HOPS: Automated detection and authentication of pathogen DNA in archaeological remains

Ron Hübler[1*] (huebler@shh.mpg.de), Felix M. Key[1,2,3*] (key@shh.mpg.de), Christina Warinner[1] (warinner@shh.mpg.de), Kirsten I. Bos[1] (bos@shh.mpg.de), Johannes Krause[1] (krause@shh.mpg.de), Alexander Herbig[1] (herbig@shh.mpg.de)

Affiliation

1. Max-Planck Institute for the Science of Human History
2. Institute for Medical Engineering and Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
3. Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

* Equal contribution

Correspondence: Felix M. Key (key@shh.mpg.de), Alexander Herbig (herbig@shh.mpg.de)

Abstract

High-throughput DNA sequencing enables large-scale metagenomic analyses of complex biological systems. Such analyses are not restricted to present day environmental or clinical samples, but can also be fruitfully applied to molecular data from archaeological remains (ancient DNA), and a focus on ancient bacteria can provide valuable information on the long-term evolutionary relationship between hosts and their pathogens. Here we present HOPS (**H**euristic **O**perations for **P**athogen **S**creening), an automated bacterial screening pipeline for ancient DNA sequence data that provides straightforward and reproducible information on species identification and authenticity. HOPS provides a versatile and fast pipeline for high-throughput screening of bacterial DNA from archaeological material to identify candidates for subsequent genomic-level analyses.

Keywords

ancient DNA, archaeogenetics, pathogen detection, metagenomics

Background

High-throughput DNA sequencing enables large-scale metagenomic analyses of environments and host tissues, providing an unprecedented understanding of life's microbial diversity. Examples of coordinated efforts to quantify this diversity include the Human Microbiome Project (1), the Tara Ocean Project (2) and the Earth Microbiome Project (3). Metagenomic data from human archaeological remains (*e.g.* bones, teeth or dental calculus), which provide a window into the individuals' metagenomic past, is a welcome addition to the wide landscape of microbial diversity now being revealed. While many ancient DNA (aDNA) studies focus on the analysis of human endogenous DNA isolated from ancient specimens (4-8), the co-recovered metagenomic aDNA can be queried to provide information related to endogenous microbial content at death, with applications ranging from characterizing the natural constituents of the microbiota to identifying systemic infectious diseases (9, 10).

Genomic-level investigations of ancient pathogens have provided valuable information about the evolution of *Yersinia pestis* (11-18), *Mycobacterium leprae* (19, 20), *Mycobacterium tuberculosis* (21, 22), pathogenic *Brucella* species (23, 24), *Salmonella enterica* (25, 26) and *Helicobacter pylori* (27), with others surely on the horizon. Notably, most studies to date have leveraged paleopathological evidence or historical context to pinpoint *a priori* involvement of a specific bacterial pathogen. However, the vast majority of infectious diseases do not lead to the formation of distinct and characteristic bone lesions, and most remains are found in contexts that lack clear associations with a particular disease. Consequently, studies of ancient pathogens must consider a long list of candidate microbes. Therefore, an automated computational screening tool that both detects and evaluates pathogen genetic signals in ancient metagenomic data is needed. Importantly, this tool should also be able to distinguish potential pathogens from the large and diverse microbial background typical of archaeological and other decomposed material, a consideration not typically required for tools developed for clinical applications.

64

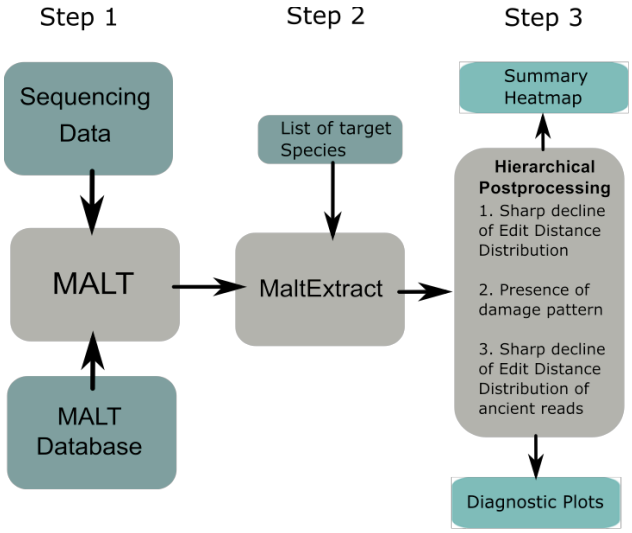
65 To save computational time and effort, most available metagenomic profiling tools
66 focus only on individual genes, such as the 16S rRNA gene used by QIIME (28), or
67 panels of marker genes, such as those used by MetaPhlAn2 (29) and MIDAS (30),
68 that are information-rich and highly species-specific. However, these genes make up
69 only a small proportion of a bacterial genome (the 16S rRNA gene, for example,
70 accounts for only ~0.2% of a bacterial genome), and if a pathogen is present at low
71 abundance compared to host and environmental DNA, these genes are likely to be
72 missed in routine metagenomic sequencing screens. As such, although these tools can
73 be specific, they lack the sensitivity required for ancient pathogen screening from
74 shallow metagenomic datasets. Screening techniques that accommodate queries of
75 whole genomes are of clear benefit for archaeological studies (25). However, while
76 some algorithms, such as Kraken (31), have been developed to query databases that
77 contain thousands of complete reference genomes using k-mer matching, this
78 approach does not produce the alignment information necessary to further evaluate
79 species identification accuracy or authenticity.
80

81 In addition to taxonomic classification (32), it is also critical to distinguish ancient
82 bacteria from modern contaminants (9, 10). Genuine aDNA, especially pathogen
83 bacterial DNA, is usually only present in small amounts and can be distinguished
84 from modern DNA contamination by applying an established set of authenticity
85 criteria (9, 10), the most important of which is the assessment of DNA damage. In
86 ancient DNA, cytosine deamination accumulates over time at DNA fragment termini
87 (33, 34), thus leading to a specific pattern of nucleotide misincorporation. The
88 evaluation of additional authenticity criteria such as edit distances and the distribution
89 of mapped reads across the reference are also recommended to mitigate against
90 database bias artifacts and to further validate taxonomic assignments (9, 10). While
91 manual evaluation of species identification and aDNA authenticity using standalone
92 tools might be feasible for a small sample set, it is impractical and too labour
93 intensive to apply to the large sample sizes typical of recent ancient DNA
94 investigations. The increased throughput of the ancient DNA field warrants an
95 automated high-throughput solution for pathogen detection in metagenomic datasets.
96 Successful ancient pathogen detection is reliant upon three criteria: (i) specificity of
97 species-level detection against a diverse metagenomic background, (ii) high

sensitivity that allows detection even with a weak signal when only trace amounts of species-specific DNA are present, and (iii) authentication of its ancient origin. However, no software currently exists that fulfills all requirements essential for reliable screening of metagenomic aDNA. Here we introduce HOPS (Heuristic Operations for Pathogen Screening), an automated computational pipeline that screens metagenomic aDNA data for the presence of bacterial pathogens and assesses their authenticity using established criteria. We test HOPS on experimental and simulated data and compare it to common metagenomic profiling tools designed for modern DNA analysis. We show that HOPS outperforms available tools, is highly specific and sensitive, and can perform reliable and reproducible taxonomic identification and authentication with as few as 50 reads.

109 # RESULTS

110 ## HOPS Workflow



111

112 **Figure 1. Schematic depiction of HOPS workflow.** First, MALT aligns the metagenomic
113 data against its reference database and has an optional mode for processing aDNA reads. MaltExtract
114 then processes the MALT output with various filters and produces various statistics. Finally, post
115 processing procedures provide a comprehensive visualization of the output which can be evaluated to
116 identify putatively positive hits.

117

118 HOPS consists of three parts (Figure 1): i) a modified version of MALT (25, 35),
119 which includes optional PCR duplicate removal and optional deamination pattern
120 tolerance at the ends of reads; ii) The newly developed program MaltExtract, which
121 provides statistics for the evaluation of species identification as well as aDNA
122 authenticity criteria for a user-specified set of bacterial pathogens, with additional
123 functionality to filter the aligned reads by various measures such as read length,
124 sequence complexity or percent identity; and iii) a post-processing script that provides
125 a summary overview for all samples and potential bacterial pathogens that have been
126 identified.

127 ## MALT

128 MALT (Megan Alignment Tool) (25, 35) is a fast alignment and taxonomic binning
129 tool for metagenomic data that aligns DNA sequencing reads to a user-specified
130 database of reference sequences. Reads are assigned to taxonomic nodes by the naïve
131 Lowest Common Ancestor (LCA) algorithm (36, 37) and are thus assigned to
132 different taxonomic ranks based on their specificity. The default version of MALT is
133 intended for the analysis of metagenomic datasets deriving from modern DNA, and

134 thus it was not designed to accommodate the specific requirements of aDNA analyses.
135 In particular, aDNA damage that manifests as miscoding lesions in sequenced
136 products can lead to an increased number of mismatches, and extensive damage has
137 the potential to prevent alignment or alter taxonomic assignment. Loss of target reads
138 due to DNA damage can hamper species detection as aDNA studies usually begin
139 with shallow sequence data during initial evaluations of sample quality. In addition,
140 archaeological remains often show low DNA yields, and library amplification can
141 result in a high number of PCR duplicates, which can falsely inflate quantitative
142 estimates of taxa.
143
144 To account for such shortcomings, we introduce a modified version of MALT that is
145 specifically tailored to the analysis of aDNA data. In this modified version, PCR
146 duplicates are removed by eliminating reads identical to those already aligned. In
147 addition, reads are optionally filtered for a minimum Wootton & Federhen complexity
148 (38) in order to remove low complexity reads. Furthermore, to accommodate aDNA
149 damage during alignment, C>T substitutions are ignored in the first five positions
150 from the 5'-end and G>A substitutions are ignored in first five positions from the 3'-
151 end.
152
153 **MaltExtract and post-processing**
154 The core of HOPS is formed by the newly developed MaltExtract module. Without
155 MaltExtract the result files produced by MALT (RMA6 format) can only be evaluated
156 manually with the metagenomic analysis tool MEGAN (39). Such analysis becomes
157 infeasible when working with large data sets, in which each sample must be
158 separately searched for a long list of candidate organisms, a process that is both
159 laborious and prone to user error. MaltExtract provides an automated approach for the
160 assessment of the alignment information stored in RMA files generated by MALT. It
161 automatically retrieves and assesses information on various evaluation criteria for all
162 taxonomic nodes that match a given list of target species.
163
164 MaltExtract obtains information on edit distance, read length distribution, coverage
165 distribution and alignment mismatch patterns in order to identify and authenticate the
166 presence of species-specific ancient DNA. Furthermore, MaltExtract allows data

filtering for maximum read length, minimum percent identity, minimum complexity, and aDNA damage patterns.

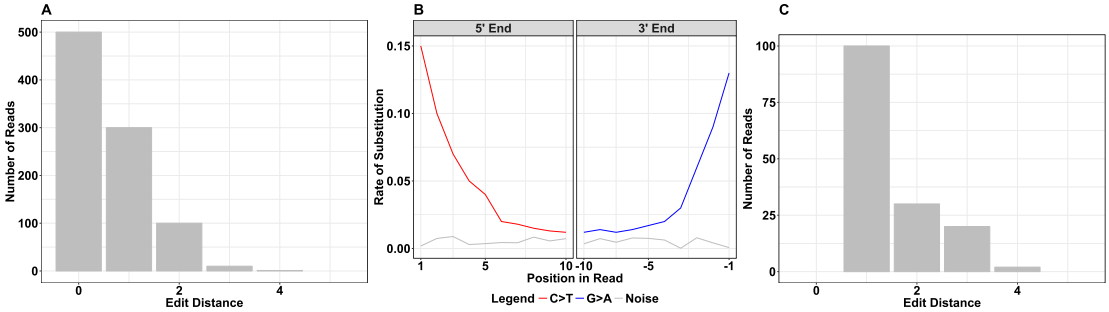


Figure 2. Post-processing steps in HOPS. Three hierarchical post-processing steps are used in HOPS. (A) First, the edit distance distribution is required to show a decline (black lines). (B) Second, the alignments are assessed for C>T and G>A mismatches typical for aDNA; by default, any such damage is considered sufficient. (C) Third, the edit distance distribution of reads showing damage is evaluated.

Accuracy in taxonomic read assignment is evaluated in a three-step procedure that includes ancient authentication criteria (Figure 2). The first step evaluates the read assignment to a taxonomic node. Incorrect read assignments can occur when databases are incomplete: many species in a metagenomic sample may have no representative reference genome in the database, and hence become erroneously assigned to the closest genetic match, which could belong to a different species, or even genus. Mapping to an incorrect species generally results in an increased number of mismatches across the read that is evident in the edit distance distribution (Figure 2A). By contrast, if the sequenced reads are assigned to the correct reference species, the edit distance distribution should continuously decline, with most of the reads showing no or only a few mismatches, mostly resulting from aDNA damage or evolutionary divergence of the modern reference from the ancient genome. We summarize the shape of the edit distance distribution by a score we term *negative difference proportion (-Δ%),* which leverages the difference in sequencing read counts between neighboring mismatch categories (Figure S1). The -Δ% takes values between 0 and 1, where 1 indicates a strictly declining edit distance distribution. While true positives have a -Δ% of 1 when enough endogenous species-specific sequencing reads are present, we use a threshold of -Δ% > 0.9 to account for possible

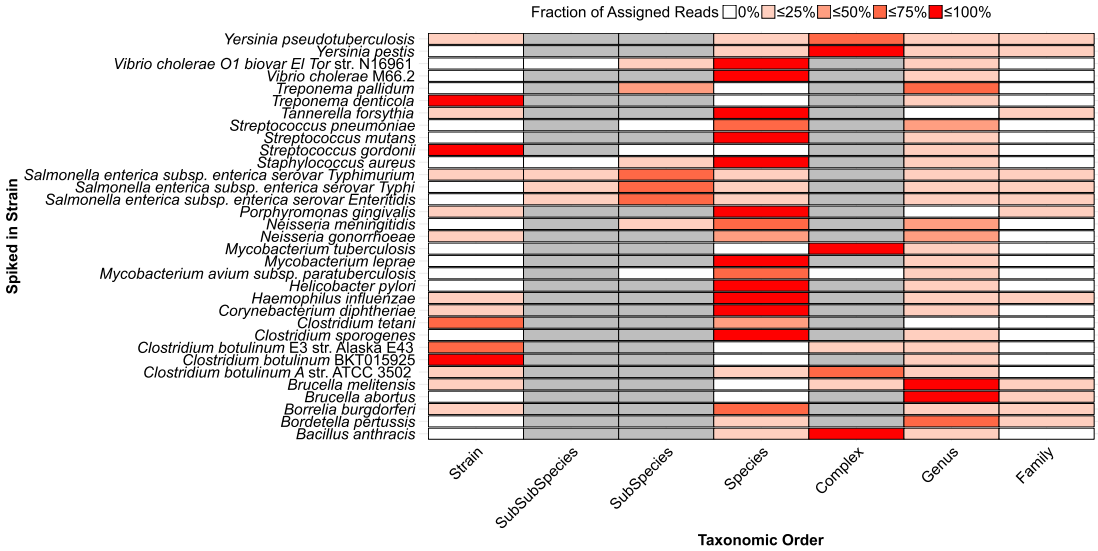
perturbations due to stochasticity in the edit distance distribution when few reads (~10-20) are present. As such, this permits the detection of even low abundant taxa.

In a second step, the ancient origin of the DNA is evaluated through analysis of DNA miscoding lesion patterns (Figure 2B). The most prominent modification observed is deamination of cytosine into uracil, which is read as a thymine by the polymerase. This leads to an overrepresentation of C>T substitutions at the 5' end and correspondingly G>A substitutions at the 3' end (34, 40). Evaluation of damage patterns is mandatory in any ancient DNA study. MaltExtract reports the rates of substitutions for the leading and trailing 10 positions of the read alignment. The default post-processing settings require only a single miscoding lesion to be present in at least one read to qualify as exhibiting damage. This maximizes sensitivity and allows authentication to function largely independently of read depth.

As a third and final criterion, we evaluate the accuracy of taxonomic assignment for all aligned reads exhibiting aDNA damage. For this we assess again the edit distance distribution using the -Δ% score, but now this is only performed for damaged reads (Figure 2C). In this step, a greater number of assigned reads (>100) is required for reliable edit distance evaluation due to the fact that not all ancient reads are expected to exhibit damage.

The MaltExtract output is saved in a structured output folder with a summary file of the processed input and subfolders for each evaluation criterion. The post-processing tool generates a summary highlighting which of the target species passed one or more evaluation criteria for each sample, as well as detailed diagnostic plots displaying the evaluation criteria for each supported target species (Figure S2).

223 **Assessment of taxonomic assignment on simulated data**



224

225 **Figure 3. Assignment of simulated reads to taxonomic levels for 33 bacterial**
226 **pathogens.** The fraction of simulated reads (red gradient) per reference (y axis) assigned to a
227 specific node across different levels of the taxonomy (x axis). The levels of taxonomy not defined for a
228 species are shown in grey.

229

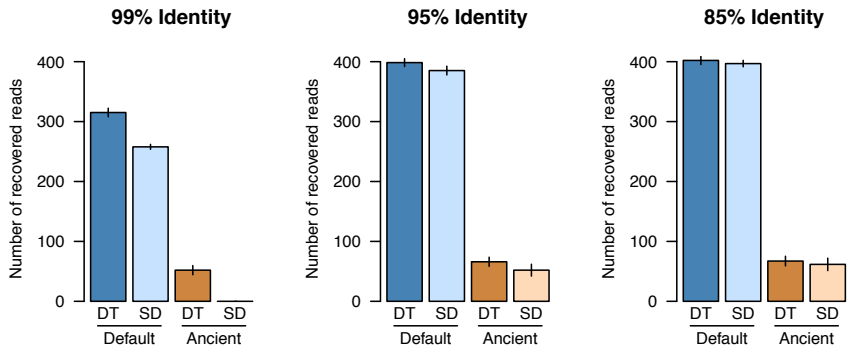
230 The naïve LCA algorithm (36), which is part of HOPS, assigns reads to different
231 taxonomic levels depending on the specificity of sequence matches. Taxonomic
232 assignment thus depends on the structure of the underlying reference database, and it
233 is critical to understand the expected taxonomic placement of sequenced reads from
234 each microbial pathogen in order to successfully identify them. To analyze the
235 taxonomic placement of a test set of 33 bacterial pathogens and to assess the
236 performance of HOPS we simulated sequencing reads that included artificial DNA
237 damage and spiked them into dentine, dental calculus, bone and soil metagenomic
238 backgrounds (see Table 1).

239

240 Applying the HOPS pipeline, we recovered 98% of the simulated reads for 32 of the
241 33 bacterial taxa of interest (see Figure 3). The one exception was *Mycobacterium*
242 *avium subspecies paratuberculosis K10* for which 23% of simulated reads were
243 assigned to an incorrect *Mycobacterium avium subspecies paratuberculosis* strain.
244 Our analysis shows that in most cases the taxonomic levels "species" and "complex"
245 (e.g. *Mycobacterium tuberculosis* complex and *Yersinia pseudotuberculosis* complex)
246 correctly accumulate the vast majority of the simulated pathogen reads. However,

247 noteworthy exceptions were *Brucella abortus*, *Brucella melitenis* and *Bordetella*
248 *pertussis*. Upon further investigation, we found that many species within the genera
249 *Brucella* and *Bordetella* show a high degree of sequence similarity, thus causing the
250 majority of the reads deriving from these pathogens to be assigned at the genus level.
251 By contrast, read assignment was found to be very specific for five taxa (*Treponema*
252 *denticola* ATCC 35405, *Clostridium tetani* E89, *Clostridium botulinum* E3 str. Alaska
253 E43*, Streptococcus gordonii* str. Challis substr. CH1 and *Clostridium botulinum*
254 BKT015925), resulting in the majority of reads deriving from these taxa to be
255 correctly assigned at the strain level. For *Salmonella enterica* subsp. *enterica* most
256 reads were assigned at the subspecies level. The results of this test provide a guide for
257 the levels of taxonomic identification that should be considered when searching for
258 any of the 33 queried bacterial species in experimental ancient datasets.

259 **Optimization of MALT for aDNA**



260
261 **Figure 4** Comparison of the number of successfully recovered *Y. pestis* reads using standard (SD) and
262 damage tolerant (DT) MALT with minimum percent identities of (A) 85%, (B) 95% and (C) 99%.
263 Shown are the recovered reads from the "default" (all reads) and "ancient" (reads with damage) modes
264 in MALT, with the same 500 reads being spiked into the metagenomic backgrounds. Error bars show
265 the standard error of five independent technical replicates for each analysis.
266
267 Because MALT was designed for taxonomic binning of modern genetic data,
268 adapting it to be used on aDNA required altering the original MALT implementation
269 to tolerate terminal substitutions consistent with aDNA damage so that they would not
270 interfere with the percent identity filter. To evaluate the efficacy of this modification,
271 we compared the performance of the modified, damage tolerant version of MALT to
272 the default version using simulated *Y. pestis* data with high damage (~40%) and three
273 different percent identity filters: 85%, 95% and 99% (Figure 4).

274 As expected, the greatest difference was observed when applying the stringent 99%
275 identity filter, for which the damage tolerant MALT version recovered ~20% more
276 reads than the standard MALT version. Additionally, only the modified version was
277 able to recover reads with simulated damage under these parameters. At 95% identity,
278 only a small difference could be observed between the two MALT versions, while
279 results were almost identical at an 85% identity level. Taken together, the damage
280 tolerant MALT version provides an advantage when searching for a given pathogen
281 using stringent filtering criteria.

282 Performance comparison of HOPS, Kraken and MIDAS on simulated 283 data

284 We next evaluated the performance of HOPS by comparing it to two commonly used
285 metagenomics profiling tools: MIDAS (30), a marker gene-based taxonomic
286 classifier, and Kraken (31), which performs taxonomic classification based on k-mer
287 matching to a database of complete genomes. The marker gene database of MIDAS
288 lacked representation for *Yersinia pseudotuberculosis, Bordetella pertussis* and
289 *Brucella melitensis.* Therefore, MIDAS could only be evaluated for 30 of the 33
290 bacterial pathogens in the simulated data sets. For Kraken, we downloaded the
291 bacterial database, which lacked a reference genome to *Clostridium sporogenses*.
292
293 HOPS consistently detected all 33 pathogens in all backgrounds and among replicates
294 with as few as 50 reads (see Figure 5A). Kraken failed to identify *Brucella abortus*
295 and *Mycobacterium tuberculosis* in some replicates with only 50 simulated pathogen
296 reads, but otherwise had a sensitivity of 100%; however, it was prone to a high false
297 positive rate (see below). The sensitivity of MIDAS was far lower than for Kraken
298 and HOPS. Even with 5000 simulated pathogen reads for each species, MIDAS
299 detected only 11 of the 30 possible bacterial pathogens. This can be explained by the
300 limited sensitivity of marker gene based approaches, which require relatively high
301 sequencing coverage in order to ensure adequate representation of the marker genes
302 needed for identification. This is further evident as MIDAS' sensitivity is more
303 heavily influenced by the number of simulated reads than Kraken and HOPS.
304

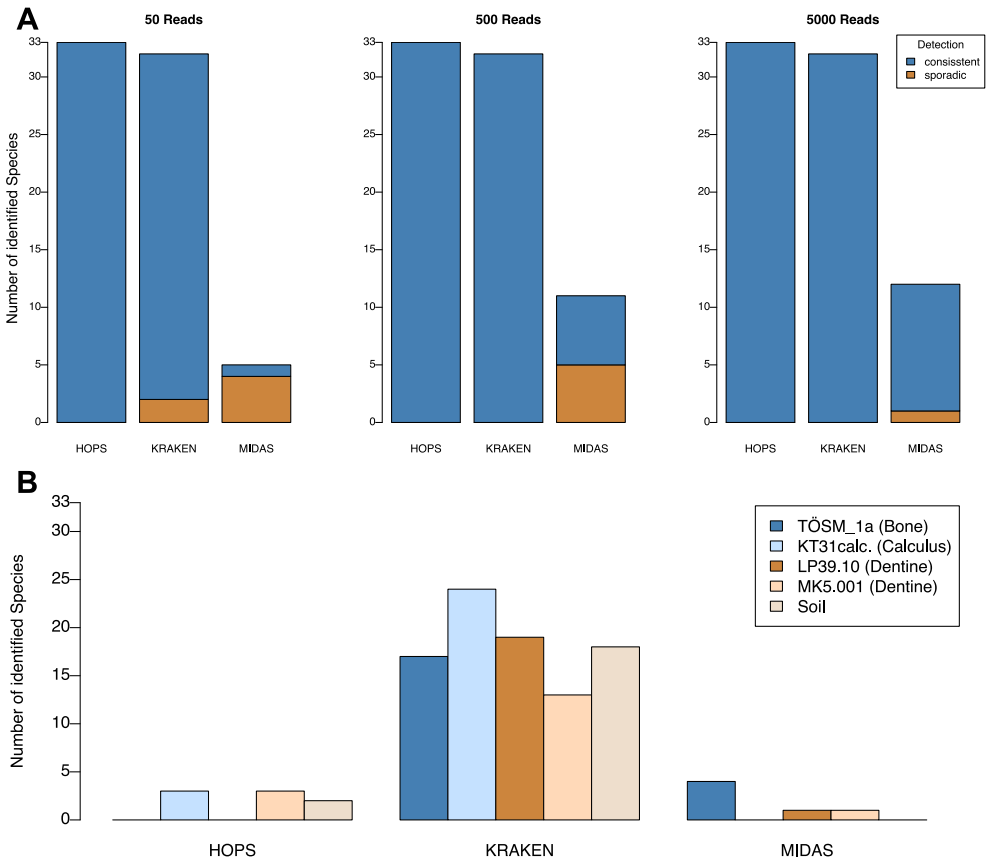


Figure 5. Performance comparison of HOPS, Kraken and MIDAS. (A) HOPS outperforms other tools, successfully and consistently identifying all 33 target bacteria, even when represented by as few as 50 reads. (B) Number of target species identified in the metagenomic background files (negative controls) for HOPS, Kraken and MIDAS.

Negative controls

To assess false positive assignments, we queried all five metagenomic datasets for detectable signatures of the 33 test bacterial pathogens using HOPS, Kraken and MIDAS in the five metagenomic backgrounds prior to the addition of simulated pathogen reads. Kraken showed the highest susceptibility to false positives (see Figure 5B; Table S1). In this analysis, Kraken detected 24 (73%) pathogens in calculus, 19 (58%) in dentine, 13 (39%) in bone and 18 (55%) in soil. Most problematically, *Mycobacterium tuberculosis* and *Bordetella pertussis* were detected by Kraken in every metagenomic background.

Unexpectedly, MIDAS detected oral streptococci, *Tannerella forsythia, Treponema denticola* and *Porphyromonas gingivalis* in the dentine samples but not in calculus, where they are normally found. Overall, MIDAS produced fewer identifications than

324 Kraken, but such a result is expected given its reliance on marker gene-based
325 detection, which limits identification to only abundant taxa.
326
327 HOPS detected four test pathogens in the metagenomic background datasets:
328 *Clostridium tetani* (soil), *Streptococcus mutans* (calculus, dentine), *Treponema*
329 *denticola* (calculus, dentine), and *Porphyromonas gingivalis* (calculus only). Because
330 *C. tetani* is ubiquitous in soil, and all other detected bacteria are commensals of the
331 human oral cavity, their identification via both MIDAS and HOPS likely reflects true
332 positives. Taken together, HOPS and MIDAS have a lower tendency toward false
333 positive assignments. Kraken's increased vulnerability for aberrant assignments likely
334 relates to the absence of an alignment step, which is necessary for reliable species
335 evaluation in both modern and ancient contexts.
336

337 **Positive Controls**

338 In addition to performing tests using simulated data, we also tested HOPS, Kraken
339 and MIDAS on 25 ancient metagenomic datasets known to be positive for bacterial
340 pathogens (Table 2). They consisted of both shotgun and capture data and they varied
341 in sequencing depth in accordance with experimental conditions and method of data
342 generation.
343
344 HOPS and Kraken share 100% sensitivity for the detection of target bacterial
345 pathogens in every sample. By contrast, MIDAS only detected the correct bacterial
346 pathogen in 22 out of 25 samples. Again, MIDAS sensitivity was likely reduced due
347 to the marker gene-based approach. These results highlight the advantage of whole-
348 genome based approaches like MALT and Kraken that take advantage of every
349 sequenced read.

350 **Runtimes**

351 To calculate the runtime for each program, we used simulated metagenomic files that
352 each contained five million sequencing reads (see Methods). For each file, HOPS
353 required an average of 3307±820 seconds for the MALT step, 16±1 seconds for the
354 MaltExtract step and 1±0 seconds for post processing, for a total of approximately 55
355 minutes of analysis time per file. Kraken took on average 72±16 seconds to run
356 *Kraken_alignment* and 22±3 for *Kraken_translate*, a total of 1.5 minutes, and the

357 MIDAS pipeline processed each file in an average of 4±1 seconds. HOPS by far
358 required the highest runtimes of the three tools, but most of this time was required for
359 sequence alignment, a step that, although time consuming, increases detection
360 sensitivity, reduces false positives, and enables the authentication of aDNA reads.
361

362 **Discussion**

363 The field of archaeogenetics faces several challenges, such as the low amount of
364 endogenous target DNA, the highly degraded nature of the DNA, and an unknown
365 and diverse metagenomic background signal that accumulates during decomposition
366 and centuries spent in a depositional environment. This makes reliable identification
367 and authentication of genuine ancient DNA challenging, particularly when targeting
368 bacterial DNA that is usually only present in small amounts. Furthermore, many
369 bacterial pathogens have close relatives in soil, which necessitates meticulous care
370 when making pathogen identifications.
371
372 HOPS provides an automated pipeline for high-throughput ancient bacterial species
373 detection and authentication from metagenomic sequencing data. We compare HOPS
374 to Kraken and MIDAS, two widely used methods for estimating both the presence
375 and abundance of bacterial taxa in metagenomic data. These tools, however, have
376 limited application to the specific challenges of aDNA in terms of degradation and
377 chemical modifications manifest as miscoding lesions. Our analyses highlight the
378 need for a pathogen identification pipeline that accommodates qualities of aDNA data
379 and includes an essential and robust authentication for all ancient read assignments.
380 HOPS provides a fast, reliable, and user-friendly solution to these established
381 limitations.
382
383 HOPS was tested on simulated ancient pathogen DNA reads, and it successfully
384 detected all targeted species spiked into metagenomic backgrounds with as few as 50
385 pathogen reads, representing less than 0.00001 % of the total dataset. In this context,
386 our modified version of MALT, which tolerates mismatches resulting from DNA
387 degradation, prevents a decrease in sensitivity even in cases of heavily damaged
388 aDNA. We demonstrate that the marker gene-based metagenomic profiling tool
389 MIDAS had a much lower sensitivity for pathogen detection compared to HOPS,
390 especially for low coverage data, which is typical of ancient DNA screening datasets.
391 Although the sensitivity of Kraken was similar to HOPS, and while Kraken's k-mer
392 matching is considerably faster than the precise alignments used in HOPS, Kraken is
393 incapable of validating species assignment and aDNA authenticity, and thus has a

394 lower specificity. This is most clearly demonstrated by our analysis of a metagenomic
395 soil sample in which Kraken detected numerous false positives, including
396 *Mycobacterium tuberculosis* and *Bordetella pertussis* (whooping cough). This is
397 likely due to many soil dwelling bacteria that harbor genetic similarities to these
398 pathogens, such as diverse mycobacerial species and *B. petrii,* a close relative to *B.*
399 *pertussis* that is a common constituent of environmental datasets. These effects are
400 further compounded by the fact that many environmental microbes have not been
401 genomically characterised and are not part of any reference database, which only
402 increases the potential of false assignments to well-sequenced pathogens. The
403 alignment-based validation procedure implemented in HOPS minimises such false
404 positive assignments, and thus offers greater accuracy in pathogen identification
405 during screening when environmental backgrounds comprise the dominant molecular
406 signal.
407
408 A previously published pipeline for the assessment of metagenomic data in
409 archaeogenetics is metaBIT (41). It implements a variety of methods for the detailed
410 assessment of metagenomic composition, which also includes validation of aDNA
411 damage patterns. As metaBIT is based on MetaPhlAn (42), which employs a marker
412 gene based approach in the initial detection step similar to MIDAS, pathogens in low
413 abundance could be missed in its initial steps when applied to shallow sequencing
414 data. An integrated approach combining HOPS and metaBIT might be a promising
415 future strategy for a detailed characterization of microbiomes while at the same time
416 proving a high level of sensitivity for the detection of pathogens. In particular, the
417 analysis of ancient samples that preserve their original microbiome signature, such as
418 dental calculus (43) or coprolites (44) would benefit from a combined application of
419 both methodologies, by using metaBIT to assess the microbial make up and using
420 HOPS for more in depth species authentication.
421
422 For all taxonomic classifiers, correct assignment of metagenomic reads is strongly
423 dependent on the quality of the underlying reference sequences. Currently we use a
424 curated database for MALT that contains completed reference sequences and
425 assemblies for bacteria from RefSeq (December 2016). Database sizes are constantly
426 increasing, but much of this growth derives from the addition of redundant sequence
427 data from model organisms, which also creates biases. In this context, methodologies

428 such as SPARSE (45) aim to mitigate against database redundancy by hierarchically
429 structuring reference sequences, which could be employed to further improve HOPS'
430 specificity and runtime.
431
432 In addition, analysis of our simulated dataset allowed for insights into the taxonomic
433 structure of each of the bacterial pathogens in our target list. It became apparent that
434 for some targets the taxonomic species level is not sufficient for identification. This
435 applies to historically important pathogens such as *Y. pestis* or *M. tuberculosis*. Here,
436 evaluation of a higher taxonomic level such as complex is more reliable, while in the
437 case of *Salmonella typhi* (typhoid fever) a lower level (subspecies) is favorable.
438 Therefore, our simulations provide a valuable resource for the optimization of
439 pathogen screening approaches in general.
440
441 Here, HOPS was evaluated for its success in screening for bacterial pathogens.
442 Because the reference database is user defined and can amended to include, for
443 example, the NCBI full nucleotide collection (46) or hand-curated sets of reference
444 genomes, tremendous flexibility exists in molecular detection, which could extend to
445 viruses, fungi, and eukaryotes.
446

447 **Conclusions**

448 We present a fast, reliable, and user-friendly computational pathogen screening
449 pipeline for ancient DNA that has the flexibility of handling large datasets. HOPS
450 successfully identifies both simulated and actual ancient pathogen DNA within
451 complex metagenomic datasets, exhibiting a higher sensitivity than MIDAS and with
452 fewer false positives than Kraken. HOPS provides a high level of automatization that
453 allows for the screening of thousands of datasets with very little hands-on time, and it
454 offers detailed visualizations and statistics at each evaluation step, enabling a high
455 level of quality control and analytical transparency. HOPS is a powerful tool for high-
456 throughput pathogen screening in large-scale archaeogenetic studies, producing
457 reliable and reproducible results even from remains with exceptionally low levels of
458 pathogen DNA. Such qualities make HOPS a valuable tool for pathogen detection in
459 the rapidly growing field of archaogenetics.

460 # Methods

461 ## Implementation of MaltExtract

462 MaltExtract is implemented in Java. It integrates parts of MEGAN's (39) source code
463 for accessing the RMA file structure and functions from *forester*
464 (https://github.com/cmzmasek/forester) for traversing the taxonomic tree.

465 ## Simulating data to analyse read assignment using the MALT LCA
466 algorithm

467 Depending on the database structure and sequence similarity between reference
468 sequences, the naïve LCA (36) algorithm will assign reads to different taxonomic
469 units. To inquire how reads are assigned to the taxonomic tree for 33 bacterial
470 pathogens (Table S2), we simulated ancient pathogen DNA reads using gargammel
471 (47) and spiked them into five ancient metagenomic background datasets obtained
472 from bone, dentine, dental calculus and soil (Table 1). The simulated reads carry a
473 unique identifier in their header in order to differentiate them from metagenomic
474 background sequences, which exhibit either full damage patterns or attenuated
475 damage patterns following UDG-half treatment (48). To simulate aDNA damage in
476 the pathogen sequences, we applied damage profiles obtained from previously
477 published ancient *Yersinia pestis* genomes with (13) and without UDG-half (18)
478 treatment. Simulated reads were processed with EAGER (49) and spiked into the
479 metagenomic backgrounds in different amounts (50, 500 or 5000 reads). For each
480 metagenomic background, a typical screening sequencing depth of five million reads
481 were used.

482

483 Table 1. Metagenomic backgrounds used for simulated data sets

ID	Source	Age (Period)	Treatment	Reference
KT31calc	Calculus	Medieval	No UDG	(50)
LP39.10	Dentine	2920-2340 BCE	No UDG	(51)
MK5.001	Dentine	3348-3035 BCE 3619-3366 BCE	UDG half	(52)
TÖSM_1a	Bone	6000-5500 BCE	UDG half	(53)
Soil	Soil	-	No UDG	(25)

484 **Evaluation of the damage tolerant version of MALT**

485 To preserve damage patterns when mapping reads with MALT, we modified the
486 source code and compared the performance of the modified and default versions.
487 We therefore created with gargammel (47) test samples that show twice the amount of
488 damage (~40%) usually found in ancient samples (13). Here, we compare both
489 MALT versions for the bacterial pathogen *Yersinia pestis* (CO92 reference). Both
490 versions of MALT were tested with 85%, 95% and 99% minimum percent identity
491 filtering, to investigate the effects of percent identity filtering on the read alignment of
492 aDNA reads.

493 **Comparison of HOPS to Kraken and MIDAS**

494 HOPS was compared to two metagenomic taxonomic classification tools: Kraken (31)
495 and MIDAS (30). We only executed the first step of MIDAS that matches reads to the
496 marker gene database to determine species abundance. This step was executed on 32
497 cores with default parameters. The first step is sufficient, as any species undetected in
498 this step would not be detected in the remaining ones. Kraken was set to use 32 cores
499 to align the sample data against its reference database with the preload parameter to
500 load the entire database into memory before starting k-mer alignment. In a second
501 step kraken-translate was executed to transform taxonomy ids into proper species
502 names. For Kraken and MIDAS, we judged a pathogen as correctly identified if at
503 least one read matches to a strain of the correct species to account for the differences
504 in the database contents, methodologies and output formats.

505 **Databases**

506 In our study, HOPS uses a database containing all complete prokaryotic reference
507 genomes obtained from NCBI (December 1st 2016) with entries containing 'multi'
508 and 'uncultured' removed (13 entries). In total, 6,249 reference genomes are included
509 in the database. For Kraken we downloaded the bacterial database with Kraken's
510 kraken-build script (June 1 2017). The Kraken database contains no strain references
511 for *Clostridium sporogenses.* Otherwise it contains at least one reference for all of the
512 simulated bacterial pathogens (Table S2). For MIDAS we used the default reference
513 database (May 24 2016), which contained no representation of *Yersinia*
514 *pseudotuberculosis, Bordetella pertussis* and *Brucella melitensis*.

515 **Positive controls**

516 We compare the sensitivity and specificity of HOPS, MIDAS and Kraken using 25
517 metagenomic datasets previously shown to be positive for one of four microbial
518 pathogens: *Yersinia pestis, Mycobacterium tuberculosis*, *Salmonella enterica* and
519 *Heliobacter pylori* (Table 2). These positive control samples represent real
520 metagenomic data and therefore contain an unknown number of modern species in
521 addition to the actual recovered bacterial pathogen. Read counts across all samples
522 ranged from 70,897 to 7,000,000 reads. While most datasets were generated by
523 shotgun library screening, four datasets were enriched for pathogen DNA prior to
524 sequencing using DNA capture methods. For all captured datasets and a subset of
525 shotgun datasets, DNA was treated with UDG prior to library construction to remove
526 DNA damage. Both types of datasets were included to evaluate the performance of
527 HOPS on samples with different levels of DNA damage and pathogen abundance.
528

529 Table 2 Metagenomic samples used as positive controls

ID	Reconstructed Bacteria	Sequencing reads	Data type	Reference
10C	Salmonella enterica	1,017,400	Shotgun	(25)
35C	Salmonella enterica	986,908	Shotgun	(25)
RK1001.C0101	Yersinia pestis	7,023,370	Shotgun	(17)
GEN_72	Yersinia pestis	7,663,408	Shotgun	(17)
549_O	Yersinia pestis	1,520,471	Shotgun	(16)
JK3031UDG	Yersinia pestis	4,059,016	Shotgun (UDG)	(16)
JK2370UDG	Yersinia pestis	52,858,027	Shotgun (UDG)	(16)
RT6	Yersinia pestis	6,706,316	Shotgun (UDG)	(18)
1343UnTal85	Yersinia pestis	3,462,216	Shotgun	(17)
6Post	Yersinia pestis	2,546,695	Shotgun	(17)
KunilaII	Yersinia pestis	1,007,417	Shotgun	(17)
RISE00	Yersinia pestis	6,000,000	Shotgun	(13)
RISE139	Yersinia pestis	6,000,000	Shotgun	(13)
RISE386	Yersinia pestis	6,000,000	Shotgun	(13)
RISE397	Yersinia pestis	6,000,000	Shotgun	(13)
RISE505	Yersinia pestis	6,000,000	Shotgun	(13)

RISE509	Yersinia pestis	6,000,000	Shotgun	(13)
RISE511	Yersinia pestis	6,000,000	Shotgun	(13)
54	Mycobacterium tuberculosis	70,897	Shotgun	(21)
58	Mycobacterium tuberculosis	114,555	Shotgun	(21)
64	Mycobacterium tuberculosis	160,310	Shotgun	(21)
54	Mycobacterium tuberculosis	5,000,000	Capture (UDG)	(21)
58	Mycobacterium tuberculosis	5,000,000	Capture (UDG)	(21)
64	Mycobacterium tuberculosis	5,000,000	Capture (UDG)	(21)
P1P2	Helicobacter pylori	5,000,000	Capture (UDG)	(27)

Runtimes

To calculate the runtimes for HOPS, Kraken and MIDAS, we used a subset of the simulated files. The subset consisted of all metagenomic background datasets spiked with 5000 reads without technical replicates resulting in a total of 330 metagenomic files. A total of 64 cores and 700 GB of RAM were allocated to each program.

Declarations

Availability of data and material

The complete source code of HOPS is available from GitHub (https://github.com/rhuebler/HOPS).

Competing interests

The authors declare that they have no competing interests

Funding

This research was funded by the Max Planck Society. The funding body had no involvement in the design of the study, collection, analysis, and interpretation of data or in writing the manuscript.

Authors' contributions

RH, FMK and AH conceived the study. RH, FMK, CW, KIB, JK and AH designed experiments. RH and FMK implemented software. RH, FMK and AH performed analyses. RH, FMK and AH wrote the manuscript with contributions from all coauthors.

Acknowledgements

We thank the Department of Archaeogenetics of the Max Planck Institute for the Science of Human History and Julian Susat for beta testing and helpful discussions.

564 **References**

565 1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett C, Knight R, Gordon JI.
566 The human microbiome project: exploring the microbial part of ourselves in a
567 changing world. Nature. 2007;449(7164):804.
568 2. Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, et al. A
569 holistic approach to marine eco-systems biology. PLoS Biol. 2011;9(10):e1001177.
570 3. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes
571 and aspirations. BMC Biol. 2014;12(1):69.
572 4. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al.
573 Massive migration from the steppe was a source for Indo-European languages in
574 Europe. Nature. 2015;522(7555):207-11.
575 5. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The
576 Simons Genome Diversity Project: 300 genomes from 142 diverse populations.
577 Nature. 2016;538(7624):201-6.
578 6. Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, et al.
579 Recalibrating Equus evolution using the genome sequence of an early Middle
580 Pleistocene horse. Nature. 2013;499(7456):74-8.
581 7. Schlebusch CM, Skoglund P, Sjodin P, Gattepaille LM, Hernandez D, Jay F,
582 et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex
583 African history. Science. 2012;338(6105):374-9.
584 8. Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak
585 M, et al. Reconstructing Prehistoric African Population Structure. Cell.
586 2017;171(1):59-71 e21.
587 9. Warinner C, Herbig A, Mann A, Fellows Yates JA, Weiss CL, Burbano HA,
588 et al. A Robust Framework for Microbial Archaeology. Annu Rev Genomics Hum
589 Genet. 2017;18(0):321-56.
590 10. Key FM, Posth C, Krause J, Herbig A, Bos KI. Mining Metagenomic Data
591 Sets for Ancient DNA: Recommended Protocols for Authentication. Trends Genet.
592 2017;33(8):508-20.
593 11. Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N,
594 Coombes BK, et al. A draft genome of Yersinia pestis from victims of the Black
595 Death. Nature. 2011;478(7370):506-10.
596 12. Wagner DM, Klunk J, Harbeck M, Devault A, Waglechner N, Sahl JW, et al.
597 Yersinia pestis and the plague of Justinian 541-543 AD: a genomic analysis. Lancet
598 Infect Dis. 2014;14(4):319-26.
599 13. Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjogren KG, et
600 al. Early divergent strains of Yersinia pestis in Eurasia 5,000 years ago. Cell.
601 2015;163(3):571-82.
602 14. Feldman M, Harbeck M, Keller M, Spyrou MA, Rott A, Trautmann B, et al. A
603 High-Coverage Yersinia pestis Genome from a Sixth-Century Justinianic Plague
604 Victim. Mol Biol Evol. 2016;33(11):2911-23.
605 15. Bos KI, Herbig A, Sahl J, Waglechner N, Fourment M, Forrest SA, et al.
606 Eighteenth century Yersinia pestis genomes reveal the long-term persistence of an
607 historical plague focus. Elife. 2016;5:e12994.
608 16. Spyrou MA, Tukhbatova RI, Feldman M, Drath J, Kacki S, Beltran de Heredia
609 J, et al. Historical Y. pestis Genomes Reveal the European Black Death as the Source
610 of Ancient and Modern Plague Pandemics. Cell Host Microbe. 2016;19(6):874-81.

611 17. Valtueña AA, Mittnik A, Key FM, Haak W, Allmńe R, Belinskij A, et al. The
612 Stone Age plague and its persistence in Eurasia. Current biology. 2017;27(23):3683-
613 91. e8.
614 18. Spyrou MA, Tukhbatova RI, Wang CC, Valtuena AA, Lankapalli AK,
615 Kondrashin VV, et al. Analysis of 3800-year-old Yersinia pestis genomes suggests
616 Bronze Age origin for bubonic plague. Nat Commun. 2018;9(1):2234.
617 19. Schuenemann VJ, Singh P, Mendum TA, Krause-Kyora B, Jager G, Bos KI, et
618 al. Genome-wide comparison of medieval and modern Mycobacterium leprae.
619 Science. 2013;341(6142):179-83.
620 20. Schuenemann VJ, Avanzi C, Krause-Kyora B, Seitz A, Herbig A, Inskip S, et
621 al. Ancient genomes reveal a high diversity of Mycobacterium leprae in medieval
622 Europe. PLoS Pathog. 2018;14(5):e1006997.
623 21. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, et al. Pre-
624 Columbian mycobacterial genomes reveal seals as a source of New World human
625 tuberculosis. Nature. 2014;514(7523):494-7.
626 22. Kay GL, Sergeant MJ, Zhou Z, Chan JZ, Millard A, Quick J, et al. Eighteenth-
627 century genomes show that mixed infections were common at time of peak
628 tuberculosis in Europe. Nat Commun. 2015;6:6717.
629 23. D'Anastasio R, Staniscia T, Milia ML, Manzoli L, Capasso L. Origin,
630 evolution and paleoepidemiology of brucellosis. Epidemiol Infect. 2011;139(1):149-
631 56.
632 24. Kay GL, Sergeant MJ, Giuffra V, Bandiera P, Milanese M, Bramanti B, et al.
633 Recovery of a medieval Brucella melitensis genome using shotgun metagenomics.
634 MBio. 2014;5(4):e01337-14.
635 25. Vagene AJ, Herbig A, Campana MG, Robles Garcia NM, Warinner C, Sabin
636 S, et al. Salmonella enterica genomes from victims of a major sixteenth-century
637 epidemic in Mexico. Nat Ecol Evol. 2018;2(3):520-8.
638 26. Zhou Z, Lundstrom I, Tran-Dien A, Duchene S, Alikhan NF, Sergeant MJ, et
639 al. Pan-genome Analysis of Ancient and Modern Salmonella enterica Demonstrates
640 Genomic Stability of the Invasive Para C Lineage for Millennia. Curr Biol.
641 2018;28(15):2420-8 e10.
642 27. Maixner F, Krause-Kyora B, Turaev D, Herbig A, Hoopmann MR, Hallows
643 JL, et al. The 5300-year-old Helicobacter pylori genome of the Iceman. Science.
644 2016;351(6269):162-5.
645 28. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello
646 EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat
647 Methods. 2010;7(5):335-6.
648 29. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al.
649 MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods.
650 2015;12(10):902-3.
651 30. Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated
652 metagenomics pipeline for strain profiling reveals novel patterns of bacterial
653 transmission and biogeography. Genome Res. 2016;26(11):1612-25.
654 31. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence
655 classification using exact alignments. Genome Biol. 2014;15(3):R46.
656 32. Velsko IM, Frantz LAF, Herbig A, Larson G, Warinner C. Selection of
657 Appropriate Metagenome Taxonomic Classifiers for Ancient Microbiome Research.
658 mSystems. 2018;3(4).

659 33. Jonsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L. mapDamage2.0:
660 fast approximate Bayesian estimates of ancient DNA damage parameters.
661 Bioinformatics. 2013;29(13):1682-4.
662 34. Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prufer K, et al.
663 Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad
664 Sci U S A. 2007;104(37):14616-21.
665 35. Herbig A, Maixner F, Bos KI, Zink A, Krause J, Huson DH. MALT: Fast
666 alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean
667 Iceman. BioRxiv. 2016:050559.
668 36. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic
669 data. Genome Res. 2007;17(3):377-86.
670 37. Droge J, McHardy AC. Taxonomic binning of metagenome samples generated
671 by next-generation sequencing technologies. Brief Bioinform. 2012;13(6):646-55.
672 38. Wootton JC, Federhen S. Statistics of Local Complexity in Amino-Acid-
673 Sequences and Sequence Databases. Computers & Chemistry. 1993;17(2):149-63.
674 39. Huson DH, Beier S, Flade I, Gorska A, El-Hadidi M, Mitra S, et al. MEGAN
675 Community Edition - Interactive Exploration and Analysis of Large-Scale
676 Microbiome Sequencing Data. Plos Comput Biol. 2016;12(6):e1004957.
677 40. Kircher M. Analysis of high-throughput ancient DNA sequencing data.
678 Methods Mol Biol. 2012;840:197-228.
679 41. Louvel G, Der Sarkissian C, Hanghoj K, Orlando L. metaBIT, an integrative
680 and automated metagenomic pipeline for analysing microbial profiles from high-
681 throughput sequencing shotgun data. Mol Ecol Resour. 2016;16(6):1415-27.
682 42. Haft DH, Tovchigrechko A. High-speed microbial community profiling. Nat
683 Methods. 2012;9(8):793-4.
684 43. Warinner C, Speller C, Collins MJ. A new era in palaeomicrobiology:
685 prospects for ancient dental calculus as a long-term record of the human oral
686 microbiome. Philos Trans R Soc Lond B Biol Sci. 2015;370(1660):20130376.
687 44. Warinner C, Speller C, Collins MJ, Lewis CM, Jr. Ancient human
688 microbiomes. J Hum Evol. 2015;79:125-36.
689 45. Zhou Z, Luhmann N, Alikhan N-F, Quince C, Achtman M, editors. Accurate
690 Reconstruction of Microbial Strains from Metagenomic Sequencing Using
691 Representative Reference Genomes. International Conference on Research in
692 Computational Molecular Biology; 2018: Springer.
693 46. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a
694 curated non-redundant sequence database of genomes, transcripts and proteins.
695 Nucleic Acids Res. 2007;35(Database issue):D61-5.
696 47. Renaud G, Hanghøj K, Willerslev E, Orlando L. gargammel: a sequence
697 simulator for ancient DNA. Bioinformatics. 2016;33(4):577-9.
698 48. Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D. Partial uracil-DNA-
699 glycosylase treatment for screening of ancient DNA. Philos Trans R Soc Lond B Biol
700 Sci. 2015;370(1660):20130624.
701 49. Peltzer A, Jager G, Herbig A, Seitz A, Kniep C, Krause J, et al. EAGER:
702 efficient ancient genome reconstruction. Genome Biol. 2016;17(1):60.
703 50. Mann AE, Sabin S, Ziesemer K, Vagene AJ, Schroeder H, Ozga AT, et al.
704 Differential preservation of endogenous human and microbial DNA in dental calculus
705 and dentin. Sci Rep. 2018;8(1):9822.
706 51. Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, et al. The
707 Beaker phenomenon and the genomic transformation of northwest Europe. Nature.
708 2018;555(7695):190.

52. Wang C-C, Reinhold SR, Kalmykov A, Wissgott A, Brandt G, Jeong C, et al. The genetic prehistory of the Greater Caucasus. bioRxiv. 2018:322347.
53. Lipson M, Szecsenyi-Nagy A, Mallick S, Posa A, Stegmar B, Keerl V, et al. Parallel palaeogenomic transects reveal complex genetic history of early European farmers. Nature. 2017;551(7680):368-72.

Supplementary Information

HOPS: Automated detection and authentication of pathogen DNA in archaeological remains

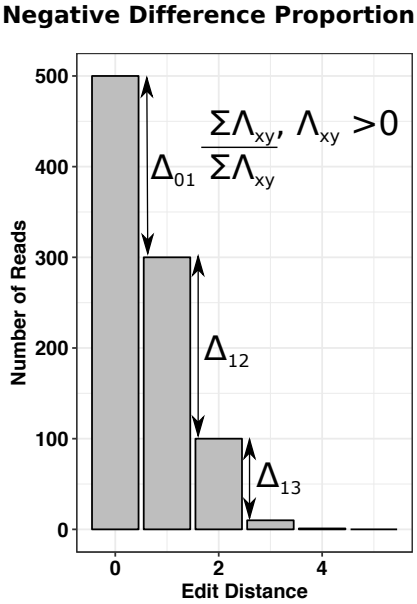
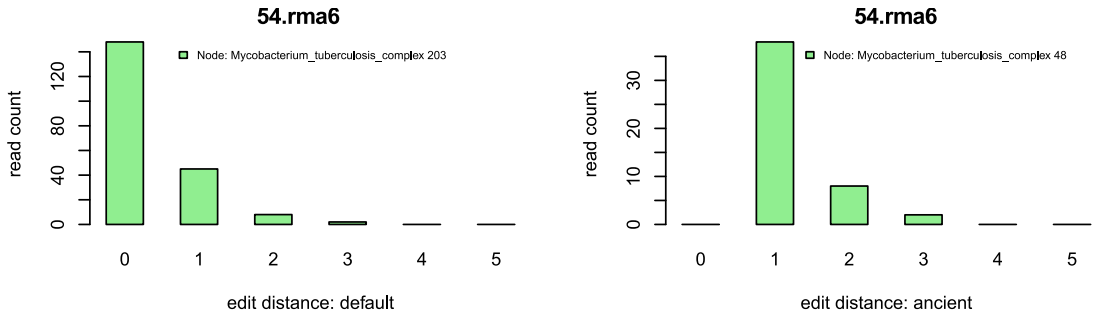
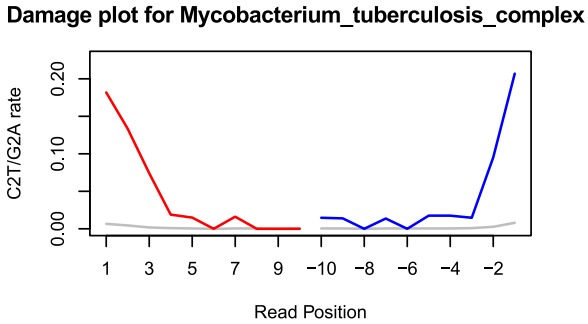


Figure S1 The first and third steps in the HOPS postprocessing protocol require a decline in the Edit Distance distribution.





Node	Mycobacterium_tuberculosis_complex
Top Reference	Mycobacterium_africanum
all reads	203
nonDup	203
readDis	1
nonStacked	203
destacking?	On
C>T_1	0.1818
G>A_−1	0.2069
mean length (sd)	42 (10.451)

Mycobacterium_africanum;_CP010334.1 100%
Mycobacterium_africanum_GM041182;_FR878060.1 099%
Mycobacterium_microti;_CP010333.1 098%
Mycobacterium_tuberculosis;_CP010335.1 098%
Mycobacterium_tuberculosis_CCDC5180;_CP002885.1 098%
Mycobacterium_tuberculosis_CCDC5180;_CP001642.1 098%
Mycobacterium_tuberculosis_CCDC5079;_CP002884.1 098%
Mycobacterium_bovis_BCG_str._Mexico;_CP002095.1 098%
Mycobacterium_tuberculosis_BT2;_CP002882.1 098%
Mycobacterium_tuberculosis_K;_CP007803.1 098%

Figure S2 HOPS summary output for a tuberculosis positive sample. Upper left: Edit distance distribution for all reads assigned to M. tuberculosis. Upper right: Edit distance distribution for assigned reads that show a possible DNA damage signal. Middle left: DNA damage plot for assigned reads. Lower left: Top ten references with percentage of aligned reads. Middle right: Summary statistics for assigned reads.

Table S1 Results for negative controls. For HOPS the step in the post processing that was reached for each species is indicated (0: No detection; 1: detected with good edit distance distribution; 2: additionally indication for damage; 3: additionally good edit distance distribution for damaged reads). For Kraken the number of k-mers assigned to the species and for MIDAS the number of assigned reads is listed.

Species	bone_ TOSM1a	calculus_ KT31calc	dentine_M K5.001	dentine_LP39 .10	soil	Software
Bacillus anthracis	0	0	0	0	0	HOPS
Bordetella pertussis	0	0	0	0	0	HOPS
Borrelia	0	0	0	0	0	HOPS

burgdorferi B31						
Brucella abortus	0	0	0	0	0	HOPS
Brucella melitensis	0	0	0	0	0	HOPS
Clostridium botulinum	0	0	0	0	1	HOPS
Clostridium sporogenes	0	0	0	0	0	HOPS
Clostridium tetani	0	0	0	0	1	HOPS
Corynebacterium diphtheriae	0	0	0	0	0	HOPS
Haemophilus influenzae	0	0	0	0	0	HOPS
Helicobacter pylori	0	0	0	0	0	HOPS
Mycobacterium avium subsp. paratuberculosis	0	0	0	0	0	HOPS
Mycobacterium leprae	0	0	0	0	0	HOPS
Mycobacterium tuberculosis	0	0	0	0	0	HOPS
Neisseria gonorrhoeae	0	0	0	0	0	HOPS
Neisseria meningitidis	0	0	0	0	0	HOPS
Porphyromonas gingivalis	0	3	0	0	0	HOPS
Salmonella enterica subsp. enterica	0	0	0	0	0	HOPS
Staphylococcus aureus subsp. aureus	0	0	0	0	0	HOPS
Streptococcus gordonii str.	0	0	0	1	0	HOPS
Streptococcus mutans	0	3	0	2	0	HOPS
Streptococcus pneumoniae	0	2	0	0	0	HOPS
Tannerella forsythia	0	0	0	0	0	HOPS
Treponema denticola	0	0	0	1	0	HOPS
Treponema pallidum subsp. pallidum	0	0	0	0	0	HOPS
Vibrio cholerae	0	0	0	0	0	HOPS
Yersinia pestis	0	0	0	0	0	HOPS
Yersinia pseudotuberculosis	0	0	0	0	0	HOPS
	bone_ TOSM1a	calculus_ KT31calc	dentine_M K5.001	dentine_LP39 .10	soil	Kraken
Bacillus anthracis	0	5	1	0	0	Kraken
Bordetella pertussis	1	12	20	4	10	Kraken
Borrelia burgdorferi B31	0	0	0	0	0	Kraken

Brucella abortus	0	0	0	0	1	Kraken
Brucella melitensis	0	4	0	0	1	Kraken
Clostridium botulinum	40	182	0	21	40	Kraken
Clostridium sporogenes	0	0	0	0	0	Kraken
Clostridium tetani	13	50	1	2	240	Kraken
Corynebacterium diphtheriae	4	79	9	13	11	Kraken
Haemophilus influenzae	2	1617	0	2	3	Kraken
Helicobacter pylori	8	12	0	0	4	Kraken
Mycobacterium avium subsp. paratuberculosis	13	8	39	6	26	Kraken
Mycobacterium leprae	16	19	22	13	21	Kraken
Mycobacterium tuberculosis	53	46	208	55	117	Kraken
Neisseria gonorrhoeae	4	2407	2	3	8	Kraken
Neisseria meningitidis	4	4448	0	9	12	Kraken
Porphyromonas gingivalis	2	13925	0	6	0	Kraken
Salmonella enterica subsp. enterica	8	54	16	24	16	Kraken
Staphylococcus aureus subsp. aureus	1	19	2	4	4	Kraken
Streptococcus gordonii	0	71751	0	15	0	Kraken
Streptococcus mutans	0	401	0	17	0	Kraken
Streptococcus pneumoniae	0	3224	0	0	0	Kraken
Tannerella forsythia	8	31294	14	18	14	Kraken
Treponema denticola	0	25264	0	4	0	Kraken
Treponema pallidum subsp. pallidum	0	0	0	0	0	Kraken
Vibrio cholerae	6	60	1	6	5	Kraken
Yersinia pestis	0	4	0	0	0	Kraken
Yersinia pseudotuberculosis	15	21	7	39	7	Kraken
	bone_ TOSM1a	calculus_ KT31calc	dentine_M K5.001	dentine_LP39 .10	soil	
Bacillus anthracis	0	0	0	0	0	Midas
Bordetella pertussis	0	0	0	0	0	Midas
Borrelia burgdorferi B31	0	0	0	0	0	Midas
Brucella abortus	0	0	0	0	0	Midas

Brucella melitensis	0	0	0	0	0	Midas
Clostridium botulinum	0	0	0	0	0	Midas
Clostridium sporogenes	0	0	0	0	0	Midas
Clostridium tetani	0	0	0	0	0	Midas
Corynebacterium diphtheriae	0	0	0	0	0	Midas
Haemophilus influenzae	0	0	0	0	0	Midas
Helicobacter pylori	0	0	0	0	0	Midas
Mycobacterium avium subsp. paratuberculosis	0	0	0	0	0	Midas
Mycobacterium leprae	0	0	0	0	0	Midas
Mycobacterium tuberculosis	0	0	0	0	0	Midas
Neisseria gonorrhoeae	0	0	0	0	0	Midas
Neisseria meningitidis	0	0	0	0	0	Midas
Porphyromonas gingivalis	0	0	0	29	0	Midas
Salmonella enterica subsp. enterica	0	0	0	0	0	Midas
Staphylococcus aureus subsp. aureus	0	0	0	0	0	Midas
Streptococcus gordonii str.	0	0	0	0	0	Midas
Streptococcus mutans	0	0	0	0	0	Midas
Streptococcus pneumoniae	0	0	0	1	0	Midas
Tannerella forsythia	1	0	1	26	0	Midas
Treponema denticola	0	0	0	42	0	Midas
Treponema pallidum subsp. pallidum	0	0	0	0	0	Midas
Vibrio cholerae	0	0	0	0	0	Midas
Yersinia pestis	0	0	0	0	0	Midas
Yersinia pseudotuberculosis	0	0	0	0	0	Midas

743 **Table S2** Genomes used to generate simulated ancient pathogen DNA data sets

Bacillus anthracis str Ames	**Neisseria meningitidis MC58**
Bordetella pertussis Tohama I	**Porphyromonas gingivalis W83**
Borrelia burgdorferi B31	**Salmonella enterica subsp enterica serovar Enteritidis str P125109**
Brucella abortus 2308	**Salmonella enterica subsp enterica serovar Typhi str CT18**
Brucella melitensis bv 1 str 16M	**Salmonella enterica subsp enterica serovar Typhimurium str LT2**
Clostridium botulinum A str ATCC 3502	**Staphylococcus aureus subsp aureus NCTC 8325**
Clostridium botulinum BKT015925	**Streptococcus gordonii str Challis substr CH1**
Clostridium botulinum E3 str Alaska E43	**Streptococcus mutans UA159**
Clostridium sporogenes NCIMB 10696	**Streptococcus pneumoniae R6**
Clostridium tetani E88	**Tannerella forsythia 92A2**
Corynebacterium diphtheriae NCTC 13129	**Treponema denticola ATCC 35405**
Haemophilus influenzae Rd KW20	**Treponema pallidum subsp pallidum str nichols**
Helicobacter pylori 26695	**Vibrio cholerae M66-2**
Mycobacterium avium subsp paratuberculosis K10	**Vibrio cholerae O1 biovar El Tor str N16961**
Mycobacterium leprae TN	**Yersinia pestis CO92**
Mycobacterium tuberculosis anc	**Yersinia pseudotuberculosis IP31758**
Neisseria gonorrhoeae FA 1090	

744