genomes for each of the species in the genus *Yersinia* (including plasmid sequences). Our method was able to confidently assign between $3,875$ and $550,200$ reads to *Y. pestis* across all seven samples (Figure 7). In addition, we found evidence for additional *Yersinia* species in all samples, including *Y. pseudotuberculosis* (311–$44,818$ reads), *Y. similis* (158–$6,263$ reads), *Y. canariae* (2–$6,029$ reads) and *Y. ruckeri* (117–566 reads). Of these, *Y. ruckeri* and *Y. similis* are known to cause infections in animals (specifically fish), but have also been reported to infect humans (De Keukeleire et al., 2014).These identifications, however, are likely surprious, since *Y. ruckeri* and *Y. similis* are phylogenetically basal to other *Yersinia* species (Reuter et al., 2014; Chen et al., 2010) and may be attracting reads belonging diverged sequences.

The posterior abundances for *Yersinia pestis* in the RISE samples ranged from 0.0054–0.66%, corresponding to between $3,875$–$550,200$ assigned reads. Despite the large number of assigned reads, posterior abundances only exceeded 0.01% for five of the seven samples (RISE00, RISE139, RISE386, RISE505 and RISE511). This highlights the need to consider both relative and absolute abundance when making positive identifications.

Case study 2: Historic dental calculus

Dental plaque is a biofilm that develops naturally on teeth, and during life it periodically mineralizes to form dental cal-

DRAFT

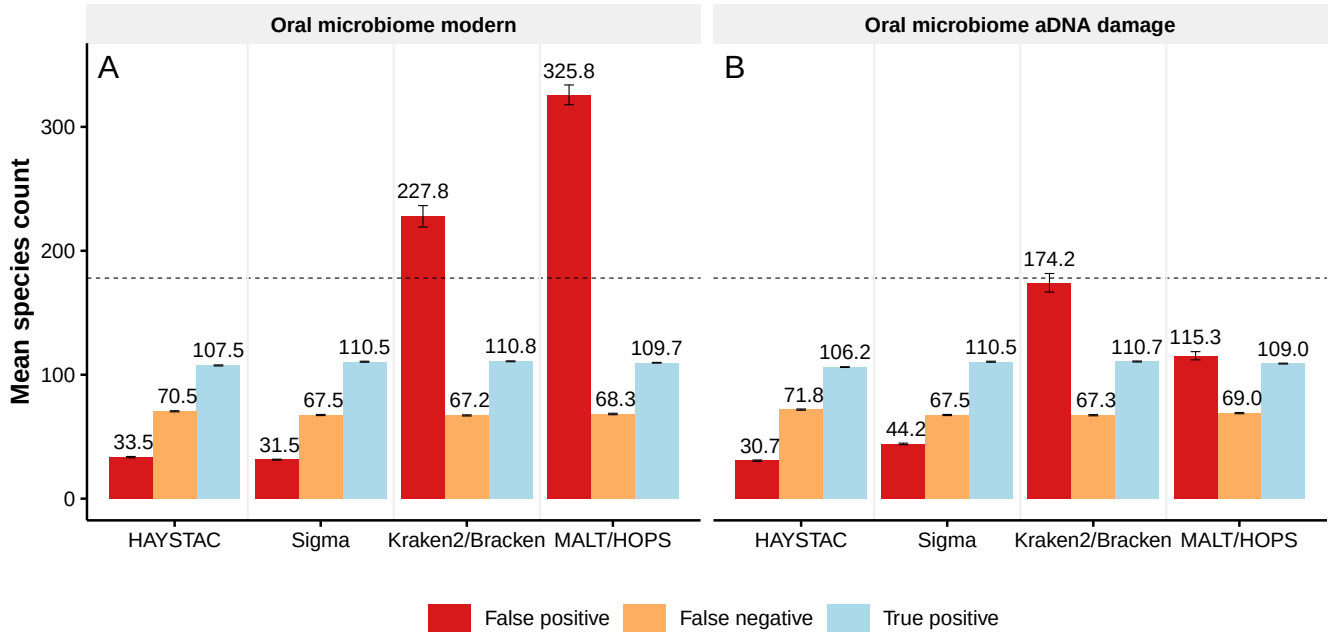

Figure 4. Accuracy of HAYSTAC and other methods for an oral microbiome simulation. Bar plot showing the mean count of false positives (red), false negatives (orange), and true detected species (blue) for two versions of the oral microbiome dataset, each with six replicates: (A) modern simulation, with fixed read lengths ($n = 6$); and (B) ancient simulation, with variable read lengths and post-mortem damage ($n = 6$). The dotted line shows the average number of simulated species in each set of samples (*i.e.*, the maximum true positive; $n = 178$), and numbers above the error bars indicate the mean species count in each category. For the modern simulation, HAYSTAC substantially outperforms both Kraken2/Bracken and MALT/HOPS with respect to false positives, and performs equivalently with Sigma. For the ancient simulation, HAYSTAC outperforms all three other methods with respect to false positives. The overall high rates of false negative identifications are due to the absence of many simulated species from the reference database for all four methods.

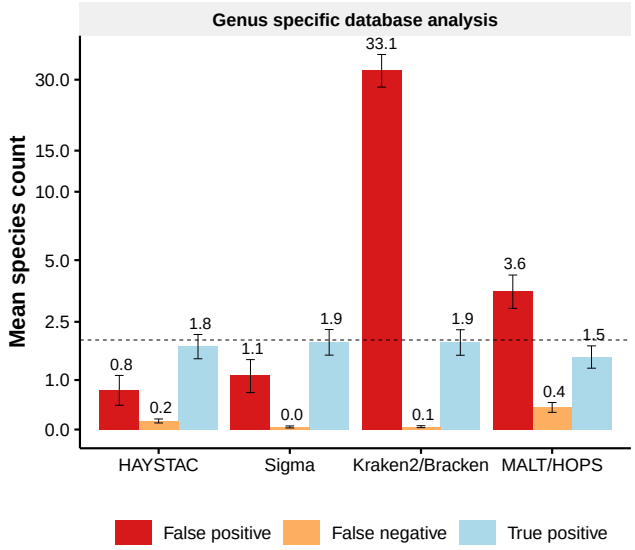

Figure 5. Accuracy of HAYSTAC and other methods using a reference database restricted to a single genus. Bar plot, with a pseudo-log10 transformed y-axis, showing the mean count of false positives (red), false negatives (orange), and true detected species (blue) for nine different genera (*Bacteroides*, *Burkholderia*, *Campylobacter*, *Clostridium*, *Corynebacterium*, *Desulfitobacterium*, *Mycobacterium*, *Solimonas* and *Streptococcus*), each with 20 samples from the general and oral microbiome datasets. The dotted line shows the average number of simulated species in each set of samples (*i.e.*, the maximum true positive; $n = 1.9$), and numbers above the error bars indicate the mean species count in each category. For the genus specific analysis, HAYSTAC substantially outperforms both Kraken2/Bracken and MALT/HOPS with respect to false positives, and performs equivalently with Sigma.

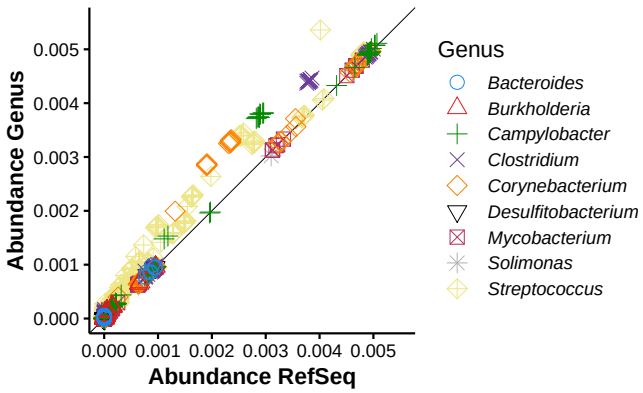

Figure 6. HAYSTAC inferred posterior abundance levels. Scatter plot showing the mean posterior abundances across all taxa ($n = 355$) and samples ($n = 20$) for either a genus specific database or the entire RefSeq representative database of prokaryotic species (plot truncated at $6 \cdot 10^{-3}$). Using a genus specific database has a small positive bias in mean posterior abundance for taxa within that genus (paired t-test p-value = $2.2 \cdot 10^{-16}$, mean of the differences = $1.4 \cdot 10^{-5}$), nevertheless the overall abundance levels are highly correlated ($R^2 = 0.985$). Computational runtime for the genus specific analyses are approximately 350 times faster and use 5 times less memory, making genus specific analyses suitable for rapid initial screening.

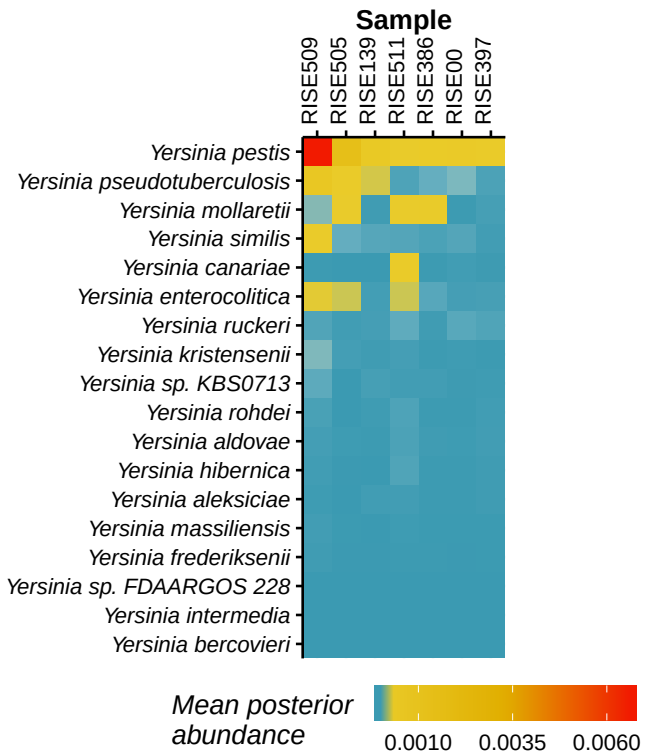

Figure 7. Posterior abundances of *Yersinia* species in Case Study 1. Heatmap showing the mean posterior abundances for the seven RISE samples, based on a genus specific analysis of 18 *Yersinia* species. *Yersinia pestis* is the species with the highest posterior abundance, followed by *Y. pseudotuberculosis*, in agreement with the results of (Rasmussen et al., 2015).

culus (Warinner et al., 2015), a robust substrate that can preserve biomolecules such as DNA from the biofilm and the host (Warinner et al., 2014). Although this substrate can act as a reservoir for respiratory pathobionts, analysis of ancient dental calculus metagenomic data is often challenging because the low abundance of these species makes it difficult to confidently identify them (Warinner et al., 2014). To test the performance of HAYSTAC on pathobiont aDNA sequences within host-associated ancient microbiomes, we analysed data from 48 dental calculus samples from a 19th century hospital in Oxford (Loe et al., 2014; Velsko et al., 2019).

We screened these libraries for potential upper respiratory pathobionts from the *Corynebacterium*, *Streptococcus*, *Klebsiella* and *Bordetella* genera (Figure 8), in an effort to test if such pathobionts could be preserved and identified in dental calculus. Members of the genera *Corynebacterium*, *Streptococcus* and *Bordetella* had also been included in the Oral Microbiome dataset. Within the *Corynebacterium* genus, we identified *Corynebacterium matruchotii* (abundance 0.01–1.63%; 36 samples) and *C. durum* (abundance 0.01–0.52%; 15 samples) both of which are oral commensal taxa. We did not, however, identify *Corynebacterium diphtheriae*, the species that causes diphtheria. We identified multiple species of *Streptococcus*, with the commensal *Streptococcus sanguinis* being identified in 33 samples at relatively high abundance (0.01–2.15%). The pathobiont *Streptococcus pneumoniae* was also positively identified at low abundance (0.01–0.07%) in seven samples. We did not find any evidence for species belonging to the genera *Klebsiella* and *Bordetella*.

Members of the *Haemophilus* genus were rarely identified, which is expected given that many species from that genus are generally low-abundance in dental plaque compared to other oral surfaces such as the tongue (Mark Welch et al., 2016), and are similarly low-abundance in dental calculus (Velsko et al., 2019). However, *Haemophilus parainfluenzae* was found in 18 samples (abundance between 0.01–0.43%), *H. haemolyticus* was identified in two samples (abundance 0.01–0.03%), and *H. parahaemolyticus* was identified only in sample CS18 (abundance 0.02%). An uncharacterized oral species, *Haemophilus oral taxon 036*, was identified in two samples (abundance 0.01–0.04%) and the commensal *H. sputorum* was also identified in two samples (abundance 0.01–0.03%).

As expected, our analysis indicates that while species naturally occurring in the oral cavity (e.g. *C. matruchotii* and *durum)*, were detected in most samples samples, upper respiratory system pathobionts (e.g. *S. pneumoniae*, *H. parainfluenzae*) were more rare. Although pathogens were found generally at very low abundance, in only a few individuals, our results indicate that HAYSTAC was able to positively identify some pathogenic species such as *Streptococcus pneumoniae* and *Haemophilus parainfluenzae*, suggesting that dental calculus is a promising substrate to study ancient respiratory pathobionts.

DRAFT

Discussion

Our results demonstrate that HAYSTAC provides a robust and computationally efficient framework to statistically assess the presence of specific species in a metagenomic sample. HAYSTAC identified fewer false positives than either Kraken2/Braken or MALT/HOPS in all simulations, and fewer than Sigma in simulations of ancient data. It runs faster than Sigma for large databases, uses less memory than Kraken2/Braken, and is both faster and uses less memory than MALT/HOPS. Furthermore, our results show that the use of reduced reference databases (e.g. a few selected genera as opposed to the entire RefSeq representative prokaryotic database), can dramatically reduce computational time (approximately 350 times faster) while still providing highly accurate species identification. Altogether, this makes HAYSTAC a flexible tool to rapidly and accurately screen metagenomic samples, a feature that is attractive for a large range of applications, including archaeogenetics and biosurveillance.

Species concept, horizontal gene transfer and species identification

Although we demonstrated that both HAYSTAC and Sigma provide more robust species assignments than MALT/HOPS or Kraken2/Bracken, their internal computation is based on

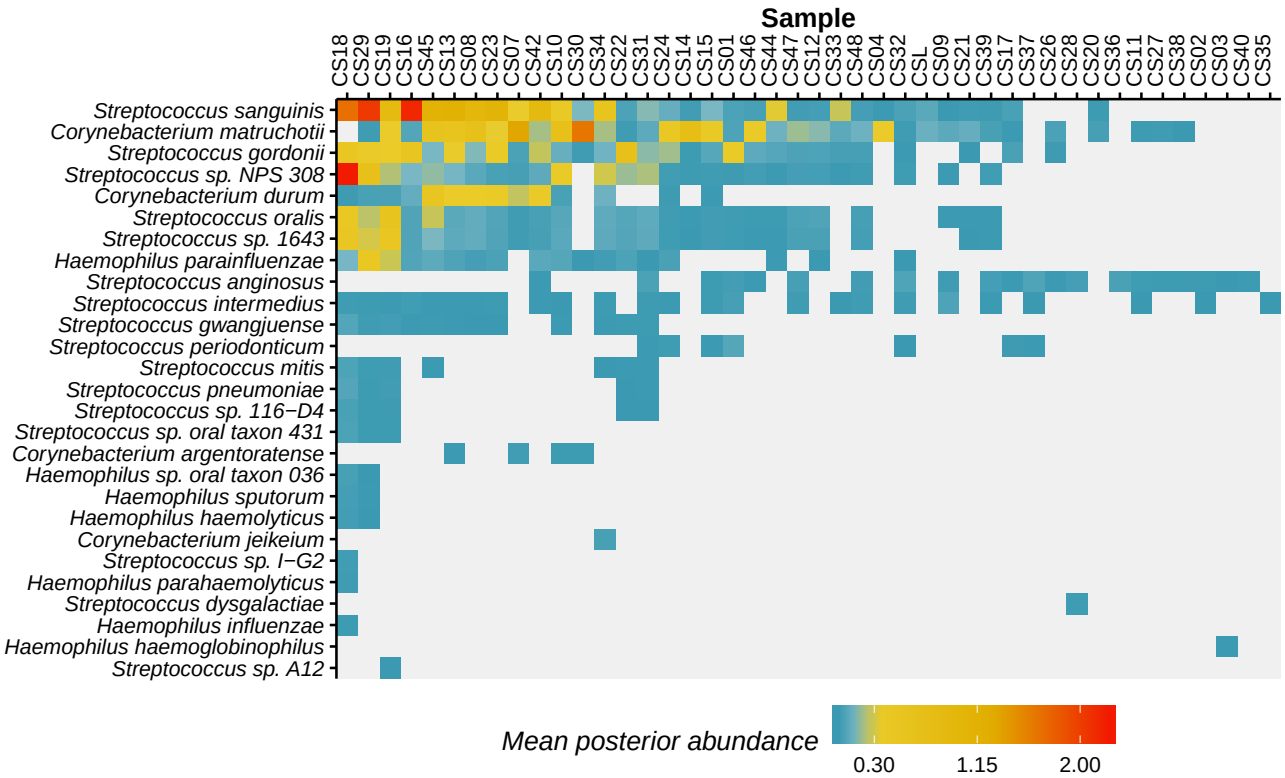

Figure 8. Posterior abundances of oral microbiome species in Case Study 2. Heatmap showing the mean posterior abundances for the 44 dental calculus samples, based on a custom database that combined the prokaryotic representative RefSeq and pathobionts with complete genomes from the following 5 genera: *Corynebacterium*, *Haemophilus*, *Klebsiella*, *Streptococcus*, and *Bordetella*. Species from these five genera of interest that naturally colonise the oral cavity can be found in more samples and at higher abundance (e.g. *C. matruchotii*) compared to pathobionts of the upper respiratory system (e.g. *S. pneumoniae*).

the assumption that intraspecific polymorphism is lower than interspecific polymorphism. While this is true in most cases, the concept of a species among microorganisms is often not straightforward. This is the case, for example, in the *Mycobacterium tuberculosis* complex in which interspecific variability is very low (Tortoli, 2012), and occasionally lower than the default value of expected mismatch (*i.e.*, expected level of intraspecific polymorphism) set in HAYSTAC ($\sigma = 0.05$; see eq. 4 below). HAYSTAC deals with this issue by introducing a "grey matter" category, to which reads that are equally likely to match two or more genomes are assigned. This ensures that they do not contribute to posterior abundance computations, which could possibly force false identifications, while also not dismissing them into "dark matter" with reads that cannot be classified. In cases where the levels of intra- and interspecific diversity are known (e.g., within a genus) it is possible to adjust the probability of the expected mismatches between a read and a genome (σ; see eq. 4), which can increase power while reducing false positive rates when analysing a specific dataset. This feature means HAYSTAC's performance can be fine-tuned to specific questions and datasets.

Identifying species within a metagenomic sample can also be complicated by horizontal gene transfer (HGT). HGT can lead to spurious, highly supported, species assignments—especially if the sequences originate from an outgroup species that is not represented in the database. HAYSTAC mitigates this issue by inspecting evenness of coverage across the genome for shallow sequencing data. Specifically, reads that belong to an outgroup species and that map to a genome in a reference database because of an HGT or contamination event are expected to map only to parts of the reference genome that resulted from such an event. HAYSTAC assesses whether reads are clustered more than expected by chance across a reference genome using a χ^2 contingency test (see Methods). This provides the user with a quantification of the evenness of spread, and thus a proxy signature of potential HGT or misassignment of reads from a taxon that is not represented in the reference database.

Issues with databases

Another issue facing metagenomics is contaminated sequences incorporated into reference genomes. For example, multiple bacterial RefSeq representative assemblies contain human sequences (Breitwieser et al., 2019). This can lead to spurious, highly supported, species assignments when profiling metagenome samples. Including genome sequences from potential contaminant sources (e.g., the human reference genome) in the database should alleviate these issues (Feuerborn et al., 2020). Alternatively, it is also possible to filter

reads before starting a taxonomic identification/assignment analysis (*i.e.*, by excluding those that map to the contaminant's reference genome). Potential contaminants, however, are not always known, and there are many gaps in reference sequence databases. In fact, inadequate species representation in databases is likely to be an important source of false positives for all species identification methods (Warinner et al., 2017).

Although reads belonging to a species that is not represented in a database are typically assigned to a higher taxonomic node ("grey matter" category in HAYSTAC), incorrect species assignment can occur if the missing taxon is closely related to another species in the database, or if the read derives from a conserved element found within a taxonomic group with sparse representation in the database. Although this is not considered a major problem for general profilers, as closely related species are likely to play a similar function in the ecosystem, such misassignments can greatly complicate the identification of specific pathogen species, especially if they have close environmental relatives.

Conclusions

Metagenomic data are now routinely produced and analysed, providing new insights into microbial communities that were previously unknown. Development of metagenomic analytical tools, however, has mostly focused on frameworks that can profile an entire microbial community. We showed that these methods, however, can produce high levels of false positives when focusing on specific species identification. Here we present a tool that can robustly assess whether a specific species is present in a metagenomic sample in one step, without the need to combine different pipelines to validate the results. HAYSTAC provides a user-friendly and automated method for rapidly constructing databases, and produces robust identifications regardless of the database size, allowing for both rapid hypothesis driven analyses and exhaustive profiling. Because of the mathematical framework it employs, HAYSTAC reliably produces the lowest number of false positive identifications, making it a valuable tool for both ancient and modern DNA microbial research.

Methods

Single Source Identification

We collect a series of observations (reads:$r_1 \cdot r_n$) and we denote the full set of observations as R. We map all the reads to a database of reference genomes and we define the event of sampling from the i-th reference genome in our database as G_j. Assuming that all reads comes from the same source (*i.e.*, a single species represented in the database), the events of sampling from a given genome are exhaustive and mutually exclusive, *i.e.*

$$G_1 \cdots G_n | G_i \cap G_j = \emptyset, \quad \forall i \neq j \quad \text{and} \quad \bigcup_i^n G_i = \Omega$$

From Bayes Theorem we have:

$$P(G_j|R) = \frac{P(R|G_j)P(G_j)}{P(R)} = \frac{P(R|G_j)P(G_j)}{\sum_k P(R|G_k)P(G_k)} \quad (1)$$

Given that each read is a different realization of sampling process (*i.e.*, an independent observation), we can express the likelihood as:

$$P(R|G_j) = \prod_i P(r_i|G_j) \quad (2)$$

Therefore, the posterior probability of sampling from a given genome G_j is:

$$P(G_j|R) = \frac{P(G_j)\prod_i P(r_i|G_j)}{\sum_k P(G_k)\prod_i P(r_i|G_k)} \quad (3)$$

Likelihood

We next derive a likelihood function to express the probability of a read given a genome $P(r_i|G_j)$. To do so, we use a previously published expression of this likelihood (Ahn et al., 2015) based on the number of mismatches between a read and a reference to which it is aligned:

$$P(r_i|G_j) = \sigma^z(1-\sigma)^{l-z}; \qquad z \leq U$$
$$P(r_i|G_j) = 0; \qquad z > U$$

where z is the number of mismatches between read r_i and genome G_j, U is the maximum number of mismatches allowed in the alignment and l is the length of read r_i (in bp).

The method assumes a uniform probability of mismatch within a given alignment (σ) to account for population level variability and sequencing errors. We expand this framework to leverage the fact that transitions and transversions occur at significantly different rates, particularly in ancient samples due to the effect of deaminations.

For each alignment we can then express the likelihood of generating a read r_i given that it has been sampled from the reference genome G_j based on both number of transitions (t) and transversions (v).

Let σ_t be the probability of observing a transition and σ_v the probability of observing a transversion, we can express the likelihood for a single read as:

$$P(r_i|G_j) = \sigma_v^{v_{i|j}} \sigma_t^{t_{i|j}} (1-\sigma_v-\sigma_t)^{l_i - v_{i|j} - t_{i|j}} \quad (4)$$

where l_i represent the length of the read r_i, while $v_{i|j}$ and $t_{i|j}$ denote the number of transversions and transitions between the read r_i and the genome G_j respectively.

Posterior

By substituting the likelihood term (eq. 4) into eq. 3, we can obtain an explicit form for the posterior probability of

DRAFT

sampling a set of reads (R) from a given reference genome:

$$P(G_j|R) = \frac{\sigma_t^{\sum_i t_{i|j}} \sigma_v^{\sum_i v_{i|j}} (1-\sigma_v-\sigma_t)^{\sum_i (l_i - t_{i|j} - v_{i|j})} P(G_j)}{\sum_{k=1}^{k=n} P(G_k) \sigma_t^{\sum_i t_{i|k}} \sigma_v^{\sum_i v_{i|k}} (1-\sigma_v-\sigma_t)^{\sum_i (l_i - t_{i|k} - v_{i|k})}} \tag{5}$$

By rescaling both numerator and denominator by a factor of $(1-\sigma_v-\sigma_t)^{\sum_i l_i}$ we get:

$$P(G_j|R) = \frac{P(G_j) \left(\frac{\sigma_t}{1-\sigma_v-\sigma_t}\right)^{\sum_i t_{i|j}} \left(\frac{\sigma_v}{1-\sigma_v-\sigma_t}\right)^{\sum_i v_{i|j}}}{\sum_{k=1}^{k=n} P(G_k) \left(\frac{\sigma_t}{1-\sigma_v-\sigma_t}\right)^{\sum_i t_{i|k}} \left(\frac{\sigma_v}{1-\sigma_v-\sigma_t}\right)^{\sum_i v_{i|k}}} \tag{6}$$

Let $V_j = \sum_i v_{i|j}$ denote the total number of observed transversions between all reads and genome G_j while $T_j = \sum_i t_{i|j}$ represents the total number of transitions. Let $\delta_v = \frac{\sigma_v}{1-\sigma_v-\sigma_t}$ and $\delta_t = \frac{\sigma_t}{1-\sigma_v-\sigma_t}$ denote instead the transversion to match ratio and the transition to match ratio, respectively. By introducing this more compact notation, the previous equation can be re-written as:

$$P(G_j|R) = \frac{P(G_j)\delta_t^{T_j}\delta_v^{V_j}}{\sum_k^n P(G_k)\delta_t^{T_k}\delta_v^{V_k}} \tag{7}$$

Numerical Representation

Our method assumes a fixed probability of mismatches ($\sigma_v + \sigma_t$) which can be specified by the user and has a default value of 5%. We estimate the transition/transversion ratio based on the total number of mismatches per categories observed in the entire dataset and we obtain the value for σ_v and σ_t by solving the following linear system:

$$\begin{cases} \sigma_t + \sigma_v = 0.05 \\ \sigma_t = \frac{T}{V}\sigma_v \end{cases} \tag{8}$$

where $V = \sum_j \sum_i v_{i|j} = \sum_j V_j$ and $T = \sum_j \sum_i t_{i|j} = \sum_j T_j$.

If we assume a uniform prior in which each genome has a probability of $1/n$ we can simplify eq. 7 even further obtaining:

$$P(G_j|R) = \frac{\delta_t^{T_j}\delta_v^{V_j}}{\sum_k^n \delta_t^{T_k}\delta_v^{V_k}} \tag{9}$$

When analysing a large number of reads, evaluating eq. 9 might become problematic due to numerical representation issues. To accommodate for this we use an equivalent expression of the posterior probability:

$$P(G_j|R) = \frac{1}{1 + \sum_{k\neq j}^n \delta_t^{T_k - T_j}\delta_v^{V_k - V_j}} \tag{10}$$

Each term in the summation at the denominator of eq. 10 represents the likelihood of genome G_k relative to genome G_J. There might be cases in which even the relative likelihood is so close to zero that cannot be represented numerically. We then take a conservative approach and assign to $\delta_t^{T_k - T_j}\delta_v^{V_k - V_j}$ the smallest number that the machine can represent (for a 64bit system it should be $2.2250738585072014 \cdot 10^{-308}$ which we roughly approximate with 10^{-300}).

Multiple Sources Identification

We then expand this method to identify multiple species from metagenomic data. The multiple source identification approach involves the following steps:

- Assigning each read to a reference genome
- Obtain the likelihood for the multinomial distribution of abundances
- Obtain the posterior distribution from the Dirichlet distribution
- Calculate the 95% credible interval around each posterior mean
- Assess significance

Operatively we apply eq.9 to each read separately and obtain a posterior probability for a read to originate from a given genome:

$$P(G_j|r_i) = \frac{\delta_t^{t_{i|j}}\delta_v^{v_{i|j}}}{\sum_k^n \delta_t^{t_{i|k}}\delta_v^{v_{i|k}}} \tag{11}$$

Thus, for each aligned read we obtain a vector of posterior probabilities (one value for each genome in our dataset). If a posterior probability value is above a given threshold (default: 0.75) we consider the read r_i to be informative and assign it to that reference genome. If instead all posterior values are below the threshold, we consider the read to be uninformative and we assign it to the $n+1$ category where n is the number of reference genomes. This means that any read r_i that has been successfully aligned, can only be assigned to a single category: either to one reference genome or to the "Grey Matter" (GM, the $n+1$ category). All reads which belong to species that do not possess a representative reference genome (or a close relative) in the database and therefore haven't been successfully aligned to any of the reference genomes, will then be assigned to the $n+2$ category called "Dark Matter" (DM).

Repeating this procedure $\forall r_i \in R$ allows us to populate a matrix $M_{N\times(n+2)}$ in which each row represents a read and each column represents a reference genome in our database or either the GM or DM categories. Hence, the assignment matrix represents a series of N observations from a multinomial distribution of $n+2$ variables (where N is the total number of reads). This distribution is parametrized by a probability mass function $q = [q_1, q_2, ..., q_n, q_{GM}, q_{DM}]$.

DRAFT

The likelihood of N observations has thus the following form:

$$M \sim Multinomial_{n+2}(N,q)$$
$$f(x_1,x_2,\cdots,x_{n+2}|N,q) = \frac{N!}{x_1!x_2!\cdots x_{n+2}!}\prod_{j=1}^{n+2} q_j^{x_j} \quad (12)$$

where x_j is the sum of the column j of M, *i.e.* the number of occurrences of genome G_j.

Dirichlet prior

The Dirichlet distribution (Dir) is a conjugate prior for the multinomial distribution and it has desirable mathematical properties which make it appropriate to model the composition of metagenomes.

Let π_0 indicate the prior distribution and π^* the posterior.

From Bayes theorem,

$$\pi^*(M|\Theta) \propto L(M|\Theta)\pi_0(\Theta) \quad (13)$$

We employ a uniform prior:

$$\begin{aligned} &\pi_0(\Theta) \sim Dir(\alpha) \\ &\alpha_0 = (\alpha_1,\cdots,\alpha_{n+2}), \\ &\alpha_i = 1 \quad \forall i = 1,\cdots,n+2 \end{aligned} \quad (14)$$

Because the Dirichlet is a conjugate prior for the multinomial likelihood (eq. 12), the posterior density has the same form, *i.e*

$$\begin{aligned} &\pi_*(M|\Theta) \sim Dir(\alpha^*) \\ &\alpha^* = (\alpha_1^*,\cdots,\alpha_{n+2}^*), \\ &\alpha_j^* = \alpha_j + x_j,\ j = 1,\cdots,n+2 \end{aligned} \quad (15)$$

where, again, x_j is the number of observations (reads) of genome G_j in the assignment matrix M. We can then use this framework to compute the minimum abundance of a species (represented by its closest reference genome) in a metagenome.

Posterior mean and Confidence interval

Let γ_j denote the posterior abundance of genome G_j. From the property of the Dirichlet we can obtain the posterior mean as follows:

$$\begin{aligned} &E[\gamma_j|X] = \frac{\alpha_j^*}{\alpha^{**}} \\ &\alpha^{**} = \sum_{k=1}^{n+2}\alpha_k^* = (n+2) + \sum_{k=1}^{n+2} x_k = n+2+N \end{aligned} \quad (16)$$

meaning that the denominator of each posterior mean corresponds to the total number of reads plus the number of categories (the number of reference genome + 2).

The marginal distribution of each abundance is:

$$\gamma_j \sim Beta(\alpha_j^*, \alpha^{**} - \alpha_j^*) \quad (17)$$

and we can use this to obtain the 95% CI numerically.

Computational architecture

HAYSTAC is written in Python and built upon the snakemake workflow engine (Köster and Rahmann, 2012). It operates as a wrapper around custom python scripts and several common bioinformatics tools, including: AdapterRemoval (Schubert et al., 2016), bowtie2 (Langmead and Salzberg, 2012), DeDup (Peltzer et al., 2016), mapDamage (Jónsson et al., 2013), samtools (Li et al., 2009), SeqKit (Shen et al., 2016), seqtk (Li, 2012) and sra-tools (Leinonen et al., 2011). It uses the conda package manager, with bioconda (Grüning et al., 2018), to manage its dependencies.

A tool for constructing databases

HAYSTAC includes a user-friendly tool to download genomes from NCBI and build customisable reference databases. After deciding *a priori* which taxa to include in their database (e.g., anywhere between a single genus to many thousands of taxa), the user constructs a database from any combination of: (i) an NCBI search query; (ii) a user specified list of NCBI accession codes; (iii) the RefSeq representative database of prokaryotic species; and/or (iv) a list of user provided reference assemblies (in FASTA format).

To construct a valid NCBI search query, visit the NCBI Nucleotide database website (`https://www.ncbi.nlm.nih.gov/nucleotide/`) and use the search feature to obtain a correctly formatted query string from the "Search details" box. This search query can then be used directly with HAYSTAC to automatically download and build a reference database based on the accession codes present in the resultset returned by the query.

We caution that the choice of reference database can dramatically affect species identification for all methods that use reference databases. For example, issues can arise if the reference genomes in the database are not complete, as reads may map uniquely to one reference genome because other incomplete genomes in the database do not possess that specific sequence. To mitigate this issue, HAYSTAC uses a simple heuristic to select the longest genome available per species in the input resultset, which limits issues arising from unequal size of reference genomes in the database.

Processing and aligning reads

Pre-processed reads can be directly passed onto the alignment step (Figure 1). This is useful when dealing with modern DNA. The user can also decide whether to remove sequencing adapters in the case of raw reads, and for ancient/degraded DNA collapse overlapping mate pairs using a wrapper for AdapterRemoval2 (Schubert et al., 2016). Reads are then aligned to a database using Bowtie2 (Langmead and Salzberg, 2012). We first align to a single Bowtie2 index that contains all the reference genomes, to obtain a BAM file that contains one single best alignment per read. For this alignment step we use the local alignment mode in Bowtie2. This allows us to discard reads that did not map to a single reference genome. This BAM is used by HAYSTAC to compute

DRAFT

the expected T_s/T_v ratio from the data, which is used as prior in eq 4. Reads that pass this first filtering step are then re-aligned to each reference genome separately using Bowtie2 (one index per reference) in end-to-end mode. We set a maximum number of mismatches to calculate the minimum alignment score during the Bowtie alignment. The maximum mismatch allowed during this second alignment step is calculated by the average read length of the input library. Allowing more mismatches results in slower run time as the number of reads included in the analytical step increases.

Computing abundance

Alignment files are then processed to calculate a likelihood score for each read/reference alignment and build a likelihood matrix that possesses $n \times r$ entries, where n is the number of reference genomes and r is the number of reads that pass the first filtering stage. This is done by computing T_s, and T_v for each alignment, and computing the likelihood using eq 4. Here the user can change the probability of mismatch within a given alignment (σ; default 0.05) (see eq 4). For each read we then compute a posterior probability using eq 7. Reads are then filtered based on their posterior probability (default 0.75) and included in the Dirichlet assignment step. Reads that have not been aligned to any taxa in the database get assigned to the "Dark Matter" category, a category that absorbs reads whose reference genomes are not included in our database. Reads that have been aligned to multiple taxa in our database but obtain a posterior probability below the threshold that is set for the Dirichlet assignment are also assigned to another special category called the "Grey Matter", a category that includes reads that were not informative enough to be uniquely aligned to a specific taxon. Both of these categories are included in the mean posterior abundance calculations.

Assessing evenness of coverage

We use a χ^2 test to assess if the reads assigned to a specific taxon represent a random sample of its genome or whether they are clustering around specific regions. In the case of a true positive identification the reads should not be clustering around specific genome regions (p-value of 0), something that would indicate an HGT event. This test is particularly useful for shallow sequencing data that do not allow for more than $1\times$ genome coverage of a given reference genome.

Representative species RefSeq database

For our analyses we build a database out of all the representative species of the prokaryotic RefSeq database. A list of all the species that were included can be found here: `https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prok_representative_genomes.txt`. In cases where the RefSeq representative database contains more than one strain for a given species, the first listed accession was picked. We also added complete genomes from the genera *Klebsiella*, *Streptococcus*, *Corynebacterium*, *Bordetella*, *Haemophilus* and *Yersinia*, for these species that were not present in the prokaryotic representative RefSeq. When possible (*i.e.*, when genome assembly with plasmid data were available) we included plasmid sequences in the database.

Simulations

We generated 3 different simulated datasets. The first dataset was generated to determine the accuracy of different methods under simple conditions. We did not apply any chemical damage pattern and the read length was kept constant at 60 bp. This was done by randomly sampling 10 species from the prokaryotic representative RefSeq database (see above) twice, to get two different sets of 10 species. For each set we then simulated 2 different sequencing library samples of 1 million single-end reads using gargammel (Renaud et al., 2017). One set had no human genome included and all taxa were present at 10% abundance. In the second set 25% of reads were from the human genome, and all remaining taxa were present at 7.5% abundance. This dataset included 4 simulated sample sequencing files.

The second dataset was produced to compare performance of the different tools when analysing ancient and modern DNA, and did not include the human genome. To do so, we first generated two sets of 100 species randomly sampled from the prokaryotic representative RefSeq database, and two sets of 500 randomly sampled species. The species were different between each of the 2 random samplings for both the 100 species and 500 species sets. For each species profile in both the 100 species and 500 species sets, we simulated two different libraries (8 total libraries) using gargammel with 10 million single-end reads of length 125bp from an Illumina HiSeq 2500 run. The species abundances were kept consistent within each library, where libraries of 100 species have all taxa at 1% abundance and libraries of 500 species have all taxa at 0.2% abundance. The first library incorporated aDNA damage patterns based on the profile of reads that mapped to the Tannerella forsythia genome from a real ancient dental calculus sample (CS21 from Velsko et al. (2019)) and the second library did not include aDNA damage patterns.

Lastly, to assess our method with more realistic data sets, we analysed simulated data previously published by Velsko et al. (2018) designed to resemble a dental plaque community, which included species that were not in the prokaryotic representative RefSeq database. Twelve datasets of 200 species each were generated: with and without ancient DNA damage, each with even species abundance (0.5%) or with abundance based on values observed in an ancient oral microbiome bacterial community (see Supplemental Table S8). The type of simulation included members of the *Corynebacterium* genus without *Corynebacterium diphtheriae*; the second type included all members of the *Corynebacterium* genus (including *Corynebacterium diphtheriae*); the third type did not include any member of the *Corynebacterium* genus. Each of these twelve samples contained 5 million paired end reads of varying length, that were collapsed with AdapterRemoval.

DRAFT

Analyses of simulated data

We trimmed adapters using AdapterRemoval v2 (Schubert et al., 2016) excluding reads that were smaller than 15 bp. Analyses of simulated data with Kraken2/Bracken (Wood et al., 2019; Lu et al., 2017), MALT (Vågene et al., 2018) were conducted using default parameters. The MALT output was subsequently analysed with HOPS (Hübler et al., 2019) with default values but providing a list of candidate species that included all the species in the custom database that we used for the rest of the methods. HAYSTAC and Sigma (Ahn et al., 2015) were run with default parameters. For HAYSTAC, Kraken2/Bracken and Sigma positive identification abundance threshold was set to 0.01% while we used 50 reads for MALT/HOPS as in Hübler et al. (2019).

In order to assess how HAYSTAC, Sigma, Kraken2/Bracken and MALT/HOPS perform with genus specific database, we analysed 20 simulated samples (based on RefSeq genomes; see above) and aligning reads to all representative genomes in a single genus. Ten genera were picked randomly for these analyses: *Bacteroides* (24 species), *Burkholderia* (3 species), *Campylobacter* (22 species), *Clostridium* (91 species), *Corynebacterium* (83 species), *Desulfitobacterium* (4 species), *Lactobacillus* (139 species), *Mycobacterium* (36 species), *Solimonas* (3 species) and *Streptococcus* (78 species).

Case study 1: Ancient *Yersinia pestis*

We tested HAYSTAC using a published dataset of ancient human bone and tooth samples in which *Yersinia pestis* was identified (Rasmussen et al., 2015), to compare performance against standard techniques used to identify specific pathogens in aDNA samples. We downloaded reads from all seven samples in which the authors detected *Y. pestis* (RISE00, RISE139, RISE386, RISE505, RISE509, RISE511) from the European Nucleotide Archive (NCBI BioProject accession: PRJEB10885). We then used HAYSTAC to build a database that contains the longest complete genome for each species of the genus *Yersinia* and aligned reads using default parameters. HAYSTAC was then used to compute posterior abundance of each species.

Case study 2: Historic dental calculus

We also tested HAYSTAC using an historic dental calculus dataset (Velsko et al., 2019) (PRJEB30331). Raw reads were processed as described in Velsko et al. (2019). We attempted to identify a set of respiratory pathogens in these calculus samples, including: *Corynebacterium diphtheriae* and *Bordetella pertussis*, the causative agents of diphtheria and whooping cough, respectively (Figure 8). We also tested for the presence of species belonging to the *Haemophilus* (ideally *Heamphilus influenzae* or *parainfluezae*), *Klebsiella* (ideally *Klebsiella pneumonae*) and *Streptococcus* (*Streptococcus pneumoniae*) genera. These species were selected to test if pathobionts of the upper respiratory system could be found in dental calculus.

Performance test

For the performance tests we created a reproducible conda environment, inside which all the different pipelines were installed and run. For each method we used the GNU time command to measure elapsed time (wall clock) and peak memory usage (maximum resident set size). Scripts to run the performance benchmarks are available from https://github.com/antonisdim/haystac_paper

We used a total of 500 accessions, one for each species, which were a subset of the accessions used for the analysis of all real and simulated datasets. We built databases with an increasing number of species (10, 100, 500) and the same database was given as input to all the methods. We only measured the database building time and maximum memory usage for each method, as HAYSTAC was used to efficiently fetch all 500 reference genomes from NCBI.

For sample inputs we used samples from the Human Microbiome Project, with accession SRS078671. We merged the forward mate and the singleton reads in one file that we then treated as single end reads. That file was subsequently subsampled to generate fastq files of 10 K, 100 K, and 1 M SE reads respectively, that were used as inputs for all the tests.

For the sample analysis we performed two tests, one that measured max memory usage and runtime for samples of increasing size against the database of 500 species, and a second where the input file size was kept constant at 1 million reads, while the database size was being varied, again measuring the same variables for memory and elapsed execution time.

ACKNOWLEDGEMENTS

E.A.D. was supported by the DTP in Environmental Research and funded by NERC award NE/L002612/1. L.A.F.F. and G.L. were supported either by a European Research Council grants (ERC-2013-StG-337574-UNDEAD and ERC-2019-StG-853272-PALAEOFARM) and Natural Environmental Research Council grants (NE/K005243/1, NE/K003259/1, NE/S007067/1, and NE/S00078X/1). E.A.D, L.A.F.F. and A.C. were also supported by the Wellcome Trust (210119/Z/18/Z). The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (http://dx.doi.org/10.5281/zenodo.22558) and Queen Mary's Apocrita HPC facility.

References

Ahn TH, Chai J, Pan C. Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics 2015 Jan;31(2):170–177.

Arora N, Schuenemann VJ, Jäger G, Peltzer A, Seitz A, Herbig A, et al. Origin of modern syphilis and emergence of a pandemic Treponema pallidum cluster. Nat Microbiol 2016 Dec;2:16245.

Borówka P, Pułaski Ł, Marciniak B, Borowska-Strugińska B, Dziadek J, Żądzińska E, et al. Screening methods for detection of ancient Mycobacterium tuberculosis complex fingerprints in next-generation sequencing data derived from skeletal samples. Gigascience 2019 Jun;8(6).

Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, Coombes BK, et al. A draft genome of Yersinia pestis from

DRAFT

victims of the Black Death. Nature 2011 Oct;478(7370):506–510.

Breitwieser FP, Pertea M, Zimin AV, Salzberg SL. Human contamination in bacterial genomes has created thousands of spurious proteins. Genome Res 2019 Jun;29(6):954–960.

Chen PE, Cook C, Stewart AC, Nagarajan N, Sommer DD, Pop M, et al. Genomic characterization of the Yersinia genus. Genome Biol 2010 Jan;11(1):R1.

Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 2018 Dec;6(1):226.

De Keukeleire S, De Bel A, Jansen Y, Janssens M, Wauters G, Piérard D. Yersinia ruckeri, an unusual microorganism isolated from a human wound infection. New Microbes New Infect 2014 Jul;2(4):134–135.

Eerkens JW, Nichols RV, Murray GGR, Perez K, Murga E, Kaijankoski P, et al. A probable prehistoric case of meningococcal disease from San Francisco Bay: Next generation sequencing of Neisseria meningitidis from dental calculus and osteological evidence. Int J Paleopathol 2018 Sep;22:173–180.

Feuerborn TR, Palkopoulou E, van der Valk T, von Seth J, Munters AR, Pečnerová P, et al. Competitive mapping allows for the identification and exclusion of human DNA contamination in ancient faunal genomic datasets. BMC Genomics 2020 Nov;21(1):844.

Gonzalez A, Vázquez-Baeza Y, Pettengill JB, Ottesen A, McDonald D, Knight R. Avoiding Pandemic Fears in the Subway and Conquering the Platypus. mSystems 2016 May;1(3).

Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods 2018 Jul;15(7):475–476.

Hübler R, Key FM, Warinner C, Bos KI, Krause J, Herbig A. HOPS: automated detection and authentication of pathogen DNA in archaeological remains. Genome Biol 2019 Dec;20(1):280.

Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 2013 Jul;29(13):1682–1684.

Key FM, Posth C, Krause J, Herbig A, Bos KI. Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication. Trends Genet 2017 Aug;33(8):508–520.

Kistler L, Ware R, Smith O, Collins M, Allaby RG. A new model for ancient DNA decay based on paleogenomic meta-analysis. Nucleic Acids Res 2017 Jun;45(11):6310–6320.

Köster J, Rahmann S. Snakemake–a scalable bioinformatics workflow engine. Bioinformatics 2012 Oct;28(19):2520–2522.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012 Mar;9(4):357–359.

Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res 2011 Jan;39(Database issue):D19–21.

Li H. seqtk Toolkit for processing sequences in FASTA/Q formats. GitHub 2012;767:69.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009 Aug;25(16):2078–2079.

Loe L, Pollard M, Gibson M, Simmonds A, Cotter J, Parkin A. Radcliffe Infirmary Burial Ground, Oxford. Post-excavation assessment and project design. Oxford; 2014.

Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci 2017 Jan;3:e104.

Mann AE, Sabin S, Ziesemer K, Vågene ÅJ, Schroeder H, Ozga AT, et al. Differential preservation of endogenous human and microbial DNA in dental calculus and dentin. Sci Rep 2018 Jun;8(1):9822.

Mark Welch JL, Rossetti BJ, Rieken CW, Dewhirst FE, Borisy GG. Biogeography of a human oral microbiome at the micron scale. Proc Natl Acad Sci U S A 2016 Feb;113(6):E791–800.

Mühlemann B, Vinner L, Margaryan A, Wilhelmson H, de la Fuente Castro C, Allentoft ME, et al. Diverse variola virus (smallpox) strains were widespread in northern Europe in the Viking Age. Science 2020 Jul;369(6502).

Müller R, Roberts CA, Brown TA. Biomolecular identification of ancient Mycobacterium tuberculosis complex DNA in human remains from Britain and continental Europe. Am J Phys Anthropol 2014 Feb;153(2):178–189.

Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res 2014 Jul;24(7):1180–1192.

Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, et al. EAGER: efficient ancient genome reconstruction. Genome Biol 2016 Mar;17:60.

Pósa A, Maixner F, Mende BG, Köhler K, Osztás A, Sola C, et al. Tuberculosis in Late Neolithic-Early Copper Age human skeletal remains from Hungary. Tuberculosis 2015 Jun;95 Suppl 1:S18–22.

Prendergast ME, Buckley M, Crowther A, Frantz L, Eager H, Lebrasseur O, et al. Reconstructing Asian faunal introductions to eastern Africa from multi-proxy biomolecular and archaeological datasets. PLoS One 2017 Aug;12(8):e0182565.

Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjögren KG, et al. Early divergent strains of Yersinia pestis in Eurasia 5,000 years ago. Cell 2015 Oct;163(3):571–582.

Renaud G, Hanghøj K, Willerslev E, Orlando L. gargammel: a sequence simulator for ancient DNA. Bioinformatics 2017 Feb;33(4):577–579.

Reuter S, Connor TR, Barquist L, Walker D, Feltwell T, Harris SR, et al. Parallel independent evolution of pathogenicity within the genus Yersinia. Proc Natl Acad Sci U S A 2014 May;111(18):6768–6773.

DRAFT

Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes 2016 Feb;9:88.

Schuenemann VJ, Bos K, DeWitte S, Schmedes S, Jamieson J, Mittnik A, et al. Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of Yersinia pestis from victims of the Black Death. Proc Natl Acad Sci U S A 2011 Sep;108(38):E746–52.

Schuenemann VJ, Kumar Lankapalli A, Barquera R, Nelson EA, Iraíz Hernández D, Acuña Alonzo V, et al. Historic Treponema pallidum genomes from Colonial Mexico retrieved from archaeological remains. PLoS Negl Trop Dis 2018 Jun;12(6):e0006447.

Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One 2016 Oct;11(10):e0163962.

Spyrou MA, Bos KI, Herbig A, Krause J. Ancient pathogen genomics as an emerging tool for infectious disease research. Nat Rev Genet 2019 Jun;20(6):323–340.

Spyrou MA, Tukhbatova RI, Feldman M, Drath J, Kacki S, Beltrán de Heredia J, et al. Historical Y. pestis Genomes Reveal the European Black Death as the Source of Ancient and Modern Plague Pandemics. Cell Host Microbe 2016 Jun;19(6):874–881.

Sumi Y, Miura H, Michiwaki Y, Nagaosa S, Nagaya M. Colonization of dental plaque by respiratory pathogens in dependent elderly. Arch Gerontol Geriatr 2007 Mar;44(2):119–124.

Tett A, Huang KD, Asnicar F, Fehlner-Peach H, Pasolli E, Karcher N, et al. The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. Cell Host Microbe 2019 Nov;26(5):666–679.e7.

Tortoli E. Phylogeny of the genus Mycobacterium: many doubts, few certainties. Infect Genet Evol 2012 Jun;12(4):827–831.

Vågene ÅJ, Herbig A, Campana MG, Robles García NM, Warinner C, Sabin S, et al. Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. Nat Ecol Evol 2018 Mar;2(3):520–528.

Velsko IM, Fellows Yates JA, Aron F, Hagan RW, Frantz LAF, Loe L, et al. Microbial differences between dental plaque and historic dental calculus are related to oral biofilm maturation stage. Microbiome 2019 Jul;7(1):102.

Velsko IM, Frantz LAF, Herbig A, Larson G, Warinner C. Selection of Appropriate Metagenome Taxonomic Classifiers for Ancient Microbiome Research. mSystems 2018 Jul;3(4).

Warinner C, Herbig A, Mann A, Fellows Yates JA, Weiß CL, Burbano HA, et al. A Robust Framework for Microbial Archaeology. Annu Rev Genomics Hum Genet 2017 Aug;18:321–356.

Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, Grossmann J, et al. Pathogens and host immunity in the ancient human oral cavity. Nat Genet 2014 Apr;46(4):336–344.

Warinner C, Speller C, Collins MJ, Lewis CM Jr. Ancient human microbiomes. J Hum Evol 2015 Feb;79:125–136.

Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J, et al. Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. N Engl J Med 2019 Jun;380(24):2327–2340.

Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol 2019 Nov;20(1):257.

Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 2014 Mar;15(3):R46.

Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell 2019 Aug;178(4):779–794.

Zhou Z, Luhmann N, Alikhan NF, Quince C, Achtman M. Accurate Reconstruction of Microbial Strains from Metagenomic Sequencing Using Representative Reference Genomes. In: Research in Computational Molecular Biology Springer International Publishing; 2018. p. 225–240.

DRAFT