Briefings in Bioinformatics, 2025, **26(1)**, bbae646
https://doi.org/10.1093/bib/bbae646
Problem Solving Protocol

Filtering out the noise: metagenomic classifiers optimize ancient DNA mapping

Shyamsundar Ravishankar[1,*], Vilma Perez[1,2], Roberta Davidson[1], Xavier Roca-Rada[1,3], Divon Lan[1,4], Yassine Souilmi[1,5,6,‡], Bastien Llamas[1,2,5,6,‡]

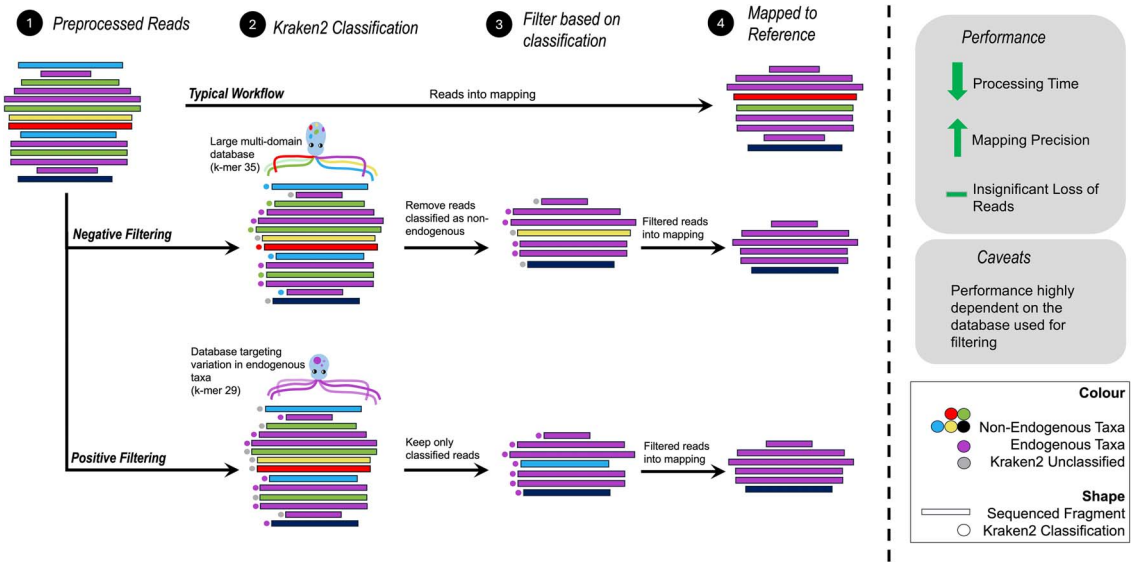
[1]Australian Centre for Ancient DNA (ACAD) and The Environment Institute, The School of Biological Sciences, University of Adelaide, Adelaide, SA, Australia
[2]Centre of Excellence for Australian Biodiversity and Heritage, University of Adelaide, Adelaide, SA, Australia
[3]Faculty of Arts and Humanities, University of Coimbra, Coimbra, Portugal
[4]Genozip Limited, Hong Kong
[5]National Centre for Indigenous Genomics, Australian National University, Canberra, ACT, Australia
[6]Indigenous Genomics, Telethon Kids Institute, Adelaide, SA, Australia

*Corresponding author. Australian Centre for Ancient DNA (ACAD) and The Environment Institute, The School of Biological Sciences, The University of Adelaide, North Terrace, Adelaide, South Australia, 5008, Australia. E-mail: shyamsundar.ravishankar@adelaide.edu.au
‡Yassine Souilmi and Bastien Llamas are equally contributing last authors.

Abstract

Contamination with exogenous DNA presents a significant challenge in ancient DNA (aDNA) studies of single organisms. Failure to address contamination from microbes, reagents, and present-day sources can impact the interpretation of results. Although field and laboratory protocols exist to limit contamination, there is still a need to accurately distinguish between endogenous and exogenous data computationally. Here, we propose a workflow to reduce exogenous contamination based on a metagenomic classifier. Unlike previous methods that relied exclusively on DNA sequencing reads mapping specificity to a single reference genome to remove contaminating reads, our approach uses *Kraken2*-based filtering before mapping to the reference genome. Using both simulated and empirical shotgun aDNA data, we show that this workflow presents a simple and efficient method that can be used in a wide range of computational environments—including personal machines. We propose strategies to build specific databases used to profile sequencing data that take into consideration available computational resources and prior knowledge about the target taxa and likely contaminants. Our workflow significantly reduces the overall computational resources required during the mapping process and reduces the total runtime by up to ~94%. The most significant impacts are observed in low endogenous samples. Importantly, contaminants that would map to the reference are filtered out using our strategy, reducing false positive alignments. We also show that our method results in a negligible loss of endogenous data with no measurable impact on downstream population genetics analyses.

Graphical Abstract



Keywords: ancient DNA; contamination; filtering, metagenomic classifiers; Kraken2

Received: July 18, 2024. **Revised:** November 3, 2024. **Accepted:** November 28, 2024

Introduction

The field of paleogenomics relies on degraded ancient DNA (aDNA) molecules extracted from historic or prehistoric biological remains to study past environments and populations. Major progress in molecular methods to isolate and sequence aDNA has enabled the recovery of high-quality ancient genomes from many different species and sources [1–5]. However, contamination from modern and ancient exogenous sources remains a challenge that requires attention to improve the reliability and interpretative power of paleogenomic research.

Sample exposure to contaminating sources of DNA happens at various stages, including microorganisms and environmental DNA in the soil matrix; DNA from people who collected and handled samples in the field and/or museums and performed laboratory work [6]; and cross-contamination from different samples in the lab during DNA extraction and contamination from DNA sequences in reagents and consumables [7]. In short, a complex mixture of ancient and modern DNA contaminants continuously accumulates in and on the sample from the time of death of the organism up to laboratory work. As a result, endogenous DNA content is often outcompeted by exogenous DNA contaminants in sequenced data [8]. In the last decade, continuing improvements in laboratory protocols and established best practice guidelines have specifically addressed the issue of contamination in aDNA sequence data. From the excavation to library preparation and target enrichment, stringent measures are applied to minimize, identify, or discard exogenous DNA contaminants during sample handling and laboratory work [7].

Beyond these practical advances, sequencing data processing could assist with managing DNA contaminants. Currently, computational methods rely on the specificity of mapping shotgun sequences to a linear reference genome of, or closely related to, the species of interest [9]. However, spurious mapping of exogenous sequences to the target reference increases with decreasing fragment length and if the exogenous sequences come from a species closely related to the target reference. Recently, Feuerborn *et al.* [10] suggested the use of competitive mapping to remove human contamination from faunal aDNA datasets by mapping aDNA sequences against a composite reference sequence file containing both human and target reference genomes simultaneously. This technique is traditionally used in microbial genomics [11]. However, competitive mapping only considers a few sources of contamination and is not easily scalable to target multiple complex eukaryotic organisms due to increasing computational demand with larger composite reference genomes. Therefore, exogenous contamination not only complicates aDNA analysis but also intensifies the computational demands during sequence mapping. Postmapping filtering tools such as PMDtools [12] rely on the presence of aDNA damage misincorporations to remove contemporary contamination. However, not all endogenous aDNA reads present base misincorporations characteristic of aDNA damage, so this approach can lead to a large loss of endogenous sequences. Moreover, contaminating sequences from exogenous aDNA would not be removed by this approach. In short, efficient computational methods are lacking to not only identify but also efficiently remove contaminant DNA during the mapping process of aDNA datasets.

In recent years, the field of metagenomics has given rise to metagenomic classifiers capable of efficiently and accurately identifying diverse taxa in sequence data. These capabilities have shown potential in ancient metagenomic studies as well, although with specific caveats for each tool [13, 14]. We hypothesize that metagenomic classifiers offer an efficient approach to removing contaminant DNA by filtering them out before the mapping stage. As a result, we predict improved mapping accuracy and a significant reduction in computational resources needed, thereby making aDNA analysis more accurate and accessible across various computing platforms. A similar approach has been successfully applied to remove human patient DNA from clinical metagenomic data [15].

Kraken2 [16] is a *k*-mer-based classifier initially designed to perform metagenomic analyses. Here, we propose an approach where *Kraken2* is used to identify and remove contaminating sequences from ancient DNA datasets of single organisms to accelerate the mapping process and improve mapping accuracy. We opted for a *k*-mer-based metagenomic classifier over alignment-based methods due to its faster processing speed [17], and we chose *Kraken2* because it presented the best balance between speed, database size, and classification accuracy compared to other metagenomic classifiers [18, 19], especially in ancient DNA contexts [20]. Using both simulated (human and dog) and empirical shotgun aDNA datasets, we show that this workflow presents a simple and efficient method that enables the removal of contaminating sequences from aDNA datasets with limited loss of endogenous DNA sequences while simultaneously reducing the overall computational resources needed during the mapping process as well as mitigating any potential errors introduced by spuriously mapping contaminant reads.

Methods

Data simulations

We simulated ancient human (*Homo sapiens*) and dog (*Canis lupus familiaris*) shotgun sequencing datasets with varying levels of contamination (Fig. 1, Table S1) using *Gargammel* 1.1.2 [21]. The human and dog reads, referred to as endogenous, were simulated from autosomal, sex, and mitochondrial contigs of *GRCh38.p14* (GCF_000001405.40) and *CanFam6* (GCF_000002285.5) genome assemblies, respectively. The exogenous contaminants consisted of modern human reads, represented by reads simulated from the same *GRCh38.p14* genome assembly; microbial contamination, represented by reads simulated from profiled microbial communities (Table S2) as presented by Seguin-Orlando *et al.* [22], including bacteria, viruses, and phages; and other contaminating reads representing common sources of contaminants found in aDNA datasets [23, 24]: sheep (*ARS-UI_Ramb_v2.0*; GCF_016772045.1), domestic cattle (*ARS-UCD1.3*; GCF_002263795.2), pig (*Sscrofa11.1*; GCF_000003025.6), goat (*ARS1.2*; GCF_001704415.2), and chicken (*GRCg6a*; GCF_000002315.6). Deamination profiles were simulated from the Loschbour individual in Lazaridis *et al.* [25] for the endogenous and microbial reads, simulating the damage profile of a single-stranded aDNA library partially treated with uracil–DNA–glycosylase (UDG) (Supplementary Fig. S1a). Finally, all the reads were simulated as paired-end Illumina HiSeqX reads with a read length of 75 bp and size distribution of the sequenced fragments simulated from the subset of a 45 000-year-old human sample from Siberia [26] (Supplementary Fig. S1b). Default *Gargammel* settings were used for the rest, including base quality distribution of the simulated bases and adding Illumina adapter sequences for fragments shorter than the read length (75 bp).

We simulated two scenarios with differing levels of modern human and other contamination in the dataset (Fig. 1A). For the ancient human dataset, the modern human portion was replaced with microbial sequences since a taxonomic classifier such as

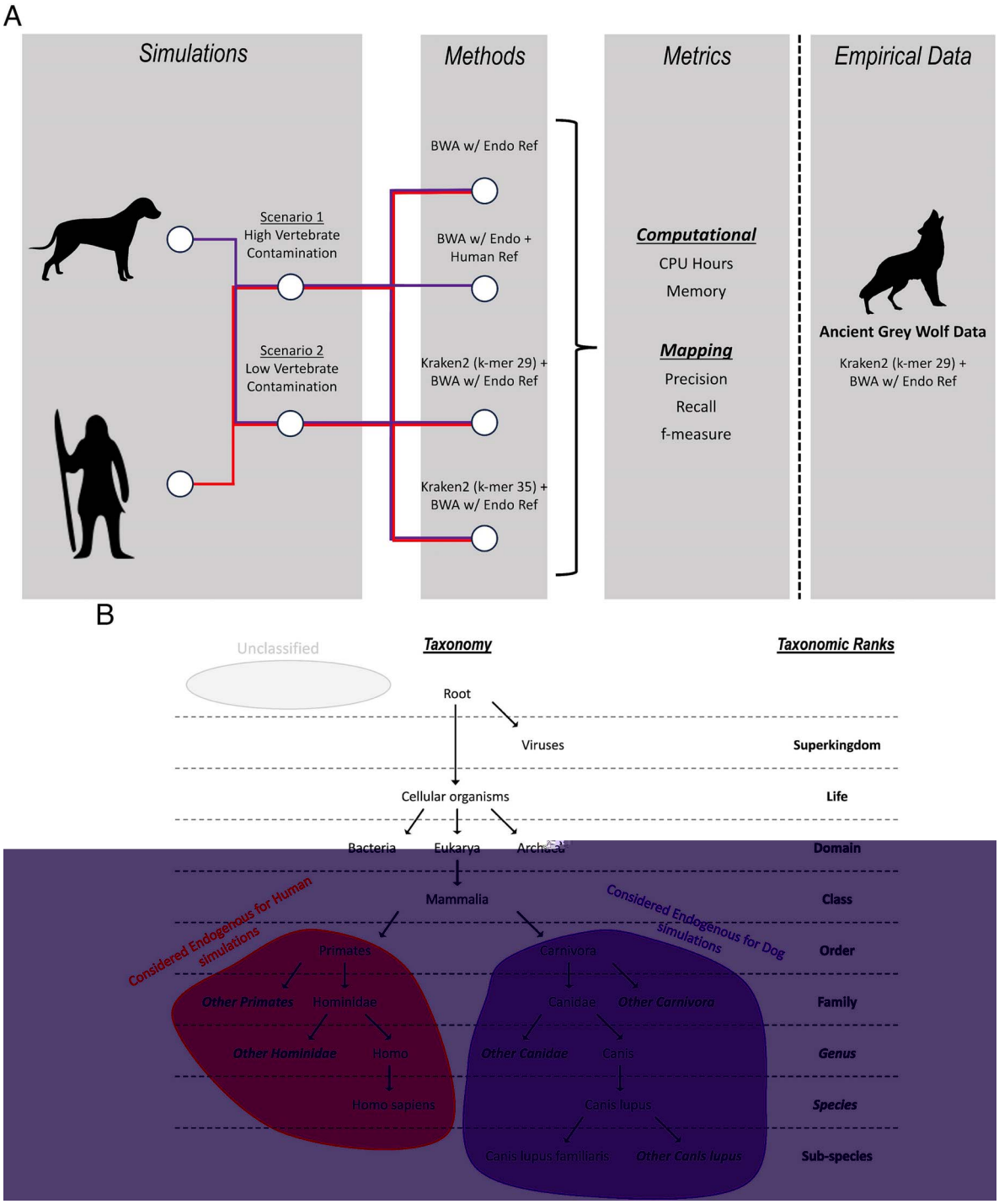


Figure 1. (A) Workflow of types of simulated data, different methods applied to the simulated data, and metrics collected. The best-performing method is applied to empirical data. (B) Ancient human reads classified at the order Primates or lower taxonomic ranks are considered endogenous reads and hence retained (red), similarly for ancient dog reads at the order Carnivora or lower (purple). Unclassified reads represented as grey are reads that could not be assigned a taxonomy.

Kraken2 will not be able to differentiate between modern and ancient reads from the same taxa. The endogenous proportion of the simulated reads ranged from 0.1% to 60%, with 20 million read pairs simulated for each (Table S1).

Data processing

Preprocessing of simulated data

The simulated datasets were processed with *AdapterRemoval* 2.3.2 [27] to trim adapters and merge paired-end reads with length and base quality filters '*–minlength* 30' and '*–minquality* 20', respectively [28]. Additionally, '*–qualitymax*' was set to 64 to account for maximum read quality in the datasets simulated using *Gargammel*.

Data processing for baseline mapping performance

We mapped the merged reads using *bwa aln 0.7.17-r1188* [29] with the '*-n* 0.01', '*-l* 1024' and '*-o* 2' ancient DNA parameters [30]. The human and dog datasets were mapped to *GRCh38.p14* and *CanFam6* genome assemblies, respectively. The mapped files

were used to establish the baseline performance of *bwa aln* for all simulated endogenous levels.

Measuring baseline mapping performance

To measure the mapping performance, the reads were then categorized as 'true positives' (**TPs**), 'true negatives' (**TNs**), 'false positives' (**FPs**), and 'false negatives' (**FNs**) depending on the source taxonomy (endogenous or contamination) of the mapped read and its mapping quality to the reference genome (Table S3). These were used to calculate *precision, recall, and f-measure* to establish the baseline mapping performance [31].

Precision $= \frac{TP}{TP+FP}$ measures the proportion of correctly predicted positive instances out of all instances predicted as positive. In this context, precision quantifies the ratio of endogenous reads to all reads mapped to the reference.

Recall $= \frac{TP}{TP+FN}$, also known as sensitivity or the true positive rate, measures the proportion of correctly identified positive instances from all positive instances in the dataset. In this context, recall quantifies the ratio of endogenous reads mapped to the reference to all endogenous reads in the dataset.

f-measure (or *f*-score): $F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ is the harmonic mean of *precision* and *recall*. It is typically used to quantify the overall performance of a classification model since optimizing only for precision or recall can have conflicting goals. A high F_1 value suggests accurate identification of positive instances while also minimizing false positives. In this context, F_1 quantifies a method's ability to identify endogenous reads and contaminants in the dataset accurately.

Measuring the performance of competitive mapping

For the dog dataset, we also benchmarked competitive mapping using a composite reference with the *CanFam6* and *GRCh38.p14* assemblies as suggested by Feuerborn *et al.* [10], which has been shown to remove contemporary human contamination in ancient faunal datasets. We then mapped the reads using the composite reference and categorized reads according to Supplementary Table S4 to calculate *precision*, *recall*, and *f-measure*.

Metagenomic classification

Measuring the performance of metagenomic filtering before mapping

We used *Kraken2* (v 2.1.3) [16], a *k*-mer-based method for taxonomic classification of sequence data. A *k*-mer refers to substrings of length *k* within a nucleotide sequence. *Kraken2* relies on the presence of exact *l*-mer (a subsequence of length *l*, where $l \leq k$) matches between sequence data and a reference database containing known sequences and taxonomies to perform taxonomic classification.

To understand the effect of database composition on taxonomic classification, we created different databases that contained single species or sequences from multiple domains of life (Table 1). All databases were built with the default *k*-mer length of 35, as well as the *k*-mer length of 29. This choice of *k*-mer lengths is motivated by the fact that while a longer *k*-mer (i.e. 35) decreases the risk of false classification compared to a shorter *k*-mer (i.e. 29) [18], aDNA datasets generally use a cut-off of 30 bp for the minimal fragment length for mapping to a reference genome [32]. Therefore, a *k*-mer of 29 is likely more appropriate in the context of aDNA datasets to prevent a bias against very short reads. Finally, we also tested a publicly available database, *k2_nt_20230502* (https://benlangmead.github.io/aws-indexes/k2), which includes a larger collection of sequences across the three domains of life and viruses from the National Centre for Biotechnology Information (NCBI), inclusive of GenBank, RefSeq, Third Party Annotation (TPA), and Protein Data Bank (PDB).

We evaluated databases based on their size and sensitivity when classifying endogenous sequences. The best-performing databases (see results) were used to filter reads from the simulated dataset before mapping. The filtered reads were then mapped to the reference with the same method as the baseline above. Finally, *precision*, *recall*, and F_1 were calculated after reads were categorized based on metagenomic filtering and mapping as per Tables S5 and S6.

Building Kraken2 databases

The *k*-mer 35 databases were built with the default options using genomes described in Table 1. The *k*-mer 29 database was built with options '*—kmer-len 29*', '*—minimizer-len 24*', and '*—minimizer-spaces 6*' and included genomes as described in Table 1.

Kraken2 filtering

The *Kraken2* classifications were run with default parameters. A nextflow pipeline is available on github (https://github.com/shyama-mama/taxonomicfiltering) to perform filtering given a database and input reads.

Empirical data

To validate our method with empirical data, we selected 10 ancient *Canis* samples from Bergström *et al.* [33] at similar endogenous proportions to the simulated data (Fig. 1A, Table S6). We mapped the data with *bwa aln* to the *CanFam6* reference genome following the parameters in the Data Processing for Baseline Mapping Performance section. We built a *Kraken2* database with a *k*-mer length set to 29 and composed of reference sequences from dog (*CanFam6*; GCF_000002285.5), grey wolf (*mCanLor1.2*; GCA_905319855.2), and dingo (*ASM325472v2*; GCF_003254725.2). Additionally, we also added a consensus *CanFam3.1* (GCF_000002285.3) reference using the alternate allele from bi-allelic single-nucleotide polymorphisms (SNPs) from 722 Canidae genomes from [34] into the database to minimize reference bias [35]. We used databases with and without the alternate allele information for filtering. We filtered the data by discarding any unclassified reads before mapping with *bwa aln* as above and compared both approaches: mapping only ($\text{Sample}_{\text{BWA}}$) and filtering before mapping ($\text{Sample}_{\text{Filt}}$).

To understand the effect of filtering on the data, for each sample, we extracted reads that met the following criteria: they mapped to the reference with a mapping quality score >20 when no premapping metagenomic filtering was performed, and they were also removed by *Kraken2* filtering when metagenomic filtering was performed. These reads were then classified using Nucleotide BLAST v2.14.1 [36]. We used the MEGAN v6.24.20 [37] suite of tools to get taxonomic abundances for each sample using the weighted lowest common ancestor (LCA) algorithm and following best-suited options for ancient DNA as suggested by Eisenhofer and Weyrich [13].

Since *Kraken2's* taxonomic assignment relies on exact *l*-mer matches (where $l \leq k$) to sequences in the database, there is a possibility to introduce reference bias when filtering relies on identifying endogenous sequences in the dataset. Hence, we used f_4 statistics from ADMIXTOOLS v7.0.2 [38] to assess if endogenous reads that were removed during *Kraken2* filtering introduced any bias in downstream analysis. The f_4 was performed with the configuration $f_4(\text{Sample}_{\text{Filt}}, \text{Sample}_{\text{BWA}}$; Basenji01, Coyote01California), where Coyote01California and

Table 1. List of Databases used, their contents, k-mer length and memory required to run the database.

Name	Contents	K-mer	Memory (GB)
k2_human	Default Kraken Human Library	29	2.9
		35	3.9
k2_microbes_human	*k2_human + Default Kraken Libraries for archaea, viruses, bacteria, fungi, and protozoa*	29	35
		35	73
k2_dog	Sequences from CanFam6 (Dog)	29	2.6
		35	3.3
k2_canis_lupus	Sequences from CanFam6 (Dog), mCanLor1.2 (Grey Wolf) and ASM325472v2 (Dingo)	29	2.8
		35	3.7
k2_canis_lupus + 722 g Variants	*k2_canis_lupus + Consensus sequence of CanFam3.1 (Dog) genome with alternate alleles from 722 g project*	29	3.7
k2_human_dog	*k2_dog + Default Kraken Human Library*	29	5.1
		35	7.2
k2_microbes_human_dog	*k2_microbes_human + k2_dog*	29	35
		35	76
k2_custom	*k2_microbes_human_dog + Default Kraken Library for Plants + Sequences from ARS-UI_Ramb_v2.0 (sheep), ARS-UCD1.3 (cow), Sscrofa11.1 (pig), ARS1.2 (goat), GRCg6a (chicken), Bison_UMD1.0 (American Bison), panTro6 (chimpanzee), Kamilah_GGO_v0 (gorilla), ponAbe3 (orangutan), EquCab3.0 (horse), UM_NZW_1.0 (rabbit)*	29	58
		35	166
k2_nt_20230502	Very large collection, inclusive of GenBank, RefSeq, TPA, and PDB from Kraken2's publicly available indexes	35	481

Basenji01, are coyote and dog, respectively, from Plassais *et al.* [34]. The effect of reference bias was tested for databases with and without alternate allele information. Pseudo-haploid genotype calls were generated for each sample using PileupCaller 1.5.2 (https://github.com/stschiff/sequenceTools) at polymorphic loci ascertained using heterozygous sites in the coyote genome [33] mapped to the *CanFam6* reference. To account for variability during random allele sampling, we replicated pseudo-haploid genotype calls three times per sample. We tested each sample for reference bias caused by filtering where a significant deviation from 0 in the positive direction indicates an introduced bias towards the reference due to filtering.

Results

Effect of database composition on taxonomic classifications

Based on simulated ancient human and dog reads with no introduced contamination, we demonstrate that database composition significantly impacts classification accuracy (Fig. 2). Databases consisting of only a single genome exhibit a bias towards classifying sequences as belonging to that genome's taxonomy. Conversely, more complex databases may compromise taxonomic resolution—i.e. the ability to classify sequences at the lowest specific taxonomic rank, as previously observed by Nasko *et al.* [39]. We observed a substantial reduction in dog reads misclassified as human from 3.17% to 0.12% and 52.52% to 3.34% for *k*-mer lengths 35 and 29, respectively, when using '*k2_human_dog*', built from human and dog reference genomes, compared to '*k2_human*', built with only the human reference (Fig. 2A). Using the combined database led to some reads classified as Boreoeutheria (0.26% and 1.77% for *k*-mer lengths 35 and 29, respectively), a classification rank that includes humans and dogs. Similarly, *Canis* sequences in '*k2_nt_20230502*' led to dog reads classified primarily as *Canis* or *C. lupus* (Fig. 2A), and the presence of primate genomes in the '*k2_custom*' database led to the classification of human reads as Hominidae or Homininae (Fig. 2B). We also note a substantial number of human reads classified as 'root' (22.22%) when using '*k2_nt_20230502*'. A 'root' classification is a result of the query sequence matching with viral and cellular organisms in the database (Fig. 1B). It is unclear if this is a result of contaminated viral sequences in the database or the presence of human endogenous retroviral sequences.

Databases built with *k*-mer length 29 decrease the size of the database and the number of unclassified reads—i.e. reads that could not be assigned a taxonomy (Fig. 1B). *Kraken2* assigns a classification to a read-only when the read contains unique *l*-mers (a subsequence of length *l*, where $l \leq k$) that exactly match the *l*-mer sequences in the database [16]. Hence, databases built with a *k*-mer length of 29 assign taxonomies to shorter reads and reads with misincorporations due to damage better than the larger default *k*-mer length (Supplementary Fig. S2). However, this comes at the price of lower taxonomic resolution, with increased ancient dog and human reads classified as 'cellular organism' or 'Eukaryota' (21.38% and 22.31% for dog and human reads, respectively, using '*k2_custom*'), as well as increased false classifications as the simulated reads were misclassified as other vertebrate species in the database (>10% in dog reads using '*k2_custom*').

Benchmarking filtering before mapping

Filtering before mapping using metagenomic classifiers primarily aims to retain as many endogenous reads as possible while discarding contaminating sequences. However, sensible taxonomic classification is highly dependent on taxa represented in the database, regardless of database complexity. As we showed above, the rate of classification bias is driven by the *k*-mer length used to build the database, with lower specificity due to a shorter *k*-mer length leading to more aberrant classifications. Therefore, we investigated two strategies that take into account the strengths and drawbacks of different types of databases.

First, we applied filtering based on identifying contaminants using databases built with *k*-mer 35 and consisting of sequences from the target taxa and as many, as possible, contaminating genomes, to remove as many contaminating sequences as possible (negative filtering). We predict that this strategy is well suited for when the database is built with larger *k*-mer sizes, which show substantially lower false classification rates (Fig. 2).

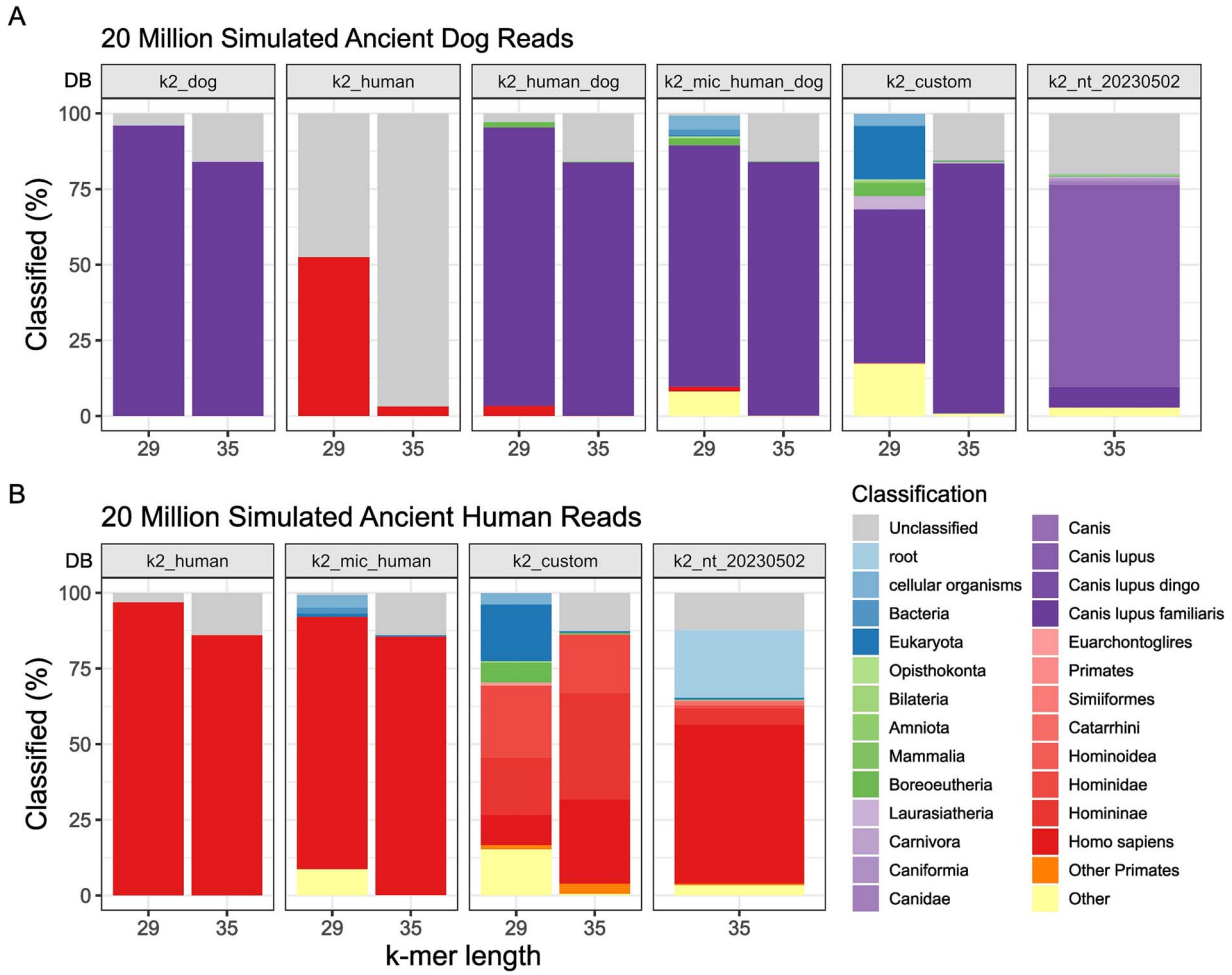


Figure 2. Impact of database choice on Kraken2 classification of (A) 20 million ancient dog reads, and (B) 20 million ancient human reads. The x-axis represents the *k*-mer length of the databases (DB; see Table 1 for descriptions) represented in the facets. The y-axis shows the proportion of reads classified as a particular taxonomy (colours).

Second, we applied filtering based on identifying endogenous reads using databases built with *k*-mer 29 and consisting of sequences from target taxa and genomes related to the target taxa to the family level, to retain as many endogenous reads as possible (positive filtering). We predict that this strategy is well suited for low *k*-mer sizes, which show higher classification rates at shorter read lengths (Supplementary Fig. S2).

The ancient human and dog reads were simulated as per the two scenarios in Fig. 1A and Table S1. Indeed, *Kraken2* is a metagenomic classifier, and it will not be able to differentiate between modern and ancient reads from the same species. Furthermore, while modern human contamination is a persistent problem in ancient human datasets, tools such as PMDtools have been shown to effectively remove them [12]. Subsequently, for the negative filtering strategy (removing contaminants), the simulated datasets were classified using the '*k2_custom*' database, built with *k*-mer 35, to select reads classified at the order rank of the species of interest—Primates for human and Carnivora for dog—or remaining unclassified, before performing mapping with *bwa aln*. For the positive filtering strategy (retaining endogenous data), we used '*k2_canis_lupus*' and '*k2_human*' built with *k*-mer 29 for the dog and human datasets, respectively, and only mapped classified reads.

A positive outcome of filtering reads before mapping is shorter processing times. We see more than a 6-fold increase in processing speed (Supplementary Figs S3A and S4A, Tables S8 and S9). Here, processing time includes the runtime to classify, filter, and map reads when filtering is applied, as opposed to only mapping time when no filtering is applied. Interestingly, despite identical total read counts across all endogenous fractions, mapping took longer for data with high contamination from vertebrate sequences (i.e. Scenario 1 in Fig. 1 and Table S1) as opposed to data with low vertebrate contamination (i.e. Scenario 2 in Fig. 1 and Table S1), when mapped solely using *bwa aln*. This suggests that contaminants from taxa closely related to the target species' reference disproportionately and negatively impact mapping time. Using a larger composite reference in competitive mapping further extended mapping times, up to 1.6-fold (Fig. S3A, Table S8). Combining competitive mapping with both *Kraken2* filtering strategies greatly improved mapping time, up to 4.4-fold and 6.8-fold faster when compared to single-reference and competitive mapping, respectively (Fig. S3A, Table S8).

By default, *Kraken2* databases are loaded into working memory and hence require, at minimum, free memory the same size as the database used. The '*k2_canis_lupus*' (*k*-mer 29), and '*k2_human*' (*k*-mer 29) databases are relatively small at 2.6 and 2.9 GB, respectively, making it possible to filter data even on personal machines. In contrast, the '*k2_custom*' (*k*-mer 35) database is substantially larger, at 166 GB, and is better suited for high-performance computing systems. Importantly, premapping filtering, regardless of

the filtering strategy, greatly reduces the read volume for highly contaminated samples (Tables S8 and S9), which, in turn, facilitates a more efficient mapping process, sparing computational resources.

Across all endogenous proportions tested, both *Kraken2* filtering strategies consistently yielded more precise mapping compared to *bwa aln* mapping alone for both the high and low vertebrate contaminations (Fig. 3A and Supplementary Fig. S5A; Tables S8 and S9). Competitive mapping was also more precise compared to *bwa aln* mapping to a single reference, indicating it can accurately remove human contaminants from the faunal data, albeit with the longest run time. Since competitive mapping was only used to remove human contamination, other microbial and vertebrate contamination introduced remained in the mappings (Figs S10–S13). Combining competitive mapping with *Kraken2* filtering improved mapping accuracy beyond what either method achieved alone while significantly reducing the processing time required for competitive mapping (Figs 3A and S5A).

We summarized the *precision* and *recall* of each method by read length (Figs S6–S9). Both competitive mapping and positive filtering showed the largest precision increase for shorter fragments (30–40 bp) compared to longer fragments when benchmarked against *bwa aln* mapping to a single reference. Interestingly, negative filtering achieved the highest precision improvement with fragments 41–50 bp in length. This is due to the *k*-mer 35 database used for negative filtering; shorter fragments <35 bp cannot be classified with this *k*-mer length. Additionally, since negative filtering retains unclassified reads for mapping with *bwa aln* to maximize endogenous read retention, the precision increase is less pronounced for 30–40 bp fragments than for 41–50 bp fragments.

We observe a greater loss of endogenous reads when applying both Kraken2 filtering strategies compared to mapping directly to the reference, with losses up to 0.99%, and 3.8% for negative and positive filtering, respectively (Tables S8 and S9). For negative filtering, this lower recall stems from endogenous reads misclassified as other taxa, a known limitation of *Kraken2*'s probabilistic compact hash table, which can lead to false classifications. This loss is more pronounced with fragments 41–50 bp in length. In positive filtering, short fragments (<50 bp) were filtered out, as the *k*-mer 29 database struggled to classify these reads, despite the shorter *k*-mer length (Figs S6–S9). Despite this loss of endogenous reads, the resulting precision increase from filtering improved f-measures for low endogenous samples with high vertebrate contamination (Scenario 1). In contrast, this same loss of endogenous reads lowered f-measure for samples with more endogenous data or less vertebrate contamination (Scenario 2) when compared to *bwa aln* mapping to a single reference (Figs 3B and S5B). Competitive mapping, in comparison, showed the lowest loss of endogenous reads across all fragment lengths and endogenous fractions. Consequently, combining *Kraken2* filtering with competitive mapping improved the *f*-measure beyond what Kraken2 filtering achieved individually.

Filtering empirical data

Finally, we mapped classified reads from 10 ancient grey wolf samples from Bergström *et al.* [33] using the positive filtering strategy with the '*k2_canis_lupus*' database built with *k*-mer 29 and including biallelic SNPs from 722 Canidae genomes [34]. We see up to a 16-fold increase in processing speed (sample 367 with 0.4% endogenous DNA) when filtering with *Kraken2* (Fig. 4A, Supplementary Table S10). The performance boost was observed across samples, with even those having a higher endogenous fraction of 20% and 50% or more experiencing a 2-fold and 1.3-fold increase in processing speeds, respectively. In line with our simulated results, we observed some loss of endogenous data (Fig. 4B, Supplementary Table S10).

Reads that were mapped to the reference with a mapping quality score above 20 when no premapping metagenomic filtering was performed but were filtered out during metagenomic filtering were assigned taxonomies using BLAST and LCA algorithms in MEGAN. Most of these reads could not be assigned a taxonomy (41.80%–76.55%), with the majority of the classified reads assigned to the Order Carnivora (> 93% of classified reads). Of note, all samples had reads classified at the order Primates (0.03%–0.70%). Competitive mapping against *CanFam6* and *GRCh38.p14* references saw the reads map preferentially to the human reference. Since these reads were originally also mapped to the *CanFam6* reference, we were able to compare the damage profiles of these reads mapping to the two references independently. Six out of the 10 samples (367, IN18-005, TU114, TU148, WOL-VAL-18A, and CANIS-ALAS-016) showed spurious C-to-T and G-to-A misincorporations when mapped to the dog reference, whereas no such misincorporations were observed when mapped to the human reference, suggesting these reads are modern human contamination (Supplementary Fig. S14). We also observed up to 3.58%, 0.82%, 0.27%, 0.26%, and 0.25% reads classified as bacteria, Artiodactyla, Rodentia, Chiroptera, and Lepidoptera, respectively (Supplementary Table S11). The authenticity of these classifications was not determined.

We see no significant bias ($|Z| < 3$) introduced by filtering with *Kraken2* when the database used to select endogenous reads contains variation information for Canids (Fig. 4C). However, it should be noted that samples CANIS-ALAS-016 and AL2744 observed low but significant reference bias when the *Kraken2* database used to identify endogenous reads did not contain the alternate allele information (Fig. 4C, Supplementary Tables S12 and S13).

Discussion

Our study presents evidence that premapping filtering using *Kraken2* not only optimizes the usage of computational resources by greatly reducing mapping time but also improves the precision of mapped aDNA reads. This is particularly evident in datasets involving samples with very low levels of endogenous sequences and high contamination from sequences closely related to the target species. By implementing a positive filtering strategy to retain putative endogenous reads using a *Kraken2* database built with *k*-mer 29 consisting of reference and alternate sequences, and genomes from closely related taxa to the species of interest, we achieve a streamlined process that is resource-efficient and suitable for a wide array of computational environments, including personal machines due to the database requiring <5 GB of memory to run.

A more thorough approach is negative filtering to remove putative contaminants. This strategy utilizes a comprehensive database built with *k*-mer 35 and encompassing a broad spectrum of contaminants and affords enhanced precision and recall. The increased memory requirements for a larger database means this approach is more suitable for researchers who have access to high-performance computing resources. However, the feasibility of this strategy depends on how well the database represents the full array of potential contaminants in a particular dataset

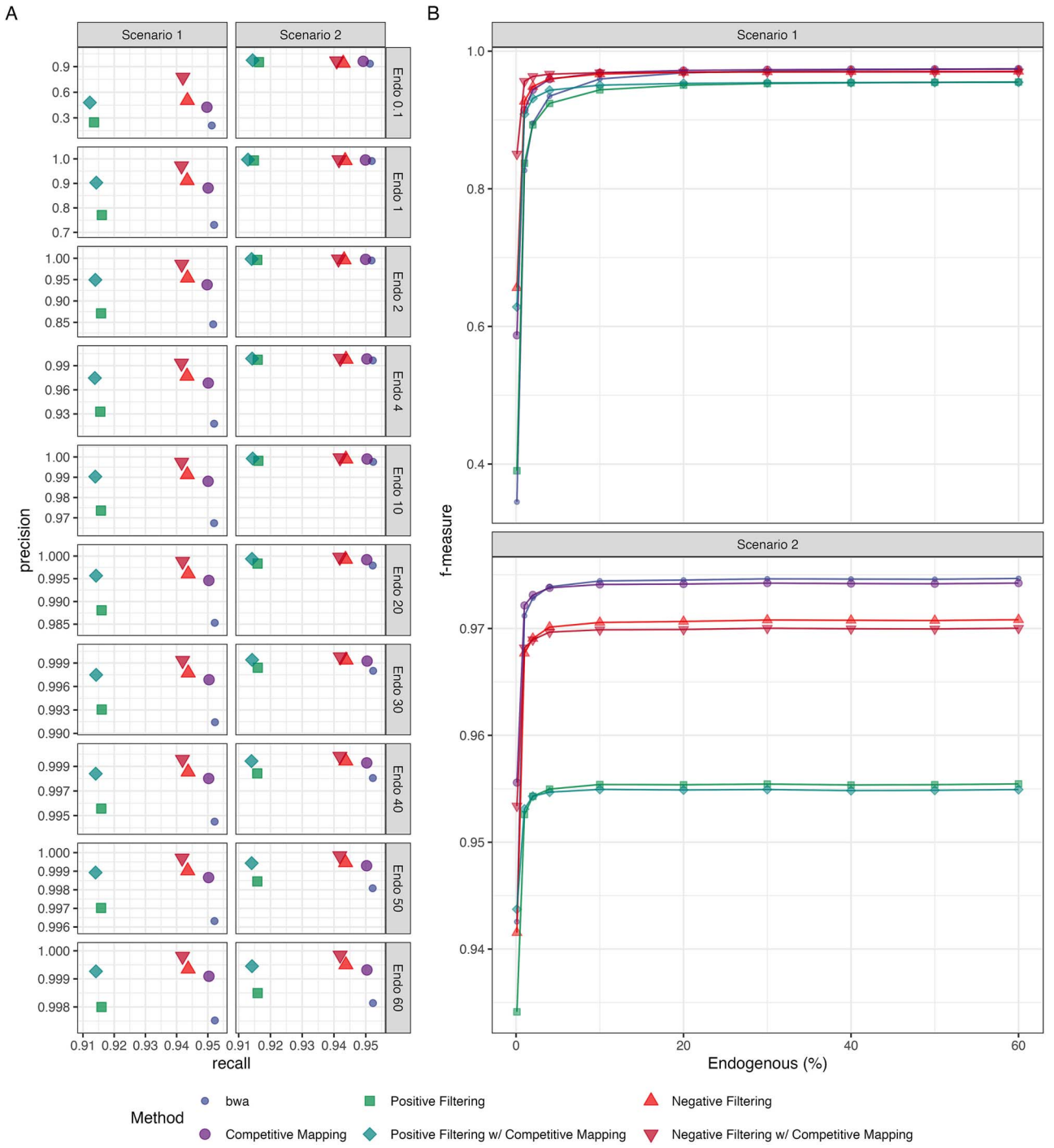


Figure 3. Precision and recall (A) and f-measure (B) of the six methods—bwa mapping to a single (colour: blue & shape: small circle) and composite dog and human reference (competitive mapping; colour: purple & shape: large circle), mapping only reads classified as Carnivora and unclassified reads by the 'k2_custom' database to a single (negative filtering; colour: light red & shape: triangle) and composite reference (negative filtering w/ competitive mapping; colour: dark red & shape: upside-down triangle), mapping only reads classified by 'k2_canis_lupus_kmer29' database to a single (positive filtering; colour: light green & shape: square) and composite reference (positive filtering w/ competitive mapping; colour: dark green & shape: diamond)—for the simulated ancient dog genome. Reads were filtered with MapQ >20 postmapping.

because the filtering depends on how many contaminants can be classified and hence removed. Identifying environmental microbes is a limitation of empirical data since most reference databases focus on human pathogens or microbial species that are of interest to humans [40–46]. However, in recent years, ancient and modern environmental microbiomes have increasingly been characterized [47, 48]. The development of resources like the Genomic Taxonomy Database [49] is helping bridge the gap between genomics and microbial taxonomy, improving microbial characterization across diverse environments, including contamination present in ancient datasets.

Our findings suggest that the choice of strategy should be guided by the available resources and specific priorities. The loss

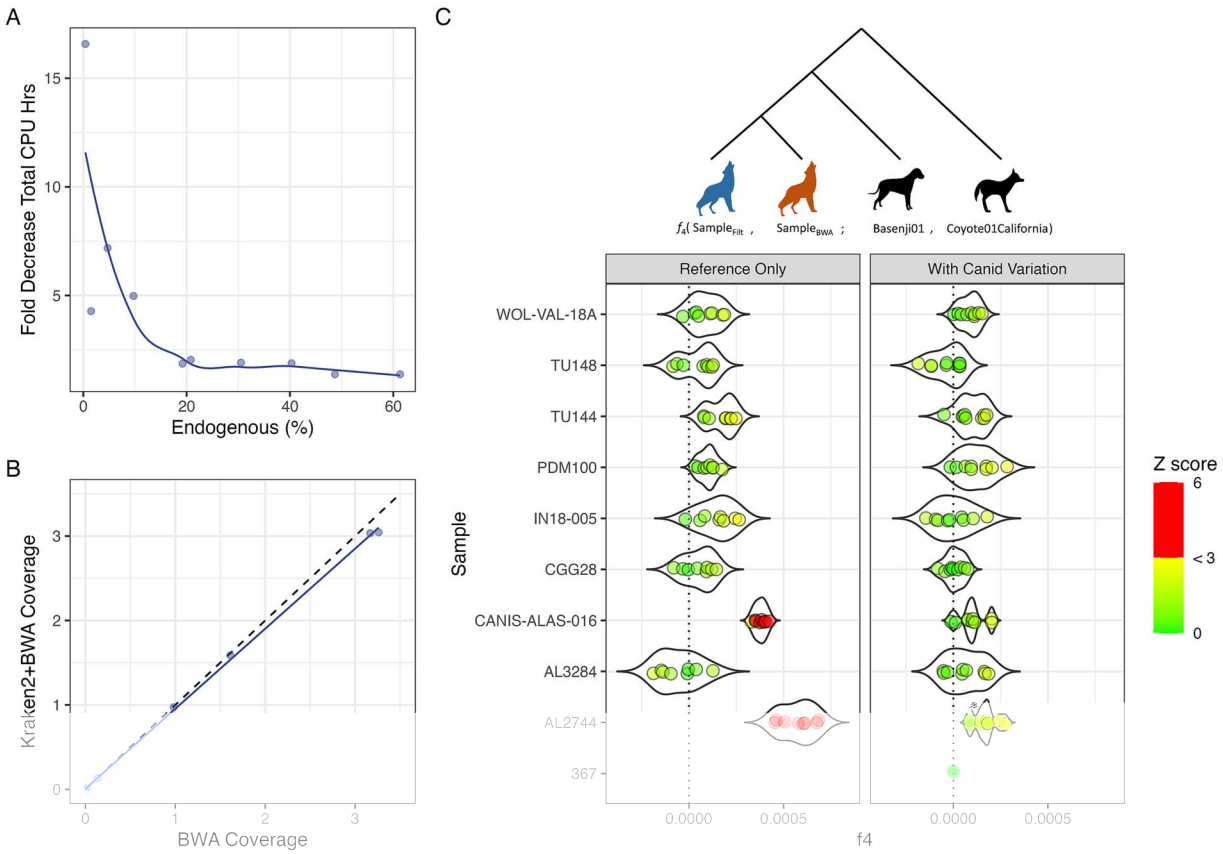


Figure 4. (A) Kraken2 + bwa aln filtering and mapping speed, normalized to bwa aln mapping only speed; (B) coverage difference between mapping only and filtering before mapping; and (C) observed f4 statistics of the configuration f4(SampleFilt (blue), SampleBWA (brown); Basenji01, Coyote01California) from pseudo-haploid genotypes. Multiple points indicate replicate pseudo-haploid calls to account for variability introduced by random pseudo-haploidization. The points are coloured by |Z| score. |Z| values below 3 are on a green-to-yellow gradient. |Z| values above 3 are denoted with red. Filtering using only the reference genome (left panel) led to samples CANIS-ALAS-016 and AL2744 being significantly biased (Z > 3) towards the reference. Adding Canid variation from the 722 g project (right panel) shows a nonsignificant deviation (|Z| < 3) from 0 for all samples.

of some negligible amounts of endogenous data is an inherent limitation of the *Kraken2* filtering approach, but this trade-off is balanced by gains in mapping efficiency and precision. Endogenous read loss occurs either through misclassifications—more common in negative filtering—or through unclassified reads in the positive filtering strategy. Misclassifications stem from the probabilistic compact hash table used by *Kraken2*, which, though memory-efficient compared to a standard hash table, sacrifices some specificity and accuracy [16]. Likewise, unclassified reads in positive filtering are also because of this, with shorter, damaged DNA fragments being especially affected (Fig. S2).

We also caution against filtering by identifying endogenous reads when studying extinct species for which no reference genomic resources exist, as it might impact the retrieval of endogenous reads and cause the reads to be biased towards the reference alleles in the database. For well-studied species such as humans and dogs, we propose adding alternate allele information from large genomic studies to better capture variation in the sequence data. Furthermore, the processing time of ancient hominin genomes enriched with the 1240 k SNP panel [50] could be greatly reduced if a database with the human reference and the expected alternate allele information captured by the panel were used to select human reads before mapping. Our findings coincide with the increasing use of pangenomic approaches for genomic analysis, such as adding alternate allele information to reduce reference bias during mapping [51], building databases from sequences from pangenomic projects for improved host removal from clinical metagenomic data [15], and novel tools such as Euka [52] that use pangenomic graphs for metagenomic classification—although currently, Euka databases are restricted to mitochondrial genomes of tetrapods and arthropods.

We also highlight that using databases with microbial and human sequences to classify human DNA can lead to some human reads being classified at the kingdom or domain rank as previously observed [18] because some microbial sequences are contaminated with human DNA. It is thus imperative to benchmark databases against the project's objectives to mitigate these issues and their effect on data interpretation, which aligns with the growing body of literature that underscores the importance of benchmarking metagenomic classifiers in different contexts [13, 14, 18–20, 53, 54]. In recent years, there have been increased efforts in characterizing and removing contaminated sequences in reference databases to reduce erroneous interpretations of metagenomic datasets [55–57].

We propose that an approach that includes classification-based filtering has the potential to refine data processing and improve overall mapping data quality. We anticipate that continued improvements in metagenomic classifiers and reference databases that can identify environmental taxa will result in increased accuracy of our proposed filtering approaches and reduce data loss, paving the way for more precise reconstructions of ancient genomes.

Key Points

- Contamination is a major challenge in paleogenomics. Computational methods are essential to distinguish between endogenous and contaminant sequences.
- We propose a new workflow relying on a metagenomic classifier to filter out contaminants prior to aligning sequences to a reference sequence.
- We provide clear strategies to build the reference database and finetune the parameters to optimize the classification.
- Our workflow significantly reduces the computational resources and overall runtime while improving mapping precision and downstream analyses.

Acknowledgements

We would like to acknowledge that the University of Adelaide and the Australian Center for Ancient DNA are located on the unceded lands of the Kaurna people, we extend our gratitude and acknowledgement to elders past and present. We sincerely thank our colleagues at the Australian Centre for Ancient DNA for their valuable feedback on the manuscript. This work was supported with supercomputing resources provided by the High Performance Computing (HPC) service at the University of Adelaide.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Author contributions

Conceptualization/Design—D.L., Y.S., B.L.; Data analysis—S.R.; Interpretation—S.R., V.P., Y.S., B.L., R.D., X.R.R.; Manuscript writing—S.R., V.P., Y.S., B.L.; Manuscript editing—S.R., V.P., Y.S., B.L., R.D., D.L., X.R.R.

Funding

B.L. was supported by funding from the Australian Research Council Centre of Excellence for Australian Biodiversity and Heritage (ARC CE170100015), Y.S. by NHMRC SYNERGY grant (GA204260) and S.R. by an Australian Government Research Training Program Scholarship, X.R.R. acknowledges the FCT–Foundation for Science and Technology, I.P./MCTES (PTDC/HAR-ARQ/6273/2020) for funding the development of his postdoctoral fellowship through the Portuguese National Funds (PIDDAC).

Data availability

All reference genomes used and their associated accession numbers can be found in Supplementary Table S7. The publicly available *Kraken2* nucleotide database, *k2_nt_20230502*, was downloaded from https://benlangmead.github.io/aws-indexes/k2 and the 2023/05/02 version used for the manuscript. The VCF file for the National Human Genome Research Institute (NHGRI) Dog Genome Project was downloaded from the National Institutes of Health (NIH) (https://research.nhgri.nih.gov/dog_genome/downloads/datasets/WGS/).

Code availability

The filtering workflow is available at https://github.com/shyama-mama/taxonomicfiltering. Other code for the manuscript can be found in https://github.com/shyama-mama/taxonomicfilteringmanuscript.

References

1. Pinhasi R, Fernandes DM, Sirak K. *et al.* Isolating the human cochlea to generate bone powder for ancient DNA analysis. *Nat Protoc* 2019;**14**:1194–205. https://doi.org.10.1038/s41596-019-0137-7.
2. Shirazi S, Broomandkhoshbacht N, Oppenheimer J. *et al.* Ancient DNA-based sex determination of bison hide moccasins indicates promontory cave occupants selected female hides for footwear. *J Archaeol Sci* 2022;**137**:105533. https://doi.org.10.1016/j.jas.2021.105533.
3. Wagner S, Lagane F, Seguin-Orlando A. *et al.* High-throughput DNA sequencing of ancient wood. *Mol Ecol* 2018;**27**:1138–54. https://doi.org.10.1111/mec.14514.
4. Warinner C, Rodrigues JFM, Vyas R. *et al.* Pathogens and host immunity in the ancient human oral cavity. *Nat Genet* 2014;**46**:336–44. https://doi.org.10.1038/ng.2906.
5. Zhang M, Cao P, Dai Q-Y. *et al.* Comparative analysis of DNA extraction protocols for ancient soft tissue museum samples. *Zool Res* 2021;**42**:280–6. https://doi.org/10.24272/j.issn.2095-8137.2020.377.
6. Peyrégne S, Prüfer K. Present-day DNA contamination in ancient DNA datasets. *Bioessays* 2020;**42**:2000081. https://doi.org.10.1002/bies.202000081.
7. Llamas B, Valverde G, Fehren-Schmitz L. *et al.* From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. *STAR Sci Technol Archaeol Res* 2017;**3**:1–14. https://doi.org.10.1080/20548923.2016.1258824.
8. Dabney J, Meyer M, Pääbo S. Ancient DNA damage. *Cold Spring Harb Perspect Biol* 2013;**5**:a012567. https://doi.org.10.1101/cshperspect.a012567.
9. Prüfer K, Stenzel U, Hofreiter M. *et al.* Computational challenges in the analysis of ancient DNA. *Genome Biol* 2010;**11**:R47. https://doi.org.10.1186/gb-2010-11-5-r47.
10. Feuerborn TR, Palkopoulou E, van der Valk T. *et al.* Competitive mapping allows for the identification and exclusion of human DNA contamination in ancient faunal genomic datasets. *BMC Genomics* 2020;**21**:844. https://doi.org.10.1186/s12864-020-07229-y.
11. Rasmussen S, Allentoft ME, Nielsen K. *et al.* Early divergent strains of Yersinia pestis in Eurasia 5,000 years ago. *Cell* 2015;**163**:571–82. https://doi.org.10.1016/j.cell.2015.10.009.
12. Skoglund P, Northoff BH, Shunkov MV. *et al.* Separating endogenous ancient DNA from modern day contamination in a Siberian neandertal. *Proc Natl Acad Sci* 2014;**111**:2229–34. https://doi.org.10.1073/pnas.1318934111.
13. Eisenhofer R, Weyrich LS. Assessing alignment-based taxonomic classification of ancient microbial DNA. *PeerJ* 2019;**7**:e6594. https://doi.org.10.7717/peerj.6594.
14. Velsko IM, Frantz LAF, Herbig A. *et al.* Selection of appropriate metagenome taxonomic classifiers for ancient microbiome research. *mSystems* 2018;**3**. https://doi.org/10.1128/msystems.00080-18.
15. Hall MB, Coin LJM. *Pangenome databases provide superior host removal and mycobacteria classification from clinical metagenomic*

data 2023;**2023**:18.558339. https://doi.org/10.1101/2023.09.18.558339.
16. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. *Genome Biol* 2019;**20**:257. https://doi.org.10.1186/s13059-019-1891-0.
17. Ainsworth D, Sternberg MJE, Raczy C. *et al.* k-SLAM: Accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. *Nucleic Acids Res* 2016;**45**:1649–56. https://doi.org.10.1093/nar/gkw1248.
18. Arizmendi Cárdenas YO, Neuenschwander S, Malaspinas A-S. Benchmarking metagenomics classifiers on ancient viral DNA: A simulation study. *PeerJ* 2022;**10**:e12784. https://doi.org.10.7717/peerj.12784.
19. Ye SH, Siddle KJ, Park DJ. *et al.* Benchmarking metagenomics tools for taxonomic classification. *Cell* 2019;**178**:779–94. https://doi.org.10.1016/j.cell.2019.07.010.
20. Pusadkar V, Azad RK. Benchmarking metagenomic classifiers on simulated ancient and modern metagenomic data. *Microorganisms* 2023;**11**:2478. https://doi.org.10.3390/microorganisms11102478.
21. Renaud G, Hanghøj K, Willerslev E. *et al.* Gargammel: A sequence simulator for ancient DNA. *Bioinforma Oxf Engl* 2017;**33**:577–9. https://doi.org.10.1093/bioinformatics/btw670.
22. Seguin-Orlando A, Korneliussen TS, Sikora M. *et al.* Genomic structure in Europeans dating back at least 36,200 years. *Science* 2014;**346**:1113–8. https://doi.org.10.1126/science.aaa0114.
23. Champlot S, Berthelot C, Pruvost M. *et al.* An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. *PloS One* 2010;**5**:e13042. https://doi.org.10.1371/journal.pone.0013042.
24. Leonard JA, Shanks O, Hofreiter M. *et al.* Animal DNA in PCR reagents plagues ancient DNA research. *J Archaeol Sci* 2007;**34**:1361–6. https://doi.org.10.1016/j.jas.2006.10.023.
25. Lazaridis I, Patterson N, Mittnik A. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 2014;**513**:409–13. https://doi.org.10.1038/nature13673.
26. Fu Q, Li H, Moorjani P. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 2014;**514**:445–9. https://doi.org.10.1038/nature13810.
27. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Res Notes* 2016;**9**:88. https://doi.org.10.1186/s13104-016-1900-2.
28. Peltzer A, Jäger G, Herbig A. *et al.* EAGER: Efficient ancient genome reconstruction. *Genome Biol* 2016;**17**:60. https://doi.org.10.1186/s13059-016-0918-z.
29. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* 2009;**25**:1754–60. https://doi.org/10.1093/bioinformatics/btp324.
30. Oliva A, Tobler R, Cooper A. *et al.* Systematic benchmark of ancient DNA read mapping. *Brief Bioinform* 2021;**22**:bbab076. https://doi.org.10.1093/bib/bbab076.
31. Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: Losada DE, Fernández-Luna JM, editors. *Adv Inf Retr.*, Berlin, Heidelberg: Springer; 2005;**3408**:345–59. https://doi.org/10.1007/978-3-540-31865-1_25.
32. Yates JAF, Peltzer A, Lamnidis TC, Borry M, Fagernäs Z, Bar I., *et al.* nf-core/eager: [2.5.1] - Bopfingen (Patch) - 2024. https://doi.org/10.5281/zenodo.10687430.
33. Bergström A, Stanton DWG, Taron UH. *et al.* Grey wolf genomic history reveals a dual ancestry of dogs. *Nature* 2022;**607**:313–20. https://doi.org.10.1038/s41586-022-04824-9.
34. Plassais J, Kim J, Davis BW. *et al.* Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nat Commun* 2019;**10**:1489. https://doi.org.10.1038/s41467-019-09373-w.
35. Günther T, Nettelblad C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet* 2019;**15**:e1008302. https://doi.org.10.1371/journal.pgen.1008302.
36. Zhang Z, Schwartz S, Wagner L. *et al.* A greedy algorithm for aligning DNA sequences. *J Comput Biol J Comput Mol Cell Biol* 2000;**7**:203–14. https://doi.org.10.1089/10665270050081478.
37. Huson DH, Auch AF, Qi J. *et al.* MEGAN analysis of metagenomic data. *Genome Res* 2007;**17**:377–86. https://doi.org.10.1101/gr.5969107.
38. Patterson N, Moorjani P, Luo Y. *et al.* Ancient admixture in human history. *Genetics* 2012;**192**:1065–93. https://doi.org.10.1534/genetics.112.145037.
39. Nasko DJ, Koren S, Phillippy AM. *et al.* RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol* 2018;**19**:165. https://doi.org.10.1186/s13059-018-1554-6.
40. Cano RJ, Rivera-Perez J, Toranzos GA. *et al.* Paleomicrobiology: Revealing Fecal microbiomes of ancient indigenous cultures. *PloS One* 2014;**9**:e106833. https://doi.org.10.1371/journal.pone.0106833.
41. Lu J, Salzberg SL. Removing contaminants from databases of draft genomes. *PLoS Comput Biol* 2018;**14**:e1006277. https://doi.org.10.1371/journal.pcbi.1006277.
42. Lugli GA, Milani C, Mancabelli L. *et al.* Ancient bacteria of the Ötzi's microbiome: A genomic tale from the copper age. *Microbiome* 2017;**5**:5. https://doi.org.10.1186/s40168-016-0221-y.
43. Ozkan J, Willcox MD. The ocular microbiome: Molecular characterisation of a unique and low microbial environment. *Curr Eye Res* 2019;**44**:685–94. https://doi.org.10.1080/02713683.2019.1570526.
44. Schulberg J, De Cruz P. Characterisation and therapeutic manipulation of the gut microbiome in inflammatory bowel disease. *Intern Med J* 2016;**46**:266–73. https://doi.org.10.1111/imj.13003.
45. Wade WG. Characterisation of the human oral microbiome. *J Oral Biosci* 2013;**55**:143–8. https://doi.org.10.1016/j.job.2013.06.001.
46. Watson RL, de KEM, Bogaert D. Characterising the respiratory microbiome. *Eur Respir J* 2019;53. https://doi.org/10.1183/13993003.01711-2018.
47. Krakau S, Straub D, Gourlé H. *et al.* Nf-core/mag: A best-practice pipeline for metagenome hybrid assembly and binning. *NAR Genomics Bioinforma* 2022;**4**:lqac007. https://doi.org.10.1093/nargab/lqac007.
48. Perfumo A, Çabuk U, Schulte L. *et al.* Paleometagenomics reveals environmental microbiome response to vegetation changes in northern Siberia over the millennia. *Environ DNA* 2023;**5**:1252–64. https://doi.org.10.1002/edn3.446.
49. Parks DH, Chuvochina M, Waite DW. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 2018;**36**:996–1004. https://doi.org.10.1038/nbt.4229.
50. Rohland N, Mallick S, Mah M. *et al.* Three assays for in-solution enrichment of ancient human DNA at more than a million SNPs. *Genome Res* 2022;**32**:2068–78. https://doi.org.10.1101/gr.276728.122.

51. Martiniano R, Garrison E, Jones ER. *et al.* Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biol* 2020;**21**:250. https://doi.org.10.1186/s13059-020-02160-7.
52. Vogel NA, Rubin JD, Swartz M. *et al.* Euka: Robust tetrapodic and arthropodic taxa detection from modern and ancient environmental DNA using pangenomic reference graphs. *Methods Ecol Evol* 2023;**14**:2717–27. https://doi.org.10.1111/2041-210X.14214.
53. Marić J, Križanović K, Riondet S. *et al.* Comparative analysis of metagenomic classifiers for long-read sequencing datasets. *BMC Bioinformatics* 2024;**25**:15. https://doi.org.10.1186/s12859-024-05634-8.
54. Odom AR, Faits T, Castro-Nallar E. *et al.* Metagenomic profiling pipelines improve taxonomic classification for 16S amplicon sequencing data. *Sci Rep* 2023;**13**:13957. https://doi.org.10.1038/s41598-023-40799-x.
55. Astashyn A, Tvedte ES, Sweeney D. *et al.* Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol* 2024;**25**:60. https://doi.org/10.1186/s13059-024-03198-7.
56. Lupo V, Van Vlierberghe M, Vanderschuren H. *et al.* Contamination in reference sequence databases: Time for divide-and-rule tactics. *Front Microbiol* 2021;**12**:12. https://doi.org.10.3389/fmicb.2021.755101.
57. Steinegger M, Salzberg SL. Terminating contamination: Large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol* 2020;**21**:115. https://doi.org.10.1186/s13059-020-02023-1.

Briefings in Bioinformatics, 2025, **26(1)**, bbae646
https://doi.org/10.1093/bib/bbae646
Problem Solving Protocol