Benchmarking metagenomics classifiers on ancient viral DNA: a simulation study

Yami Ommar Arizmendi Cárdenas[1,2], Samuel Neuenschwander[1,3*],
Anna-Sapfo Malaspinas[1,2*]

[1] Department of Computational Biology, University of Lausanne, 1015, Lausanne, Switzerland
[2] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[3] Vital-IT, Swiss institute of Bioinformatics, 1015 Lausanne, Switzerland

*Corresponding authors
Samuel Neuenschwander
Email: samuel.neuenschwander@unil.ch

Anna-Sapfo Malaspinas
Email: annasapfo.malaspinas@unil.ch

Abstract

Owing to technological advances in ancient DNA, it is now possible to sequence viruses from the past to track down their origin and evolution. However, ancient DNA data is considerably more degraded and contaminated than modern data making the identification of ancient viral genomes particularly challenging. Several methods to characterise the modern microbiome (and, within this, the virome) have been developed; in particular, tools that assign sequenced reads to specific taxa in order to characterise the organisms present in a sample of interest. While these existing tools are routinely used in modern data, their performance when applied to ancient microbiome data to screen for ancient viruses remains unknown.

In this work, we conducted an extensive simulation study using public viral sequences to establish which tool is the most suitable to screen ancient samples for human DNA viruses. We compared the performance of four widely used classifiers, namely Centrifuge, Kraken2, DIAMOND and MetaPhlAn2, in correctly assigning sequencing reads to the corresponding viruses. To do so, we simulated reads by adding noise typical of ancient DNA to a set of publicly available human DNA viral sequences and to the human genome. We fragmented the DNA into different lengths, added sequencing error and C to T and G to A deamination substitutions at the read termini. Then we measured the resulting sensitivity and precision for all classifiers.

Across most simulations, more than 228 out of the 233 simulated viruses are recovered by Centrifuge, Kraken2 and DIAMOND, in contrast to MetaPhlAn2 which recovers only around one third. Overall, Centrifuge and Kraken2 have the best performance with the highest values of sensitivity and precision. We found that deamination damage has little impact on the performance of the classifiers, less than the sequencing error and the length of the reads. Since Centrifuge can handle short reads (in contrast to DIAMOND and Kraken2 with default settings) and since it achieves the highest sensitivity and precision at the species level across all the simulations performed, it is our recommended tool. Regardless of the tool used, our simulations indicate that, for ancient human studies, users should use strict filters to remove all reads of potential human origin. Finally, we recommend to verify which species are present in the database used, as it might happen that default databases lack sequences for viruses of interest.

Introduction

The human body is home to different species of microorganisms (including bacteria, archaea, viruses and eukaryotes). The composition of these microorganisms is called microbiota, and the union of all their genomes is referred to as the microbiome. The microbiome is being characterised since the last few decades and is known to play an important role in human health (Anon 2019). The majority of the microbiome studies have concentrated on the most abundant organisms of the microbiota, namely the bacteria (Stern et al. 2019). However, the rest of the microbiome, especially the virome, the viral fraction of the microbiome, has gained more attention in recent years as it is tightly linked to our welfare (Pérez-Brocal and Moya 2018; Siqueira et al. 2018; Stern et al. 2019).

Ancient DNA (aDNA) sequencing allows for the reconstruction of ancient microbial genomes and has opened the door to an entire new field sometimes dubbed "paleomicrobiology". The availability of aDNA has opened a unique window into the past allowing to study the evolution of viruses and other pathogenic microbes (Taubenberger et al. 2005; Duggan et al. 2016; Worobey et al. 2016; Krause-Kyora et al. 2018; Mühlemann, Jones, et al. 2018; Mühlemann, Margaryan, et al. 2018; Vågene et al. 2018; Rascovan et al. 2019).

Even though the potential gains are enormous, the ancient microbiome is both challenging to retrieve and to analyse as aDNA is degraded and contaminated. More specifically, aDNA molecules present characteristic patterns caused by post-mortem molecular damage: fragmentation and substitutions (especially occurring at the end of the molecules). Such damage (and the associated patterns) depends on environmental factors such as humidity, temperature, salinity, pH, and microbial growth (Briggs et al. 2007; Allentoft et al. 2012; Sawyer et al. 2012; Dabney et al. 2013). Fragmentation is the consequence of depurination with abasic sites leading to single-strand breaks and, subsequently, to double-strand breaks, which cause the DNA molecule to break into small fragments (Dabney et al. 2013). Note that a large fraction of the DNA fragments is considerably shorter than 100 bp resulting in sequenced reads that often include some of the sequencing adapters. Substitutions at the end of the DNA fragments are caused by deamination resulting into an increase cytosines to thymines (C to T) substitutions at the 5' end of the fragments and guanines to adenines (G to A) substitutions at the 3' ends of the fragments for double stranded libraries (Briggs et al. 2007; Carøe et al. 2018).

In most microbiome aDNA studies the first step conducted is to assign a taxon to each of the sequenced reads using a classifier (i.e. classify the reads). The classification of the reads results in a list of microbes that are potentially ancient having infected the host prior to the death of the sampled

organism. From this list of potential candidates some microbes (usually pathogens) are selected for follow up analyses to corroborate if they are truly ancient microbes and "endogenous". The classification of reads and selection of candidates is known as the screening or scanning step (Warinner et al. 2017). The steps following the screening are computationally intensive, for instance, they generally include a mapping step (BWA (Li and Durbin 2009), BowTie2 (Langmead and Salzberg 2012)) and a blasting step (Altschul et al. 1990) to determine if a particular genome is the most likely source of the sequenced reads (Krause-Kyora et al. 2018; Mühlemann, Jones, et al. 2018; Vågene et al. 2018; Mühlemann et al. 2020).

For the screening step there are many options of classifiers available. The classifiers are diverse and differ from each other in the underlying database, sequence search algorithm, and taxonomic binning (the process of assigning a taxonomic rank to each of the reads sequenced from a sample). In the specific case of ancient microbiome studies, benchmarks have been conducted to choose an appropriate classifier (Velsko et al. 2018) or to determine the most adequate parameters to enhance the performance of specific classifiers (Eisenhofer and Weyrich 2019). Nevertheless, these benchmark studies only focus on the bacterial fraction of the microbiomes. Considering this, we have decided to perform a benchmark study to assess the performance of different classifiers in screening ancient microbiome samples for viruses; specifically, our study focuses on DNA viruses with human as their host (human DNA viruses), since most of the ancient viral DNA studies published so far target such viruses. In short, we classified simulated samples with four classifiers: Centrifuge, Kraken2, DIAMOND and MetaPhlAn2 (for a fifth classifier: MALT, see Supplementary Note 1). The classified samples consisted of simulated ancient-like DNA reads, with different length, deamination damage and sequencing error values. These samples represent an ideal case scenario as they come from a single human DNA viral species and have a high coverage. As we knew the true species in the sample, we scored the classifications per sample; we evaluated the effect of read length, deamination and sequencing error; and we compared the overall performance of all classifiers. In addition, we evaluate the effect of human contamination by classifying simulated human reads with the four classifiers.

The four classifiers assessed in the present study: Centrifuge (Kim et al. 2016), Kraken2 (Wood et al. 2019), DIAMOND (Buchfink et al. 2015) and MetaPhlAn2 (Truong et al. 2015) differ on their approaches. These classifiers were not specifically developed for aDNA; fragmentation and cytosine deamination may present a challenge for the taxonomic assignment since reads with post-mortem alterations could lead to misclassifications. In the subsequent paragraphs we will describe the

characteristics in terms of database, algorithms and taxonomic binning of these four classifiers (see Table 1 for a summary).

The four classifiers databases are characterised by the type of molecule and the genomic region(s) that are included. Among the four classifiers, two different types of molecules are used: DNA and protein sequences. In terms of genomic regions, the databases either include multiple loci from each organism (such as a set of proteins or marker genes) or whole genomes. Centrifuge uses whole genome DNA databases. Kraken2 is versatile in database usage as it can use protein or DNA, whole genome or single locus databases. DIAMOND uses a protein database and can use as query both DNA or amino acid sequences. MetaPhlAn2 relies on a multiple loci DNA database consisting of core genes that are shared within a clade but not outside of it (clade specific marker genes). In contrast to its previous version, MetaPhlAn2's database includes marker genes for viruses and eukaryotic microbes.

For the sequenced search algorithm, classifiers either rely on alignments or on exact k-mer matches. Centrifuge, DIAMOND and MetaPhlAn2 depend on alignment algorithms. Centrifuge uses the Burrows-Wheeler Transform (Burrows and Wheeler 1994) and Ferragina-Manzini index (Ferragina and Manzini 2000). These algorithms allow to create a data structure which facilitates fast alignments with efficient memory usage. Centrifuge works by searching for exact matches (i.e. no mismatches or gaps are allowed). DIAMOND implements double index alignment (both query and reference are indexed), looking for matches of seeds (short subsequences of fixed length, with default lengths of 15-24 bp) and then extending the alignment. DIAMOND's seeds are spaced seeds, meaning that some positions in the seeds are treated as wildcards. Finally, MetaPhlAn2 performs alignments using BowTie2 (Langmead and Salzberg 2012). BowTie2 is an index-assisted aligner that allows gaps in the alignment. Its algorithm has two phases: ungapped seed match, and an extension that permits gaps. Kraken2 is in contrast an alignment-free classifier which is based on exact matches of k-mers (sequence substrings of length k) (Wood and Salzberg 2014). Kraken2 builds an index of k-mers (of length 35 by default for nucleotide sequences) from the database. K-mers are first obtained from the query read and these k-mers are then looked up in the database (an exact k-mer match is performed).

Finally, the classifiers use different strategies for taxonomic binning. DIAMOND uses the lowest common ancestor (LCA) algorithm: that is, if a read matches several species, it will be assigned to the most specific ("lowest") taxonomic rank shared by all the matching species. If the read only matches one species, the taxonomic assignation will remain at the level of that species. As a consequence, conserved sequences will be assigned to higher taxonomic levels, and specific sequences will be

assigned to species level (Huson et al. 2007). Centrifuge and Kraken2 perform taxonomic binning using scoring schemes. Centrifuge ranks the alignments done with a score that favours longer hits. The query read is attributed to the taxon in the database with the highest score. Similarly, Kraken2 assigns a read to the taxon which has the most k-mer matches in common. In this case, a root-to-leaf path is used to sum k-mers coming from higher taxonomic ranks. MetaPhlAn2 relies on a clade specific marker catalogue. If there is an alignment to a marker, the read is assigned to the taxonomic clade associated with that marker. A clade can be as specific as a strain, or as broad as a phylum (Segata et al. 2012).

Materials & Methods

In short, the analyses consisted in simulating reads from reference viral sequences, classifying them with four widely used classification tools and quantifying the performance of the classifiers under different conditions. In particular, we investigated how varying the read lengths, adding DNA damage (features typical of ancient DNA) and sequencing error impact the classifications made by the tools. Finally, we analysed the assignments made by the classifiers on simulated short human reads.

Reference viral sequences

A total of 238 reference viral sequences from 233 different human DNA viruses were used in this study. To get the aforementioned sequences we first queried the Viral-Host DB (downloaded on 25 September 2019) (Mihara et al. 2016) for viruses that infect *Homo sapiens*, obtaining a list of 1,315 viruses. We downloaded the reference sequences of such viruses in genbank format from NCBI RefSeq (Brister et al. 2015; O'Leary et al. 2016) and kept the sequences if their "molecule type" record was "DNA" or "ss-DNA", obtaining a total of 250 DNA viral sequences. Finally, we removed duplicated sequences with the same taxID, representing viruses with different isolates or alternative genomes assemblies; we also removed the Human endogenous retrovirus K113 (HERV-K113) sequence (NC_022518.1), as this virus is a provirus in the human reference genome (Turner et al. 2001). This gave us a total of 233 viruses with 238 different sequences (one of the viruses is stored in RefSeq in six different contigs). A table with information about the viruses used in this study is available on Table S1; it includes the following: virus name, sequence name, sequence length, type of molecule (DNA or ss-DNA), type of genome (linear or circular), sequence accession number, virus taxID and Baltimore group. For each selected viral sequence, we generated a set of simulated reads as described below.

Simulation of viral reads and read length

ART (ART-MountRainier-2016-06-05, art_Illumina Q Version 2.5.8) (Huang et al. 2012) was used to simulate sequencing reads starting from each of the 233 human DNA viruses' sequences as reference. By default, ART simulates reads with sequencing error typical for the specified sequencer. The sequencing machine was set to Illumina HiSeq 2500 (*art_illumina -ss HS25*) for all simulations, and the parameter *qShift* (that controls the amount of sequencing error, see below) was set to 0 for most of the simulations (but also to different values when assessing the effect of sequencing error) resulting in typical Illumina HiSeq 2500 error profiles. In an initial simulation, for each of the 233 viruses, single-end reads of a fixed length were simulated emulating the short read length observed in typical ancient DNA data, see e.g. (Green et al. 2008). First, the read length was set at 60 bp (*-l 60*) and a coverage of 10× was requested (*-f 10*) resulting in 280 to 39,270 reads per virus, depending on the viral sequence length. Second, the read length was varied from 30 to 150 bp while keeping the coverage at 10×. We set the coverage of 10× (an unrealistically high number for ancient DNA data) as we wanted to compare the classifiers under ideal conditions. As the reads are classified independently of each other, we do not expect that the relative performance of the classifiers (point estimates) would change with a lower depth of coverage.

Each set of reads from the 233 selected human DNA viruses was classified using four different classifiers: Centrifuge, Kraken2, DIAMOND and MetaPhlAn2 (Table 1). The parameters for each of the classifiers are specified below.

Classification of reads

As much as possible, we used the default parameters suggested by the developers of each tool to build the databases and perform the classification as the performance of each tool is likely optimal for those parameters. Otherwise the parameters are stated.

Centrifuge

Centrifuge version 1.0.3-beta was used. To build a custom viral database we downloaded all the viral sequences stored in RefSeq on 12 September 2021 from https://ftp.ncbi.nlm.nih.gov/refseq/release/viral/viral.${i}.1.genomic.fna.gz (${i} has values from 1 to 3) and NCBI's taxonomy nodes information (files nodes.dmp and names.dmp) from https://ftp.ncbi.nih.gov/pub/taxonomy/taxdmp.zip on 14 October 2021. We created our own "conversion

table" file required by Centrifuge which links the accession number to its taxID using the following command of NCBI's E-utilities (Anon):
`esearch -db nuccore -query "${accession}[Accession]" | elink -target taxonomy | efetch -format xml | xtract -pattern Taxon -element TaxId`, where "${accession}" represents the accession number. We then created the database with the following command:
`centrifuge-build --conversion-table my_viral_accession2taxid.map --taxonomy-tree nodes.dmp --name-table names.dmp all_viral.genomic.fna Centrifuge_viral`, where "my_viral_accession2taxid.map" is our "conversion-table" file; "nodes.dmp" and "names.dmp" are the taxonomy files downloaded from NCBI Taxonomy; "all_viral.genomic.fna" contains all RefSeq viral sequences in fasta format; and "Centrifuge_viral" is the name of the created database.
We also created Centrifuge's default database which includes all RefSeq sequences from bacterial, archaeal, viral and human genomes (built on 2 September 2021) following the software instructions). The results obtained with the default database are shown in Supplementary Note 2.
When running Centrifuge to classify the reads a maximum of one taxonomic assignment per read was retrieved (*-k 1*), to obtain an output comparable to the other classifiers. All other parameters were left with default values.

Kraken2

Kraken2 version 2.0.7-beta was run with default parameters. To build our viral custom database first we downloaded the taxonomy using the following command `kraken2-build --download-taxonomy --db Kraken2_viral_db`, where Kraken2_viral_db is the name of the database. Then we replace all the internal files (except "Kraken2_viral_db/taxonomy") to have the following directory structure:

```
Kraken2_viral_db/
|
|-- library
|   \-- viral
|       \-- library.fna
|
|-- seqid2taxid.map
|
\-- taxonomy
```

Where "library.fna" is the same as "all_viral.genomic.fna" and "seqid2taxid.map" is the same as "my_viral_accession2taxid.map"; both files used to build Centrifuge's custom viral database. To build the database the following command was used: `kraken2-build --build --db Kraken2_viral_db`.

We also downloaded Kraken2 the prebuilt "standard" database from https://benlangmead.github.io/aws-indexes/k2 dated to 17 May 2021. This database includes the RefSeq of bacterial, archaeal, viral and human genomes plus UniVec_Core. The results obtained with this default database are shown in Supplementary Note 2.

We also performed analyses building a database with a shorter k-mer length in order to classify short (30 bp) reads, for these results see Supplementary Note 3.

DIAMOND

DIAMOND version 0.9.22 was used with the RefSeq viral protein database. The protein sequences were downloaded from https://ftp.ncbi.nlm.nih.gov/refseq/release/viral/viral.${i}.protein.faa.gz (${i} has values from 1 to 3); the taxonomy nodes where downloaded from https://ftp.ncbi.nih.gov/pub/taxonomy/taxdmp.zip; and the file which links accession number to taxID was downloaded from https://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/prot.accession2taxid.gz. All files were downloaded on 1 September 2021. The database was built following the manual's instructions.

In addition, we built a "full" database with archaeal, bacterial, human and viral protein sequences. The viral sequences and taxonomy files are identical as the previous database. We downloaded the archaeal sequences from https://ftp.ncbi.nlm.nih.gov/refseq/release/archaea/archaea.nonredundant_protein.${i}.protein.faa.gz (${i} has values from 1 to 12); the bacterial from https://ftp.ncbi.nlm.nih.gov/refseq/release/bacteria/bacteria.nonredundant_protein.${i}.protein.faa.gz (${i} has values from 1 to 1396); and human from https://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/human.${i}.protein.faa.gz (${i} has values from 1 to 11). Archaeal and bacterial protein sequences were downloaded on 1 September 2021; human protein sequences were downloaded on 6 September 2021. The results obtained with this "full" database are shown in Supplementary Note 2.

DIAMOND was run in the taxonomic classification mode (*-f 102* option) in order to perform the taxonomic assignments; all other parameters were set as default. As the input data is DNA, the query is translated into protein for each of its six reading frames and aligned against the protein database.

We also performed analyses changing the default seed length used by DIAMOND in order to classifiy short (30 bp) reads, for these results see Supplementary Note 3.

MetaPhlAn2

MetaPhlAn2 version 2.6.0 was run with default parameters. The default microbe clade-specific marker genes database was used (mpa_v20_m200 version), it includes marker genes from bacteria, archaea, viruses and eukaryotic microbes.

Summary statistics

To evaluate the performance of the classifiers, each read was assigned to one of the following categories: i) correctly classified read at the species level ("correct species"), ii) correctly classified read at a higher taxa level ("correct higher"), iii) misclassified read at any taxa level ("incorrect") and iv) unclassified read ("unclassified"). Figure 1 shows a schematic example of each of the four categories for *Variola virus*. The correctly classified reads are divided into two sets hereafter: a set including only the reads classified correctly at the species level ("correct species", abbreviated "s" below) and a set including all the reads classified correctly at the species and at higher taxa ("correct species" and "correct higher", abbreviated "s&h" below). Further, we computed the widely used statistics sensitivity and precision to compare the classifiers (Wood and Salzberg 2014). Sensitivity is the proportion of correctly classified reads over all simulated reads; precision is the proportion of correctly classified reads over all classified reads. We computed two sets of sensitivity and precision measures depending on whether "correct species" is assumed to be correct or "correct species" and "correct higher" are assumed to be correct. More specifically, sensitivity and precision are defined as follows for each viral sequence v_i:

$$Sensitivity_s^{v_i} = \frac{r_{cs}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i} + r_u^{v_i}}$$

$$Sensitivity_{s\&h}^{v_i} = \frac{r_{cs}^{v_i} + r_{ch}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i} + r_u^{v_i}}$$

$$Precision_s^{v_i} = \frac{r_{cs}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i}}$$

$$Precision_{s\&h}^{v_i} = \frac{r_{cs}^{v_i} + r_{ch}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i}}$$

where $r_{cs}^{v_i}$ are the number of "correct species" reads for the viral sequence v_i, $r_{ch}^{v_i}$ are the number of "correct higher" reads for the viral sequence v_i, $r_i^{v_i}$ are the number of "incorrect" reads for the viral sequence v_i, $r_u^{v_i}$ are the number of "unclassified" reads for the viral sequence v_i. For some of the viral

sequences v_j, all reads are "unclassified" and $r_{cs}^{v_j} + r_{ch}^{v_j} + r_i^{v_j} = 0$. In this case, the $Precision_s^{v_j}$ and the $Precision_{s\&h}^{v_j}$ are undefined.

We summarize the results by computing the mean sensitivities and precisions across the simulated viral sequences. For the sensitivities, we have:

$$Sensitivity_s = \frac{1}{233}\sum_{i=1}^{233}\frac{r_{cs}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i} + r_u^{v_i}}$$

$$Sensitivity_{s\&h} = \frac{1}{233}\sum_{i=1}^{233}\frac{r_{cs}^{v_i} + r_{ch}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i} + r_u^{v_i}}$$

For the precisions, we have:

$$Precision_s = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{r_{cs}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i}}$$

$$Precision_{s\&h} = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{r_{cs}^{v_i} + r_{ch}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i}}$$

where n_c is the number of viral sequences for which there is at least one read classified (correctly or not), i.e. for which $r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i} > 0$.

In addition, we counted and characterised the number of spurious extra taxa identified by the classifiers. These taxa are associated with one or more "incorrect" reads. They include any taxa that were not used to simulate the sequencing reads or taxa not included in the lineages of the simulated viruses.

Effect of adding deamination damage

Besides DNA fragmentation, another molecular characteristic of ancient DNA is the deamination process which, for double stranded libraries, results in an increase of C to T substitutions at the 5' ends and G to A substitutions at the 3' ends (Briggs et al. 2007; Carøe et al. 2018). We simulated reads of 60 bp length and added different levels of deamination using deamSim (June 06 2016 version) gargammel sub-program (Renaud et al. 2017). The following parameters were applied: a nick frequency of 0.03, a geometric parameter of 0.25 for the average length of overhanging ends, a probability of deamination in the double-stranded portions of DNA of 0.01, and 11 different values for the probability of deamination in the single stranded portions of the DNA (0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45 and 0.5). Note that the resulting damage pattern is similar to what has been observed for real data (Figure 2) with

higher deamination fractions at read termini. The resulting observed average number of substitutions was computed for the case of NC_006273.2 (*Human betaherpes virus 5*) for all tested single-stranded probabilities of deamination (Figure 3).

Effect of increasing substitution sequencing error

To better understand the effect of the distribution of errors across the reads, substitution errors were also simulated following an Illumina-like error profiles. To do so, sequencing errors were added to the simulated reads using ART (Huang et al. 2012) indirectly varying the overall error rate by changing the parameter *qShift* of ART. Specifically, seven different values for the parameter *qShift* were tested (0, -1, -3, -5, -7 and -9). This parameter changes the quality score of the simulated reads, leading to more substitutions in the simulated reads with lower values. Each *qShift* value corresponds to an overall fold increase of substitution error (Huang et al. 2012):

$$fold_increase = \frac{1}{10^{\frac{qShift}{10}}}$$

For instance, a *qShift* of 0 corresponds to a one-fold increase (i.e. no increase or in other words typical sequencing error), -3 to a ~two-fold increase and, finally, -9 to a 7.9-fold increase. The effect of decreasing *qShift* on the observed number of substitutions in simulated reads is shown in Figure 3 for the case of NC_006273.2 (*Human betaherpes virus 5*) values. For this analysis also the simulated read length was set at 60 bp.

Human simulations

When human remains are sequenced, the resulting data contains microbial and human DNA. To study the microbiome, ideally, reads mapping to the human genomes should be first removed. However, such cleaning steps are imperfect and can lead to false classifications. To evaluate how the tools classify potential human sequences that were not properly filtered out, human reads were simulated using ART (ART-MountRainier-2016-06-05, art_Illumina Q Version 2.5.8) with the human genome (GRCh37) as reference. As above, the read length used was 60 bp, the coverage 10×, and the sequencing technology was set to Illumina HiSeq 2500. Centrifuge, Kraken2, DIAMOND and MetaPhlAn2 were then used to classify the reads as above. The proportions and the number of correctly (i.e. classified as human) and incorrectly (i.e. not classified as human or at as a taxon within the human lineage) classified reads were calculated; the number of extra taxa were counted and characterised by determining the assigned kingdom and the number of assigned reads per taxon.

Results

Overall, the performance of the four classifiers is surprisingly distinct across simulations (Figure 4-9). The results for Centrifuge and Kraken2 are the most similar to each other, MetaPhlAn2 is the most distinct compared to Centrifuge and Kraken2, while DIAMOND occupies an intermediate position.

Classification results for the 60 bp read set

Figure 4 and Figure S1 summarise the results for the simulated 60 bp read set. As discussed above, the simulated reads could be either correctly classified at the species level ("correct species"), correctly classified at a higher taxa level ("correct higher"), misclassified ("incorrect"), or "unclassified" (Figure 1).

When considering each virus separately, most reads are classified correctly at the species level ("correct species") for most of the simulated viral sequences for Centrifuge and Kraken2 (Figure S1). In contrast, for DIAMOND, the proportion of "correct species" varies substantially with many viruses with low, intermediate and high "correct species" proportions while the remaining reads are mostly "unclassified" or "correct higher". MetaPhlAn2 stands out by having roughly two thirds of the viruses with no correctly classified reads (neither as "correct species" or "correct higher"). As a result, the means across viruses of the proportion of "correct species" are 24.41% (MetaPhlAn2), 53.61% (DIAMOND), 93.39% (Kraken2) and 94.48% (Centrifuge). Unlike the other classifiers, DIAMOND classifies a large proportion of reads correctly but at a higher taxonomic rank ("correct higher") for many of the viruses. The other three classifiers have a handful of viruses with a large proportion of "correct higher". Overall, most of the simulated viral sequences have zero or a small proportion of "incorrect" reads, especially in the case of Centrifuge and Kraken2 (Figure S1, for detailed list of the viruses with the highest incorrect proportions per classifiers see Supplementary Note 4; Table S2 contains the raw numbers of reads assigned to each taxa per classifier). The resulting averages of the proportions of "incorrect" reads are 0.01% (Centrifuge), 0.26% (Kraken2), 0.75% (DIAMOND) and 2.79% (MetaPhlAn2). Finally, Centrifuge and Kraken2 classify most of the simulated reads for all viruses (with a mean "unclassified" of 0.01% and 0.44%), while both DIAMOND (mean of 20.4%) and MetaPhlAn2 (mean of 59.8%) do not classify a large proportion of reads for most viruses (Figure S2).

The sensitivity and precision reflect the previous "correct species", "correct higher", "incorrect" and "unclassified" overall values (Figure 4A circles and triangles). At the species level, both means are considerably higher for Centrifuge and Kraken2 with mean values above 93% for Sensitivity_s and

Precision_s. In contrast, MetaPhlAn2's sensitivity and precision are the lowest with values of 24.4% (Sensitivity_s) and 46.3% (Precision_s). MetaPhlAn2's low Sensitivity_s value is a consequence of the high proportions of "unclassified" reads for this classifier. Note that for this classifier, the precision is computed by excluding the viruses without any classified reads, i.e. by excluding most of the data (see Material and Methods). At the species level, DIAMOND has intermediate values, with a Sensitivity_s of 53.6% and a Precision_s of 67.4%; both lower than Centrifuge and Kraken2's, but higher than MetaPhlAn2's (Figure 4A circles).

When considering the "correct species" and "correct higher" reads to compute the precision and the sensitivity (Figure 4A triangles), the order of the classifiers stays the same, however the performance of DIAMOND and MetaPhlAn2 are greatly improved. In particular, DIAMOND has many viruses with a large fraction of "correct higher" reads. As a result, DIAMOND's Sensitivity_s&h is considerably higher than its Sensitivity_s with values of 78.9% and 53.6%, respectively. Similarly, its Precision_s&h (98.6%) is higher than its Precision_s (67.4%). DIAMOND's Sensitivity_s&h is lower than Centrifuge's and Kraken2's, and higher than MetaPhlAn2's. Centrifuge and Kraken2 achieve the highest Precision_s&h values with 99.9% and 99.7%, respectively; followed closely by DIAMOND (98.6%), and lastly by MetaPhlAn2 (74.4%) (Figure 4A triangles).

Beyond the proportion of correctly classified reads per virus, we also considered which viruses used in the simulations were recovered with at least one correctly classified read (Figure 4B). Encouragingly, despite large differences observed at the read level, three classifiers detected almost all the 233 viruses tested: Centrifuge and Kraken2 identified 232 viruses, and DIAMOND detected 228. In contrast MetaPhlAn2 reported only 87 viruses.

In addition, we computed how many spurious extra taxa were reported, these are taxa that were neither the virus of interest nor in the lineage of that virus (Figure 4C). Encouragingly, no classifier reports more than 2 extra taxa per virus on average. We found that MetaPhlAn2 is the classifier with the highest number of spurious extra taxa (on average 1.45 across tested viruses), followed by DIAMOND (1.06), Kraken2 (0.58) and, finally, Centrifuge (0.41).

Effect of increasing read length on the classification performance

To better understand the effect of the length of the sequenced reads on the classification results, we then investigated the performance of the classifiers for read lengths ranging from 30 to 150 bp (Figure 5, Figures S2 & S3).

In general, by increasing the read length, the classifications improve. We observe higher proportions of "correct species" reads (Figure S2 & S3), fewer "incorrect" reads, higher sensitivity and precision, and a decrease in the number of spurious extra taxa (Figures S2 & S3).

The most drastic result across classifiers for the read length simulations is that DIAMOND and Kraken2 do not classify very short reads. Specifically, no 30 bp reads are classified for DIAMOND and for Kraken2 given the default parameters used. We have changed their default settings to classify these short 30 bp reads, however we found that the number of classified reads was minimal for DIAMOND. In the case of Kraken2, it classifies less than half of the reads and presents an increase of spurious extra taxa reported, the results are shown in Supplementary Note 3. Similarly, DIAMOND classifies only very few reads for 40 bp and 50 bp compared to higher read lengths. This contrasts with the results for Centrifuge and MetaPhlAn2 whose performance do not significantly deteriorate with shorter reads in the case of Centrifuge, or it remains stable across all lengths in the case of MetaPhlAn2 (Figure 5).

From 60 bp onwards, most of the observations discussed above hold; Centrifuge and Kraken2 have the highest "correct species" mean proportions, followed by DIAMOND and MetaPhlAn2 (Figure S3). DIAMOND has the highest "correct higher" proportions; and Centrifuge has the lowest "incorrect" mean proportions. Consequently, from 60 bp onwards, the highest sensitivities (Sensitivity_s) and precisions (Precision_s) are observed for Centrifuge and Kraken2 followed by DIAMOND and MetaPhlAn2. Regarding the "correct higher" classifications, Centrifuge, Kraken2 and DIAMOND have the highest Precision_s&h; as DIAMOND classifies correctly many reads at higher taxonomic level its precision is greatly benefited. DIAMOND's has a Sensitivity_s&h that is lower than Centrifuge's and Kraken2's but higher than MetaPhlAn2. Moreover, as above, regardless of the simulated read length, the number of detected viruses remains the same. Most viruses are detected by Centrifuge, Kraken2 and DIAMOND while an important fraction of the tested viruses are missed by MetaPhlAn2 (Figure 5C).

A sharp decrease in the number of extra taxa for increased read lengths is observed for MetaPhlAn2 (Figure 5D), the most constant classifier across read lengths for sensitivity. A slighter decrease of spurious extra taxa is observed for Centrifuge and Kraken2. In the case of DIAMOND, we see an increase up to 90 bp, followed by a decrease. The highest number of spurious extra taxa is found in MetaPhlAn2 up to 60 bp; for read lengths above 60 bp DIAMOND present the highest number (Figure 5D).

When looking more in detail per classifier, for read lengths of 40 bp and longer, the proportion of "correct species" reads increases for Centrifuge and Kraken2, while the proportions of "correct higher"

and “incorrect” reads decreases (Figure S2 & S3). This results in a slight increase in the sensitivity and precision with longer reads (Figure 5A & B). For DIAMOND, the fraction of “unclassified” reads decreases dramatically from 100% (30 bp read lengths) to less than 25% (150 bp read lengths). Among the reads that are classified, the “correct species” and “correct higher” ones increase substantially up to 90 bp while the proportion of “correct higher” slightly decreases above 90 bp (Figure S2 & S3). Unlike other classifiers, substantially more reads are classified with longer read lengths, and a slight increase in the proportion of “incorrect” reads is observed as well (Figure S2 & S3). As a result, the sensitivity improves substantially for longer reads for DIAMOND, especially for the Sensitivity_s&h while the precision’s increase is slighter (Figure 5).

MetaPhlAn2 is the most stable classifier when varying read length with the classification categories “correct species” and “correct higher” remaining almost constant (Figure S2 & S3). The only summary statistics that changes is the proportion of “incorrect” reads which slightly decrease when increasing read length, resulting also in a decrease in the number of spurious extra taxa. Consequently, the sensitivity remains constant and the precision increases with longer reads (Figure 5A & B).

Effect of the deamination damage on the classification performance

To further assess the performance of the classifiers for ancient DNA like data, we added deamination damage characteristic of ancient double-stranded DNA libraries (increase of C to T substitutions at the 5’ read termini and G to A substitutions at the 3’ read termini, see Figure 2) to the simulated 60 bp reads and evaluated the impact of deamination on the taxonomic assignments (Figure 6). The probability of deamination in the double-stranded portions of DNA was set at 0.01 for all simulations, while we varied the different probabilities of single stranded deamination from 0 to 0.5 (see Material and Methods). Those parameters result in realistic damage profiles, with 3’ values of up to 0.45 being on the higher end of what can be observed (Figure 3), see e.g. (Malaspinas et al. 2014; Allentoft et al. 2015).

Overall, reassuringly, deamination has little to no effect on the classifiers’ performance. The slight reduction of “correct species” reads leads to a hardly noticeable decrease of the sensitivities (Figure 6A). Similarly, the precision of all classifiers remains constant for all values of the single-stranded deamination probabilities (Figure 6B). The deamination damage has no effect, as well, in the number of correctly detected viruses (among the 233 viruses tested) (Figure 6C).

Deamination only impacts the average number of spurious extra taxa for Centrifuge, Kraken2 and DIAMOND, which increases (Figure 6D). In contrast, for MetaPhlAn2 the average number of

spurious extra taxa presents a small reduction. For single-stranded probability of deamination values ranging from 0 to 0.15 Centrifuge reports the lowest number of spurious extra taxa; from 0.2 and above DIAMOND has the lowest numbers. MetaPhlAn2 reports the highest number of spurious extra taxa up to 0.35; above that value Centrifuge presents higher numbers.

Effect of the substitution sequencing error on the classification performance

As the effects of ancient DNA deamination were minor, we indirectly investigated whether this was the result of the unusual ancient DNA-like distribution and type (C to T in 5' ends and G to A in 3' ends) of the substitutions compared to standard sequencing error. For these simulations, errors were added using ART which assumes a profile similar to the ones observed for Illumina Sequencing machines (HiSeq 2500) by decreasing the *qShift* parameter resulting in an increase of 1 to 7.9-fold of the overall error rate. As with damage, a higher number of errors leads to divergent reads (reads with different sequences compared to the reference from which they were generated, Figure 3).

As for damage, DIAMOND and MetaPhlAn2 are essentially not affected by the sequencing error across simulations with results for a 7.9-fold increase in error rate essentially identical to the ones at 1-fold across all statistics (Figure 7, Figures S2 & S3).

The trends observed for Centrifuge and Kraken2 (increase of "unclassified" reads and small decrease of the rest of the categories) are exacerbated with Illumina-like sequencing error compared to damage (Figure S2 & S3). This result is an almost linear decrease in sensitivity when compared to the fold increase in error for Centrifuge and Kraken2 (Figure 7). However, Centrifuge and Kraken2 precision values remain constant, the reason for this is that the absolute number of correctly classified ("correct species" and "correct higher") is high relative to "incorrect" along the increased substitutions (Figure 7B).

When considering not only the proportion of reads but also the viruses identified, we observe that the number of recovered true viruses is constant regardless of the error rate substitutions. Moreover, as for deamination, the number of spurious extra taxa increases substantially for Centrifuge and Kraken2 with increased number of substitutions, while DIAMOND's increase is less sharp and MetaPhlAn2 presents an overall small reduction (Figure 7 D).

Classification of human reads

In recent years, many ancient microbes have been detected in ancient human remains. In those studies, one of the challenges is to disentangle human from microbial DNA. To assess the effect of not

filtering out effectively human DNA, simulated reads were generated from the reference human genome and used as input to the classifiers.

Almost the totality of reads (more than 98%) remains unclassified for all four classifiers (Figure 8A), which is expected as none of them have human sequences in the databases used. MetaPhlAn2 and DIAMOND have the highest proportion of unclassified reads (99.98%), followed by Kraken2 (99.5%) and Centrifuge (98.4%). Among all classifiers, Centrifuge presents the highest number of incorrectly assigned reads (Figure 8B). Above 78% of the classifications were made at the species level for all classifiers; and about a quarter of the assignments were within the Heunggongvirae kingdom (Figure 9).

Interestingly, the number of spurious extra taxa varied greatly across classifiers (Figure 8C), being Centrifuge the tool with the largest number of spurious extra taxa with 7,118 identified by at least 1 read, followed by Kraken2 (1,390), DIAMOND (528) and MetaPhlan2 (91). Table S3 contains a list of each taxon identified by at least one read for each classifier, we observed that several taxa have a large number of assigned reads (several thousands) for all classifiers.

Discussion

By providing a direct window into the past, ancient virome studies have the potential to shed light into the pathogens responsible for historical epidemics, to uncover prehistorical epidemics, but also to provide clues about the molecular biology and the evolutionary history of ancient viruses. Yet, finding ancient viruses is akin "finding broken needles in noisy haystacks". Ancient genomes are fragmented, affected by post-mortem damage, incomplete and contaminated by ancient and present organisms. Furthermore, the DNA of viruses represent only a tiny fraction of the DNA extracted from the host. Thus, the recovery of ancient viruses' genomes is an experimental and computational challenge and, given how rare ancient samples are, it is crucial to recover as much ancient DNA as possible.

Our approach and its limitations

To get a sense of the best suited classifier to screen for viruses in ancient microbiome samples, we compared state-of-the art classifiers under controlled conditions, i.e. in silico simulations. To do so, we selected 238 viral sequences from 233 human DNA viruses, fragmented them and added ancient DNA-like noise, and classified the resulting short and damaged reads with Centrifuge, Kraken2, DIAMOND and MetaPhlAn2. These classifiers (that were not specifically developed for ancient DNA) differ in essentially all possible ways including the underlying databases, the sequence search algorithms and the taxonomic binning strategies (Table 1). Hereafter we provide some clues to explain the main

results that were obtained for each classifier and conclude with some recommendations. However, it is important to note that our results and recommendations are limited to the four tested classifiers and to the databases available at the time of the analyses. Moreover, the simulations are a proxy of a real-life situations (i.e. a screening step in a microbiome study). In other words, our recommendations are a good starting point for anyone wishing to study ancient viromes. Yet future work should allow to update and to further refine them.

Unidentified and incorrectly classified viruses

In most simulations with reads over 50 bp long, three of the classifiers, namely Centrifuge, Kraken2 and DIAMOND, successfully detect at the species level more than 228 out of the 233 tested viruses. These encouraging results suggests that if viral genomes were present in a sample with sufficient coverage, they would be detected by Centrifuge, Kraken2 and DIAMOND even if the reads are fragmented and damaged. In contrast, MetaPhlAn2 missed 146 viruses. Regarding the incorrect assignments, encouragingly, for almost all viruses across all classifiers (except for MetaPhlAn2), the incorrect proportion does not exceed 43%. For a list as well as a discussion of the unidentified viruses and viruses with the highest incorrect proportions see Supplementary Note 4.

Classification of very short (30 bp) reads

One of the main characteristics of ancient DNA is the highly fragmented nature of the molecules, generally shorter than the sequencing read length (resulting in the sequencing of the adapters). The classifiers we tested were not developed to tackle such short reads and as expected, longer reads, which contain more information, positively impact the overall classification results. Our simulations suggest that DIAMOND and Kraken2's performance is substantially affected by a reduction in read length. This contrasts with the result for Centrifuge and especially MetaPhlAn2 showing these two classifiers can still handle shorter reads. In particular, we observed that 30 bp long reads are not classified at all by DIAMOND and Kraken2, using their default parameters. DIAMOND's performance is still substantially reduced at 40 bp and 50 bp while Kraken2's sensitivity is close to its highest value at 40 bp. In other words, a significant number of reads shorter than 40 bp would be lost when using Kraken2; and in the case of DIAMOND, shorter than 60 bp. These results can be explained by the underlying algorithms and databases. Kraken2's default k-mer length (35 bp) was used to build its database, making an exact match impossible for 30 bp, as the k-mers are longer than the reads. Similarly, DIAMOND uses as seeds 4 shapes of length 15-24 and weight 12 by default which translates into 45-72 bp and weight 36. This explains why it failed to classify any read of 30 bp long, and why it has issues classifying 40 bp reads

(the reads being only 4 bp longer than the default seed weight). Relatedly, DIAMOND is the only classifier depending on protein alignments and longer DNA reads are required to align the same number of amino acids as nucleotides. Even when changing the default settings of DIAMOND and Kraken2 the number of short reads classified is minimal (DIAMOND) or it comes with a trade-off: higher number of spurious taxa reported (Kraken2), see Supplementary Note 3.

A high sensitivity and precision for Centrifuge and Kraken2

As aDNA studies are usually limited in the amount of starting biological material, it is crucial to assign correctly as many reads as possible and to minimise the number of errors, i.e. to use a classifier with a high sensitivity and precision. With sensitivity and precision values at the species level around 90% and above, Centrifuge and Kraken2 outperform DIAMOND and MetaPhlAn2. In other words, Centrifuge and Kraken2 classify correctly more reads among the simulated reads but also classify correctly more reads among the classified ones. In comparison, DIAMOND and MetaPhlAn2 detect correctly less than 65% of the reads (sensitivity) on average across viruses (Figure 4A). For the sensitivity, this can likely be explained in part by the differences in database; Centrifuge and Kraken2 include whole genomes, while DIAMOND's database contains proteins and MetaPhlAn2's contains custom clade specific markers, i.e. only a fraction of the genomes in both cases. The latter two classifiers have therefore a large proportion of unclassified reads. DIAMOND's precision is not higher at the species level as it classifies correctly a large fraction of reads at higher taxonomic ranks ("correct higher", see below). MetaPhlAn2's mean precision is considerably lower than the other classifiers because it has high proportions of "incorrect" reads for some viruses, and because the number of viruses used to compute it is smaller (only the viruses with classifications are considered, and many viruses had 0 classified reads). As mentioned in Supplementary Note 4, three and 145 out of the 233 tested viruses do not have sequences stored in DIAMOND's and MetaPhlAn2's databases, respectively. As a result, both sensitivity and precision are reduced for both tools. When computing the statistics considering only the viruses that are present in their databases their performance improves: with 60 bp long reads DIAMOND reaches a Sensitivity_s of 54.31%, Sensitivity_s&h of 79.58%, Precision_s of 68.24% and Precision_s&h of 99.24%; while MetaPhlAn2 reaches a Sensitivity_s of 65.37%, a Sensitivity_s&h of 69.67%, a Precision_s of 89.9% and a Precision_s&h of 96.92%. Despite this improvement both of them are still outperformed by Centrifuge and Kraken2.

High proportions of "correct higher" reads for DIAMOND

DIAMOND exhibits a markedly higher number of reads correctly classified at a higher taxonomic rank compared to the other classifiers. Protein changes have the potential to directly impact the phenotype of an organisms and are therefore more conserved (Li 1997) and DIAMOND is the only classifier based on a protein database. Thus, the protein database (which benefits from the redundancy of the genetic code and the high degree of conservation of proteins) but also the LCA binning algorithm implemented in DIAMOND may explain why this classifier has a much larger fraction of reads classified as "correct higher", and as a result a great improvement of its precision when considering the higher taxa. As an example, the virus with the highest proportion of reads assigned to higher levels in DIAMOND, human erythrovirus V9 (NC_004295.1) has 84.8% of its reads correctly assigned to the *Erythroparvovirus* genus. All these reads aligned to proteins of the correct species human erythrovirus V9, but also to proteins from species within the genus. Given the use of the LCA algorithm, DIAMOND reported the reads at genus level. Consequently, when using DIAMOND, one should consider jointly the classifications at the species level and above.

Sensitivity and precision are generally robust to deamination damage

Besides fragmentation, deamination damage is another key feature of ancient DNA that could negatively impact classifications. Encouragingly, our simulations suggest that ancient DNA-like damage has little to no effect on the classifiers' performance aside for an increased number of spurious extra taxa for Centrifuge, MetaPhlAn2 and Kraken2 (see below). To avoid a high number of spurious extra taxa, the use of DIAMOND would be convenient, as long as the reads are long enough. Interestingly, when comparing deamination and sequencing error simulations resulting in similar number of mismatches (i.e. when the single stranded probability of deamination is set at 0.5 and the *qShift* value at -9), deamination has little effect, while sequencing error reduces the sensitivities of Centrifuge and Kraken2. This suggests having errors mostly concentrated at the end of the reads (ancient DNA-like errors) do not hinder their classification as much as Illumina-like errors, which are any kind of substitutions more evenly distributed across the read (Pfeiffer et al. 2018).

Extra spurious taxa

Across all simulations with reads above 30 bp, Centrifuge and Kraken2 achieved the lowest fraction of misclassified reads across viruses and MetaPhlAn2 the highest. Besides the proportion of "incorrect" reads, we also investigated the number of spurious extra taxa reported since they provide a

clue of the amount of follow-up work to confirm candidate microbes. In length simulations MetaPhlAn2 and DIAMOND have the highest number of spurious extra taxa. In contrast, Centrifuge is the classifier with the lowest number of spurious taxa for the same simulations; as the length increases the number of spurious extra taxa decreases. This could be explained by Centrifuge's taxonomic binning strategy with scores favouring longer hits. The contrasting high numbers of MetaPhlAn2 could be explained by its database: with MetaPhlAn2's database missing markers for 145 of the tested viruses and having markers for diverse organisms (archaea, bacteria and eukaryotes, besides the viruses). However, it is important to mention that in the human simulations, with the true species absent in all databases, Centrifuge is the tool with the highest number of spurious taxa.

For high values of deamination and sequencing error MetaPhlAn2 and DIAMOND are exceeded by Centrifuge and Kraken2. It seems that DIAMOND and MetaPhlAn2 alignment query algorithms prevent them of increasing the number of spurious taxa as much as Centrifuge or Kraken2. DIAMOND, as well, could be more robust to noise generated by the substitutions as it depends on a protein database.

Human reads can be classified as viruses, archaea, bacteria or other eukaryotes

Ancient human reads mistakenly classified as viruses can be very problematic in studies of human remains as both viral and human reads would exhibit characteristic ancient DNA features such as fragmentation and deamination damage. To assess whether human reads had any chance of being classified as viruses if they were not properly cleaned out, we classified simulated human reads. The results show that encouragingly, most reads remain as unclassified (the species is not in the database). Nevertheless, thousands of reads are misclassified. Most of those incorrectly classified reads are assigned to a wrong species (false positive), giving the user a sense of false confidence in the classification as it is so specific. The assignments are generally hard to interpret except for a retrovirus identified by Centrifuge, Kraken2 and DIAMOND. In this case, one of the top hits (i.e. the taxa with the highest numbers of reads, see Table S3) was human endogenous retrovirus K113 (HERV-K113). As its name suggests this viral genome is integrated in the human genome (Turner et al. 2001). As a result, and as we did not filter any of the simulated human data prior to classification, some simulated human reads from the reference assembly are assigned to the HERV-K113 present in the databases. The simulated reads coming from the location in the human genome where the endogenous retrovirus was integrated aligned to the HERV-K113 in the database. Note that if there had been a filtering step for human reads (such as the removal of reads that map to the reference human genome), this virus would not have been

recovered (the reads would have mapped to the proviral form of HERV-K113 in the reference human genome, hence would have been filtered out).

Recommendations and future directions

In the context of ancient virome studies, the ideal classification tool is one that allows us to recover all viruses present in the studied sample but also that does not identify any spurious extra viruses. In other words, we hope to find the right trade-off between having a high number of true positives and a low number of false positives. In this study, we considered both the fraction of reads per virus that were correctly classified, as well as the number of recovered viruses and the number of spurious extra taxa with at least one read assignment.

Sequenced ancient genomes are generally incomplete and being able to correctly classify most regions in the viral genomes will be an advantage to recover ancient viruses. Given the difficulty in finding ancient viruses and in being certain they are actually ancient, every lead in the screening phase is generally followed by extensive work. These steps generally include the mapping of sequenced data to individual candidate genomes identified by classifiers (Krause-Kyora et al. 2018; Mühlemann, Jones, et al. 2018; Vågene et al. 2018; Mühlemann et al. 2020). Any classification is treated as a candidate and reducing the number of false positives is key to achieve a manageable workload, to decrease the computational resources needed and to minimize false claims.

We find that, when considering the fraction of correctly classified reads, Centrifuge and Kraken2 exhibit the highest sensitivity and precision at the species level. When including higher taxonomic ranks, DIAMOND reaches a precision close to those of Centrifuge and Kraken2; although the sensitivity remains considerably lower. Moreover, Centrifuge, Kraken2 and DIAMOND recover essentially all the simulated viruses for most simulations. Regarding Kraken2 and DIAMOND, very short fragments cannot be classified with these two classifiers when using default parameters (shorter than 35 bp in the case of Kraken2, and shorter than 45 bp for DIAMOND). We performed additional analyses by adjusting some of the parameters (using a seed of 9 amino acids for DIAMOND and reducing the k-mer size to 29 bp when building the database for Kraken2) so these tools can handle shorter sequences (see Supplementary Note 3). In summary, DIAMOND classifies very few additional reads. Moreover, for Kraken2, there is a trade-off: the number of spurious extra taxa increases. Considering all those results together, Centrifuge and Kraken2 are likely the better choice among the four tools to increase the likelihood of recovering ancient viruses.

One caveat with Centrifuge is that it outputs the largest numbers of spurious extra taxa across when the species tested is not present in its database (see Figure 8 and Supplementary Note 2) or when the reads are highly deaminated. Hence, if the list of candidates is very wide (probably due to the extra taxa) and option to optimise computational resources for downstream analyses would be to focus on the candidates with the highest number of assignments (as we saw in the human simulations, an important proportion of extra taxa has less than 10 assigned reads). For studies with high deamination, DIAMOND could be a good choice to reduce the number of candidates for further downstream analyses (as long as the reads have the appropriate length) as its sensitivity and precision are stable (Figure 6). We also recommend DIAMOND for fast screenings given its high Precision_s&h and as its database is smaller than Kraken2's or Centrifuge's and is easier to build (see Table 1 and Materials and Methods).

MetaPhlAn2 did not perform well in our analyses. While MetaPhlAn2 has advantages, such as having stable sensitivity and precision values in different conditions (Figures 5-7), the large number of viruses that would be missed in ancient virome studies suggest it is not well-suited for such analyses. However, its small database makes this tool convenient if researchers are after specific viruses, as long as the virus of interest has a marker gene in the database or the researchers define its marker genes and add them to the database following the developer's instructions.

For human studies, our simulations suggest that, even when the human genome is included in the database, cleaning as many human reads as possible would be necessary prior to classification to minimize the number of false positives. This would come at the expense of losing true candidates. However, we found hundreds of detected spurious extra taxa across different kingdoms within the viruses, and this could lead to false claims. To avoid having to handle so many false positives, our suggestion would be to map the data to the human genome and remove all reads that get assigned coordinates including those with low mapping qualities prior to classification using e.g. Sunbeam (Clarke et al. 2019).

The results presented here are focused on DNA viruses that infect humans. However, we expect they likely hold for viruses from other hosts as well. The reasons for the difficulty to classify some viruses for specific classifiers are mainly the specific databases used and the viral taxonomy of the tested viruses (see Supplementary Note 4). Hence, if the intention is to answer a specific biological question, i.e. to identify a given set of viruses with any of the classifiers tested here, we recommend to first make sure the viruses of interest are present in the classifier's database. To help users in that regard, we have

also included Table S2, that lists (for 60 bp) the raw number of reads per taxon for each classifier for all human DNA viruses included in this study.

To conclude, the four classifiers that we compared performed remarkably well considering they were not specifically developed to handle aDNA data and are robust to ancient DNA deamination damage. In our simulations, Centrifuge and Kraken2 outperforms the other classifiers as Centrifuge can handle short fragments, and both of them have high sensitivity and precision values across read lengths. Moreover, for human studies, human ancient DNA contamination could lead to a very large number of false positives if human reads are not filtered properly. Finally, an area of future research would be to adapt the databases and default parameters of the classifiers to improve their precision and sensitivity values, and to increase the number of identified viruses, especially when handling short reads.

Acknowledgments

We would like to thank the former Vital-IT platform of the SIB Swiss Institute of Bioinformatics (SIB) and the DCSR for maintaining the computer infrastructure at UNIL. We thank Barbara Mühlemann, Nicolas Rascovan, María C. Ávila-Arcos, Diana Ivette Cruz Dávalos, Christophe Dessimoz and Christina Warinner for helpful discussions and Florian Clemente for comments on earlier versions of this manuscript. The work was supported by the Swiss National Science Foundation (SFNS) and a European Research Council (ERC) grant to ASM.

References

Allentoft ME, Collins M, Harker D, Haile J, Oskam CL, Hale ML, Campos PF, Samaniego JA, Gilbert MTP, Willerslev E, et al. 2012. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. B Biol. Sci.* 279:4724–4733.

Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L, et al. 2015. Population genomics of Bronze Age Eurasia. *Nature* 522:167–172.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

Anon. 2019. The Integrative Human Microbiome Project. *Nature* 569:641.

Anon. Entrez Programming Utilities Help. :173.

Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, et al. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci.* 104:14616–14621.

Brister JR, Ako-adjei D, Bao Y, Blinkova O. 2015. NCBI Viral Genomes Resource. *Nucleic Acids Res.* 43:D571–D577.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12:59–60.

Burrows M, Wheeler DJ. 1994. A block-sorting lossless data compression algorithm.

Carøe C, Gopalakrishnan S, Vinner L, Mak SST, Sinding MHS, Samaniego JA, Wales N, Sicheritz-Pontén T, Gilbert MTP. 2018. Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* 9:410–419.

Clarke EL, Taylor LJ, Zhao C, Connell A, Lee J-J, Fett B, Bushman FD, Bittinger K. 2019. Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. *Microbiome* 7:46.

Dabney J, Meyer M, Pääbo S. 2013. Ancient DNA Damage. *Cold Spring Harb. Perspect. Biol.* 5:a012567.

Duggan AT, Perdomo MF, Piombino-Mascali D, Marciniak S, Poinar D, Emery MV, Buchmann JP, Duchêne S, Jankauskas R, Humphreys M, et al. 2016. 17th Century Variola Virus Reveals the Recent History of Smallpox. *Curr. Biol.* 26:3407–3412.

Eisenhofer R, Weyrich LS. 2019. Assessing alignment-based taxonomic classification of ancient microbial DNA. *PeerJ* 7:e6594.

Ferragina P, Manzini G. 2000. Opportunistic data structures with applications. In: Proceedings 41st Annual Symposium on Foundations of Computer Science. p. 390–398.

Green RE, Malaspinas A-S, Krause J, Briggs AW, Johnson PLF, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, et al. 2008. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* 134:416–426.

Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator. *Bioinformatics* 28:593–594.

Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome Res.* 17:377–386.

Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* [Internet]. Available from:

http://genome.cshlp.org/content/early/2016/11/16/gr.210641.116
Krause-Kyora B, Susat J, Key FM, Kühnert D, Bosse E, Immel A, Rinne C, Kornell S-C, Yepes D, Franzenburg S, et al. 2018. Neolithic and medieval virus genomes reveal complex evolution of hepatitis B. *eLife* 7:e36666.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9:357–359.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25:1754–1760.
Li W-H. 1997. Molecular Evolution. Sunderland, Mass
Malaspinas A-S, Lao O, Schroeder H, Rasmussen M, Raghavan M, Moltke I, Campos PF, Sagredo FS, Rasmussen S, Gonçalves VF, et al. 2014. Two ancient human genomes reveal Polynesian ancestry among the indigenous Botocudos of Brazil. *Curr. Biol.* 24:R1035–R1037.
Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, Hingamp P, Goto S, Ogata H. 2016. Linking Virus Genomes with Host Taxonomy. *Viruses* 8:66.
Mühlemann B, Jones TC, Damgaard P de B, Allentoft ME, Shevnina I, Logvin A, Usmanova E, Panyushkina IP, Boldgiv B, Bazartseren T, et al. 2018. Ancient hepatitis B viruses from the Bronze Age to the Medieval period. *Nature* 557:418–423.
Mühlemann B, Margaryan A, Damgaard P de B, Allentoft ME, Vinner L, Hansen AJ, Weber A, Bazaliiskii VI, Molak M, Arneborg J, et al. 2018. Ancient human parvovirus B19 in Eurasia reveals its long-term association with humans. *Proc. Natl. Acad. Sci.* 115:7557–7562.
Mühlemann B, Vinner L, Margaryan A, Wilhelmson H, Castro C de la F, Allentoft ME, Damgaard P de B, Hansen AJ, Nielsen SH, Strand LM, et al. 2020. Diverse variola virus (smallpox) strains were widespread in northern Europe in the Viking Age. *Science* [Internet] 369. Available from: https://science.sciencemag.org/content/369/6502/eaaw8977
O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44:D733–D745.
Pérez-Brocal V, Moya A. 2018. The analysis of the oral DNA virome reveals which viruses are widespread and rare among healthy young adults in Valencia (Spain). *PLOS ONE* 13:e0191867.
Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, Mayer G. 2018. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* 8:10950.
Rascovan N, Sjögren K-G, Kristiansen K, Nielsen R, Willerslev E, Desnues C, Rasmussen S. 2019.

Emergence and Spread of Basal Lineages of Yersinia pestis during the Neolithic Decline. *Cell* 176:295-305.e10.
Renaud G, Hanghøj K, Willerslev E, Orlando L. 2017. gargammel: a sequence simulator for ancient DNA. *Bioinformatics* 33:577–579.
Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. 2012. Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. *PLOS ONE* 7:e34131.
Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 9:811–814.
Siqueira JD, Dominguez-Bello MG, Contreras M, Lander O, Caballero-Arias H, Xutao D, Noya-Alarcon O, Delwart E. 2018. Complex virome in feces from Amerindian children in isolated Amazonian villages. *Nat. Commun.* 9:4270.
Stern J, Miller G, Li X, Saxena D. 2019. Virome and bacteriome: two sides of the same coin. *Curr. Opin. Virol.* 37:37–43.
Taubenberger JK, Reid AH, Lourens RM, Wang R, Jin G, Fanning TG. 2005. Characterization of the 1918 influenza virus polymerase genes. *Nature* 437:889.
Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* 12:902–903.
Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J. 2001. Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr. Biol.* 11:1531–1535.
Vågene ÅJ, Herbig A, Campana MG, García NMR, Warinner C, Sabin S, Spyrou MA, Valtueña AA, Huson D, Tuross N, et al. 2018. Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat. Ecol. Evol.* 2:520.
Velsko IM, Frantz LAF, Herbig A, Larson G, Warinner C. 2018. Selection of Appropriate Metagenome Taxonomic Classifiers for Ancient Microbiome Research. *mSystems* [Internet] 3. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6050634/
Warinner C, Herbig A, Mann A, Fellows Yates JA, Weiß CL, Burbano HA, Orlando L, Krause J. 2017. A Robust Framework for Microbial Archaeology. *Annu. Rev. Genomics Hum. Genet.* 18:321–356.
Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20:257.
Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46.

Worobey M, Watts TD, McKay RA, Suchard MA, Granade T, Teuwen DE, Koblin BA, Heneine W, Lemey P, Jaffe HW. 2016. 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature* 539:98–101.

Table 1. Characteristics of the four classifiers.

Database, query algorithm and taxonomic binning strategy for each classifier. In the main text the results for MetaPhlAn2 and for Centrifuge, Kraken2 and DIAMOND with custom viral databases are shown. The results for Centrifuge and Kraken2's default databases as well as DIAMOND database with archaeal, bacterial, human and viral protein sequences are shown in Supplementary Note 2. Results for MALT are shown in Supplementary Note 1.

Classifier	Database					Query algorithm	Taxonomic binning strategy
	Organisms included		Loci	Molecule	Size		
Centrifuge	Default ("full db")	RefSeq: archaea, bacteria, viruses and human	Whole genomes	DNA	24G	Exact alignment (no mismatches or gaps)	Score that favours longer hits
	Custom ("viral db")	RefSeq: viruses			213M		
Kraken2	Default ("full db")	Refseq + UniVec_Core: archaea, bacteria, viruses, human and (vectors, adapters, linkers and primers)	Whole genomes	DNA	89G	Exact k-mer matching	Highest number of k-mer matches considering a root-to-leaf path
	Custom ("viral db")	RefSeq: viruses			34G		
DIAMOND	Custom ("full db")	RefSeq: archaea, bacteria, viruses, human and viruses	Protein coding regions	Proteins	72G	Alignment using spaced seeds	Lowest common ancestor
	Custom ("viral db")	RefSeq: viruses			117M		
MetaPhlAn2	Default	Archaea, bacteria, viruses and eukaryotic microbes	Clade specific markers	DNA	1.2G	Alignment with BowTie2: ungapped seed match followed by extension	Clade specific marker
MALT	Custom ("viral db")	RefSeq: viruses	Whole genomes	DNA	7.3G	Banded alignment with spaced seeds	Lowest common ancestor

Figure 1. Visual representation of the different classification categories defined in this study (“correct species”, “correct higher”, “incorrect” and “unclassified”).

The *Variola virus* is used here as an example to illustrate the simulation steps and the four classification categories used to summarize the simulation results. Three schematic steps are shown to summarize the simulations: (1) The random selection of a virus (here the *Variola virus*). (2) The simulation of reads from the viral sequence: the selected virus is used as reference genome to simulate sequencing reads with ART (Huang et al. 2012). (3a) The classification of the reads: all reads are then classified with Centrifuge, Kraken2, DIAMOND and MetaPhlan2 (Buchfink et al. 2015; Truong et al. 2015; Kim et al. 2016; Wood et al. 2019: 2) and then assigned to one of the four categories (3b): i) “correct species”: reads correctly classified as the virus of interest (dark green), ii) “correct higher”: reads classified as higher taxonomic ranks included in the lineage of the virus of interest (light green), iii) “incorrect”: reads classified as a taxon not included in the lineage of the virus of interest, d) “unclassified”: reads not classified.

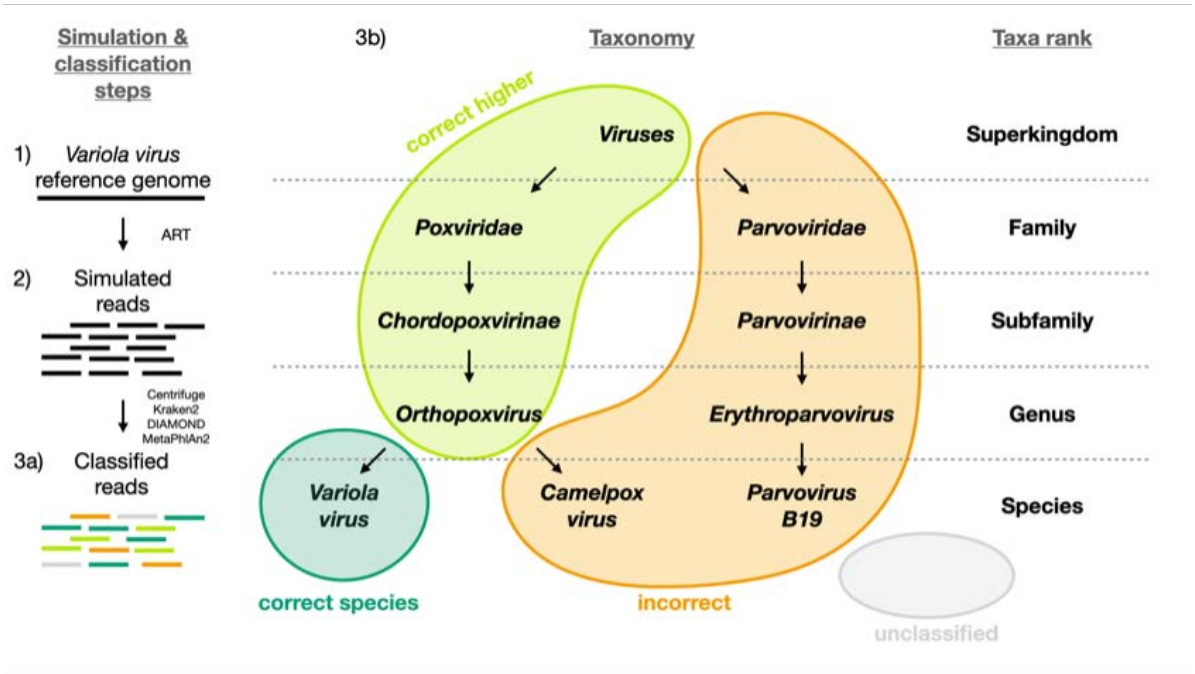


Figure 2. Observed deamination damage across the reads.

Observed frequency of substitutions across the reads simulated from the viral sequence NC_006273.2 (Human betaherpes virus 5) for different probabilities of single stranded deamination ranging from 0 to 0.5 (the range used in the simulations). The nick frequency was set at 0.03, the geometric parameter for the average length of overhanging ends at 0.25, and the probability of deamination in the double-stranded portions of DNA at 0.01. The frequency of substitutions is shown on the y-axis; the distance from the read termini is shown on the x-axis. C to T substitutions are depicted in red; G to A substitutions are depicted in blue; in grey all other substitutions. Plots on the left represent the 5' end of the reads, plots on the right represent the 3' end. Note that similar patterns have been observed in real data (Briggs et al. 2007; Carøe et al. 2018).

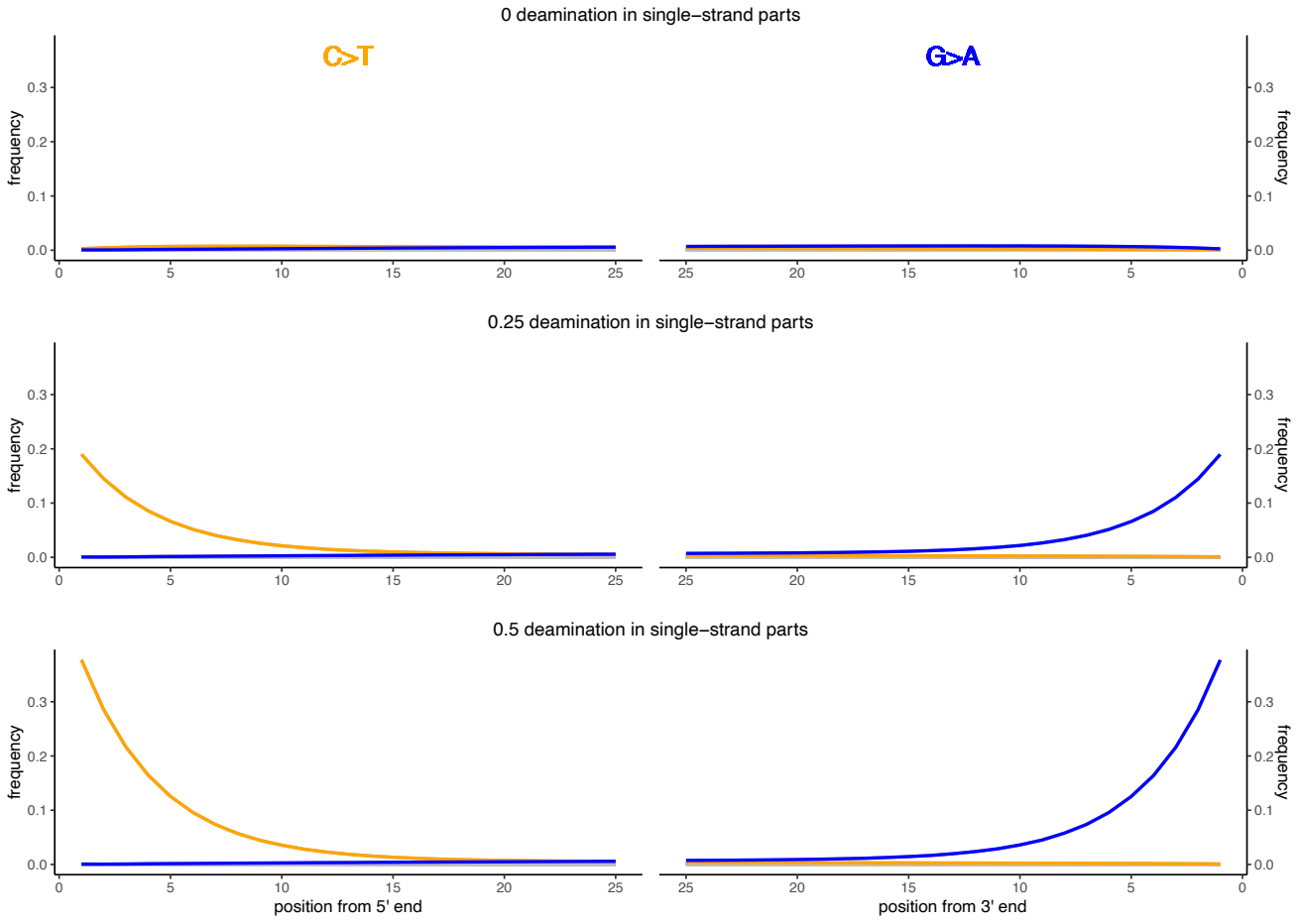


Figure 3. Average number of substitutions for the deamination and sequencing error simulations.

Observed average number of substitutions for different simulation parameter values for the deamination and sequencing error simulations. Shown here are the observed values for the case of NC_006273.2 (*Human betaherpes virus 5*) with 39,270 simulated reads of length 60 bp. For the deamination simulations, the single-stranded probability (see Figure 2) of the deamSim gargamel subprogram (Renaud et al. 2017) is increased from 0 to 0.5. Note that deamination takes place with a probability of 0.01 across the read for all the deamination simulations so that there are additional errors even with a single stranded probability set at 0. For the sequencing error simulations, the parameter *qShift* of ART was decreased from 0 to -9 which corresponds to a 1 fold (*qShift* of 0) to 7.9 fold (*qShift* of -9) increase compared to standard Illumina 2500 sequencing (Huang et al. 2012).

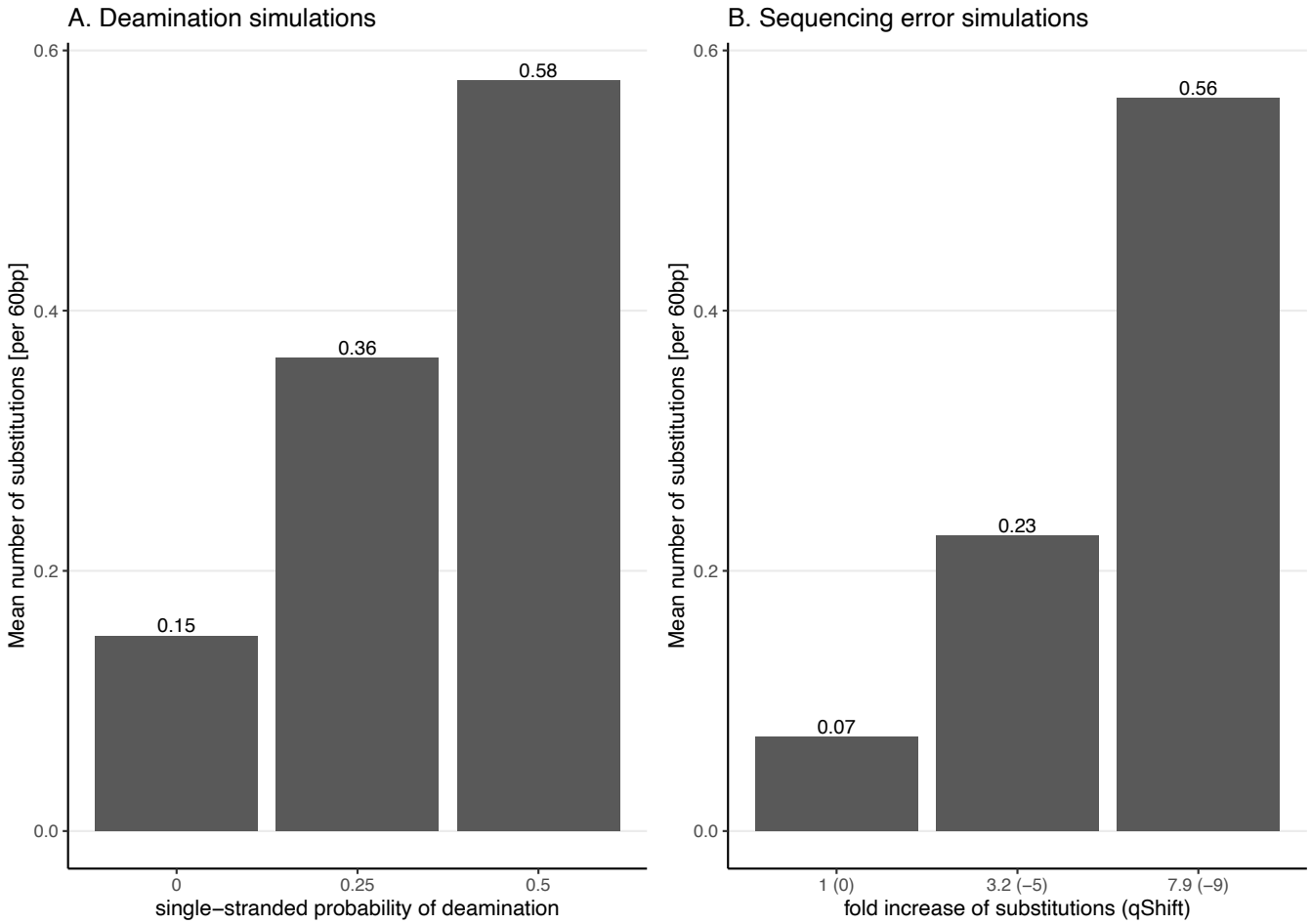


Figure 4. Classification results for the 60 bp read set.

A) Mean sensitivity versus mean precision. The mean sensitivities (Sensitivity_s & Sensitivity_s&h) are the means of the proportions of reads correctly classified over the total number of simulated reads across viruses. The mean precisions (Precision_s & Precision_s&h) are the means of the proportions of reads correctly classified over the number of classified reads across viruses. Circles denote the values if only "correct species" reads are considered as correctly classified reads; triangles denote the values if "correct species" and "correct higher" reads are considered as correctly classified reads (see Materials and Methods). The perfect classifier would have 100% sensitivity and 100% precision. B) Total number of viruses recovered for each classifier when correctly identifying at least 1 read per virus. The dashed line indicates the total number of tested viruses (233). C) Mean number of spurious extra taxa per classifier. In this plot, a taxon is assumed as identified by a classifier if at least 1 read is assigned to it.

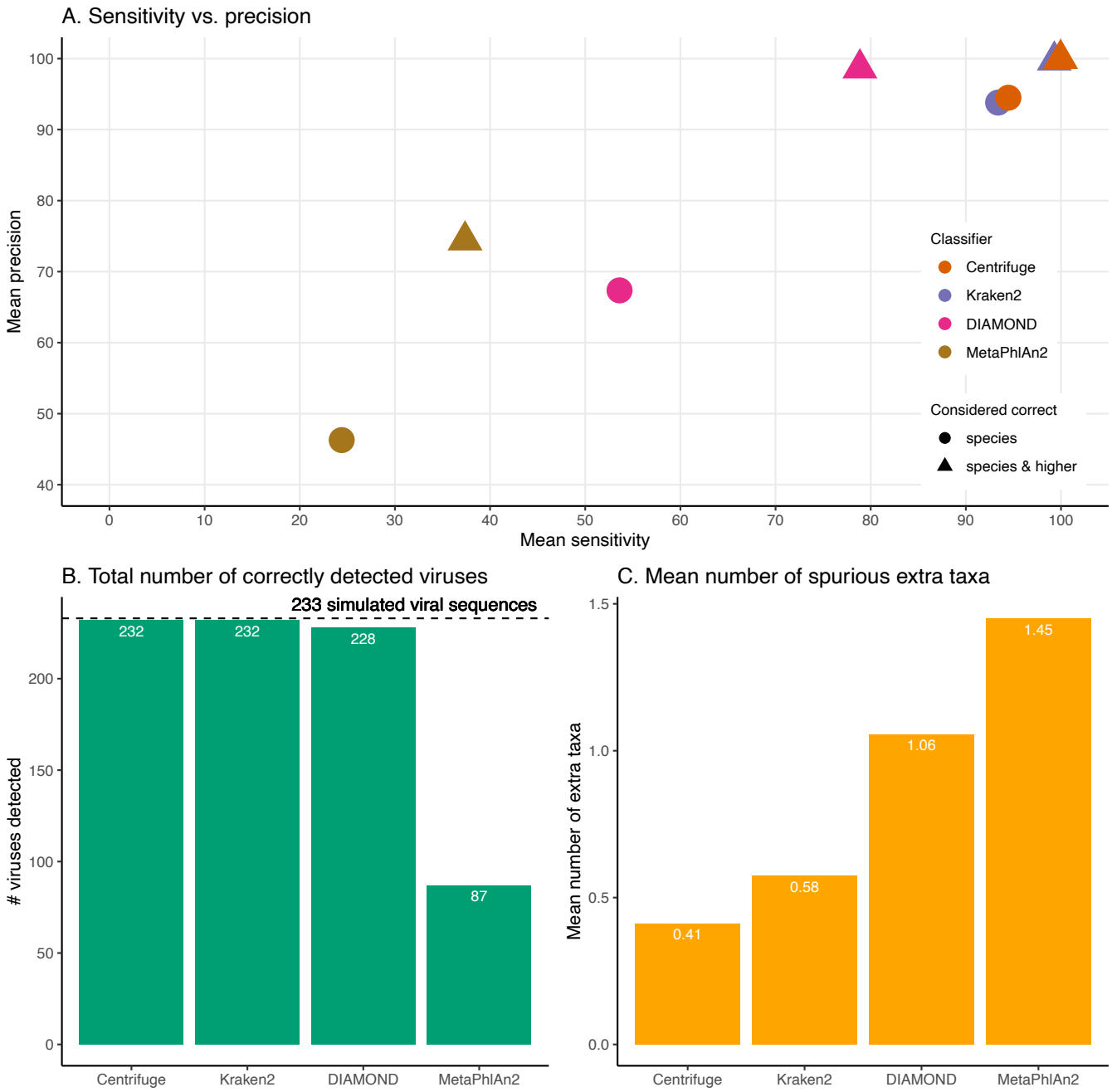

A. Sensitivity vs. precision
Mean precision
Mean sensitivity
Classifier
Centrifuge
Kraken2
DIAMOND
MetaPhlAn2
Considered correct
species
species & higher
B. Total number of correctly detected viruses
233 simulated viral sequences
viruses detected
232
232
228
87
Centrifuge
Kraken2
DIAMOND
MetaPhlAn2
C. Mean number of spurious extra taxa
Mean number of extra taxa
0.41
0.58
1.06
1.45
Centrifuge
Kraken2
DIAMOND
MetaPhlAn2

Figure 5. Effect of the read length on the classification performance.

For these simulations, read length was varied from 30 to 150 bp. A) Average Sensitivity_s (continuous lines) and Sensitivity_s&h (dashed lines) for each classifier. B) Average Precision_s (continuous lines) and Precision_s&h (dashed lines) for each classifier. C) Total number of viruses detected out of the 233 tested. The dashed line shows the maximum number of detectable viruses. D) Average number of spurious extra taxa across simulated viral sequences. The vertical dashed line indicates the initial 60 bp read set.

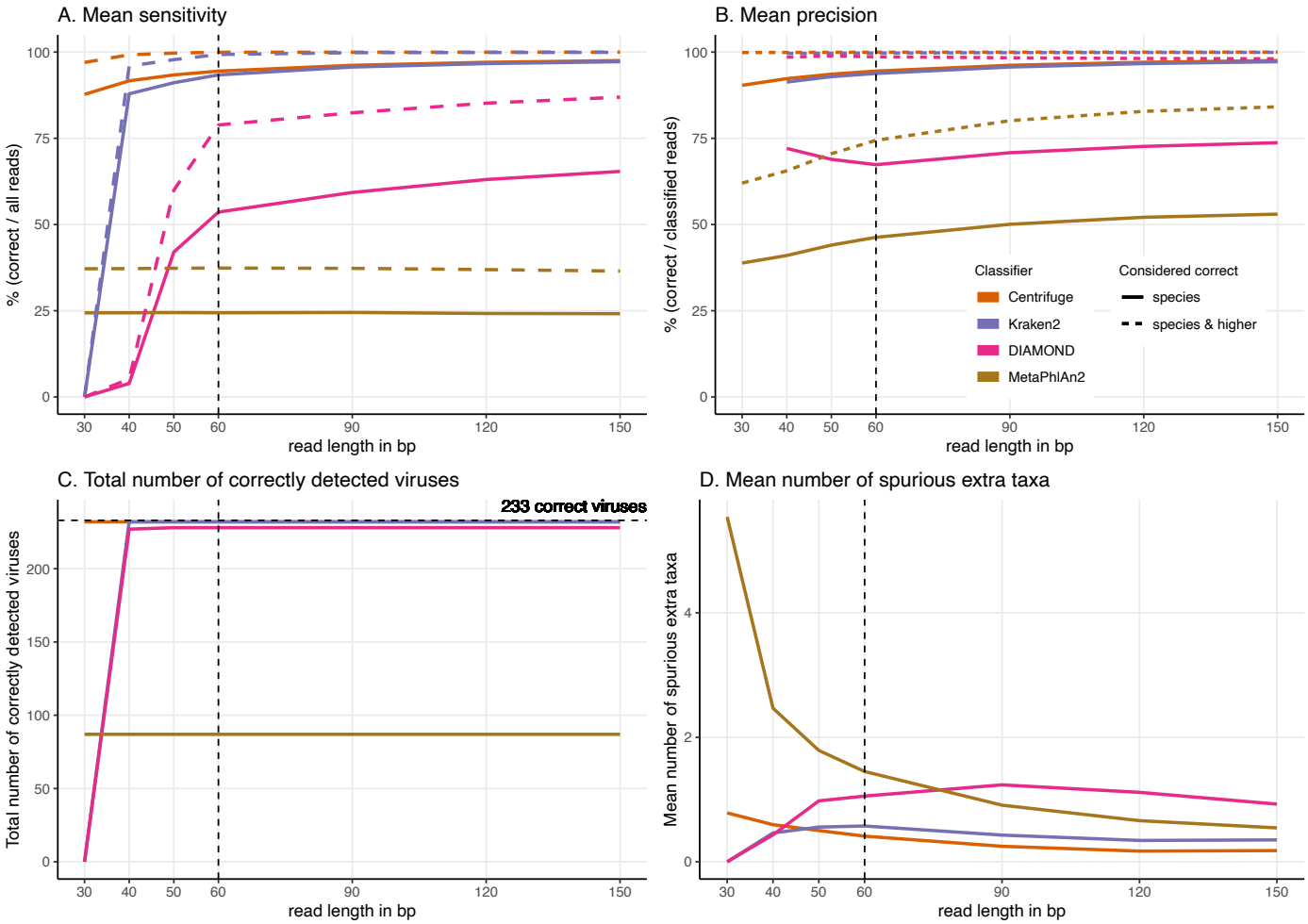


Figure 6. Effect of the deamination damage on the classification performance.

For these simulations, errors were added using deamSim gargamel subprogram which assumes an ancient DNA deamination-like distribution and were added in addition to the ART Illumina like sequencing errors. The results shown here correspond to a single-stranded probability of deamination varying from 0 to 0.5. For all the results the nick frequency is set at 0.03, the average length of overhanging ends is set at 0.25, and the probability of deamination in the double-stranded portions of DNA is set at 0.01. A) Average Sensitivity_s (continuous lines) and Sensitivity_s&h (dashed lines) for each classifier. B) Average Precision_s (continuous lines) and Precision_s&h (dashed lines) for each classifier. C) Total number of viruses detected out of the 233 tested. The dashed line shows the maximum number of detectable viruses. D) Average number of spurious extra taxa across simulated viruses.

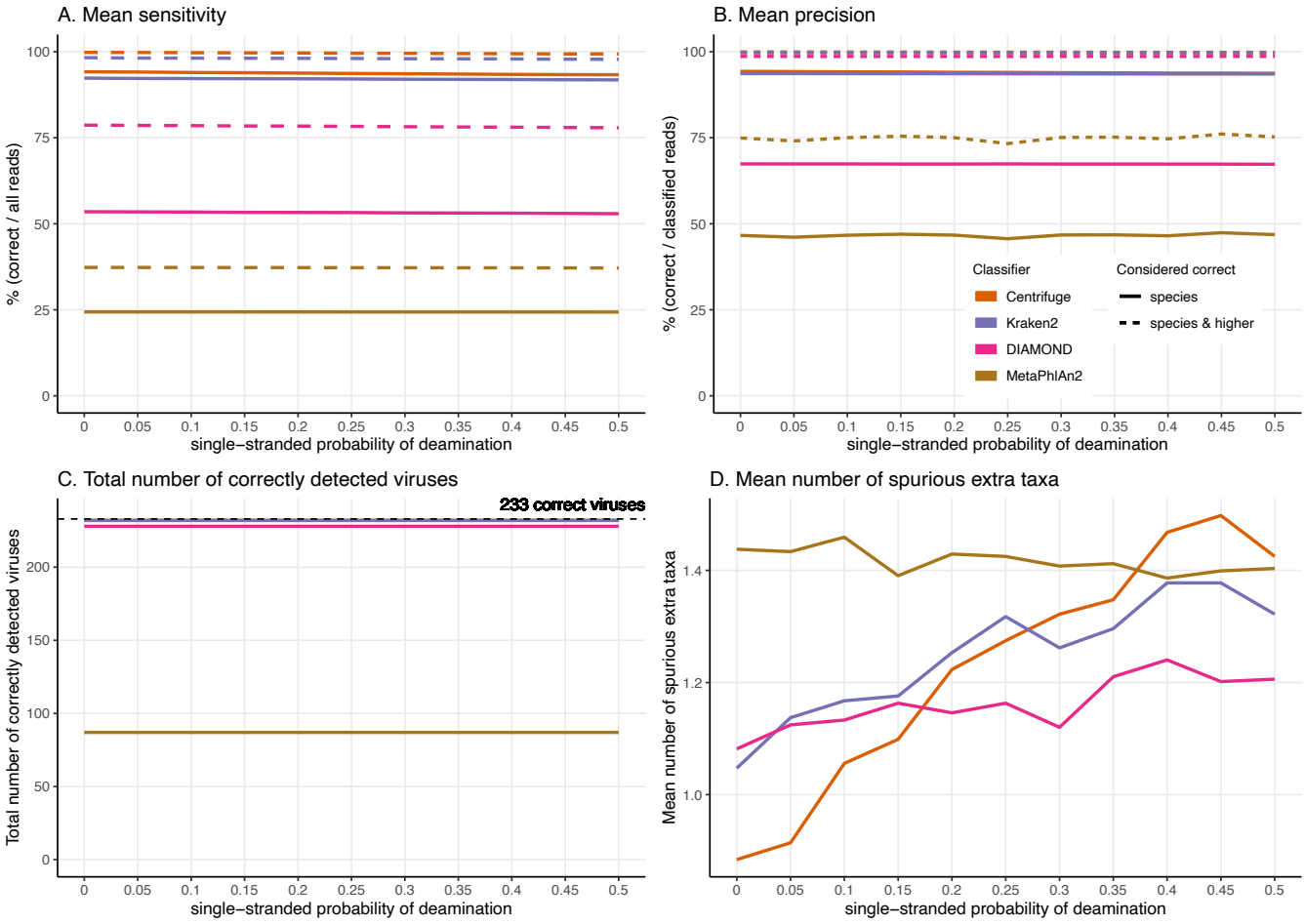


Figure 7. Effect of the substitution sequencing error on the classification performance.

For these simulations, errors were added using ART which assumes a profile similar to the ones observed for Illumina Sequencing machines (HiSeq 2500). The results shown here correspond to increasing the overall sequencing error rate ranging from 1 to 7.9-fold (qShift values from 0 to -9). On the x-axis, the first number correspond to the expected fold increase in error rate while the parameter that was varied, *qShift*, is shown in parenthesis. A) Average Sensitivity_s (continuous lines) and Sensitivity_s&h (dashed lines) for each classifier. B) Average Precision_s (continuous lines) and Precision_s&h (dashed lines) for each classifier. C) Total number of viruses detected out of the 233 tested. The dashed line shows the maximum number of detectable viruses. D) Average number of spurious extra taxa across simulated viral sequences. The vertical dashed line indicates the initial 60 bp read set.

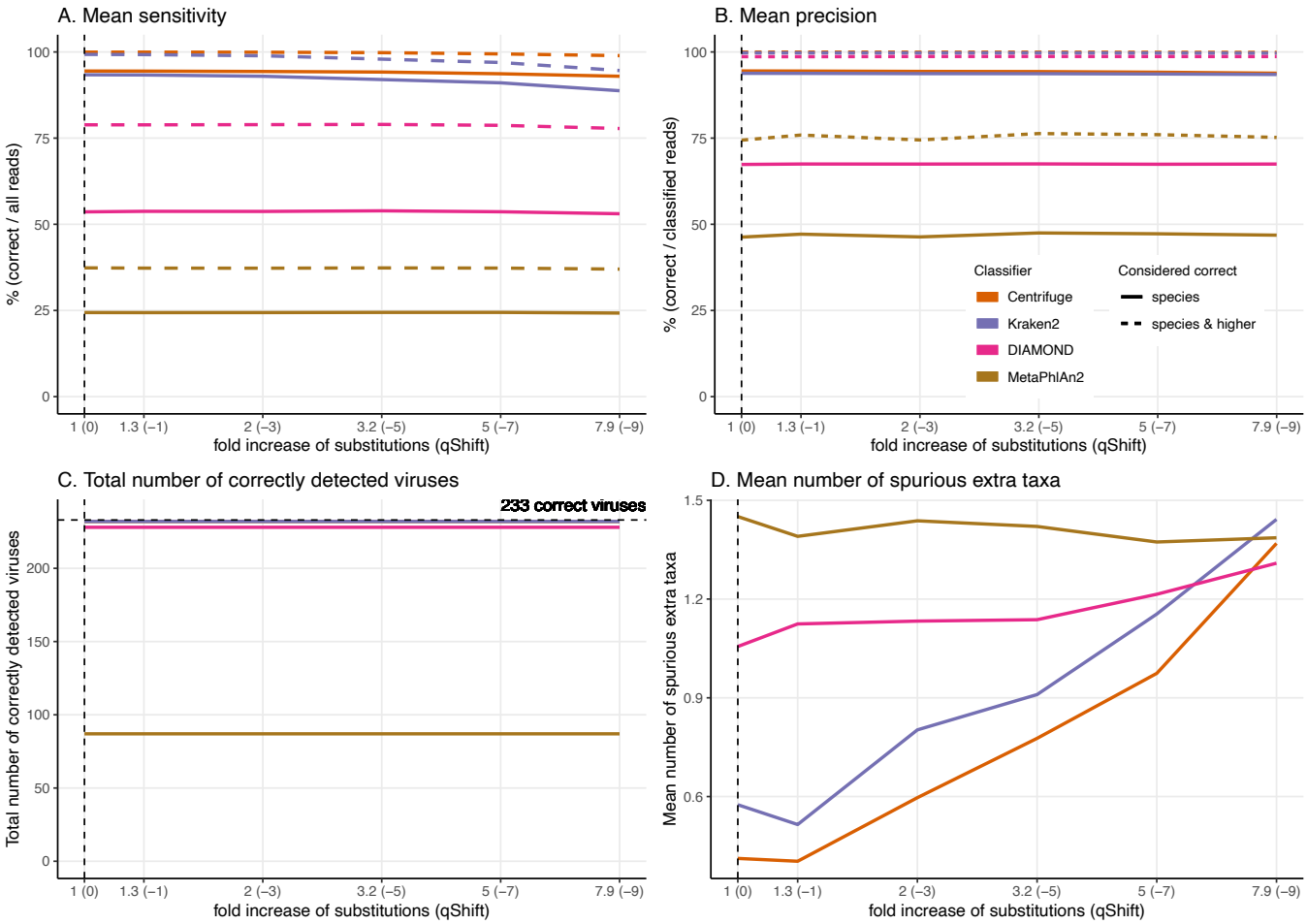


Figure 8. Classification of human reads.

A) Proportion of reads for each classification category (see Figure 1). Note that none of the databases used include human (see Table 1). B) Total number of incorrectly classified reads. C) Number of spurious extra taxa reported by each classifier.

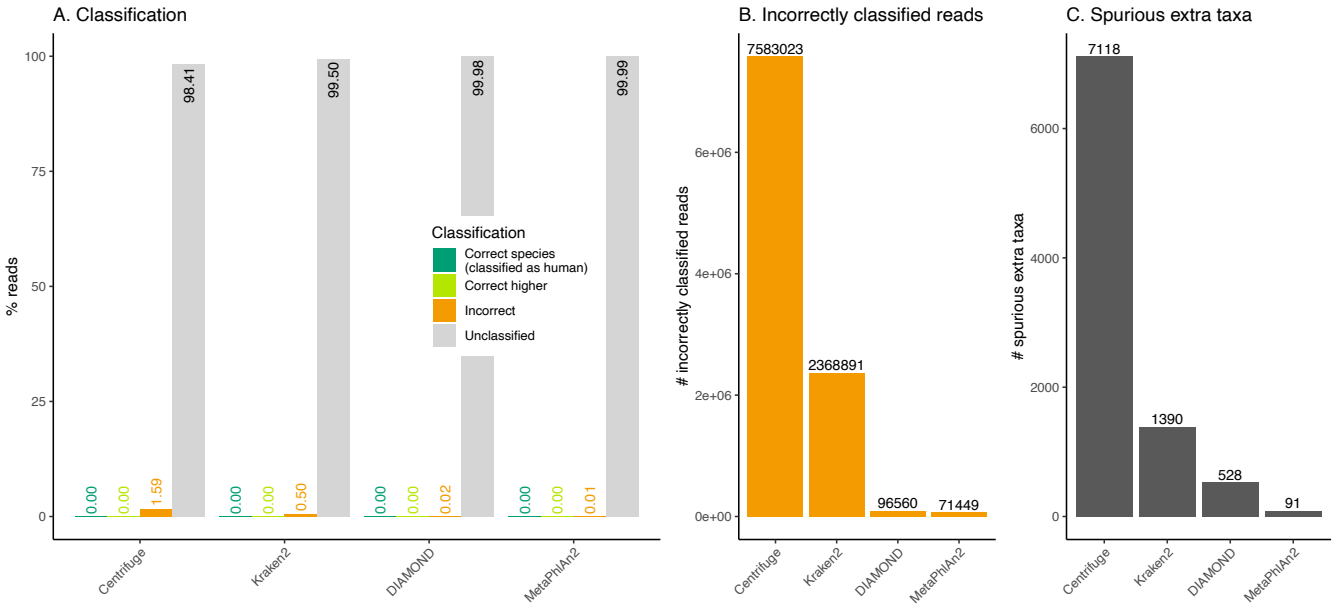


Figure 9. Taxonomic ranks represented in the classification of human reads.

Number of reads and the proportion of reads per classifier (bar plots). Proportion of taxa classified as a particular pair of superkingdom-kingdom (pie charts).

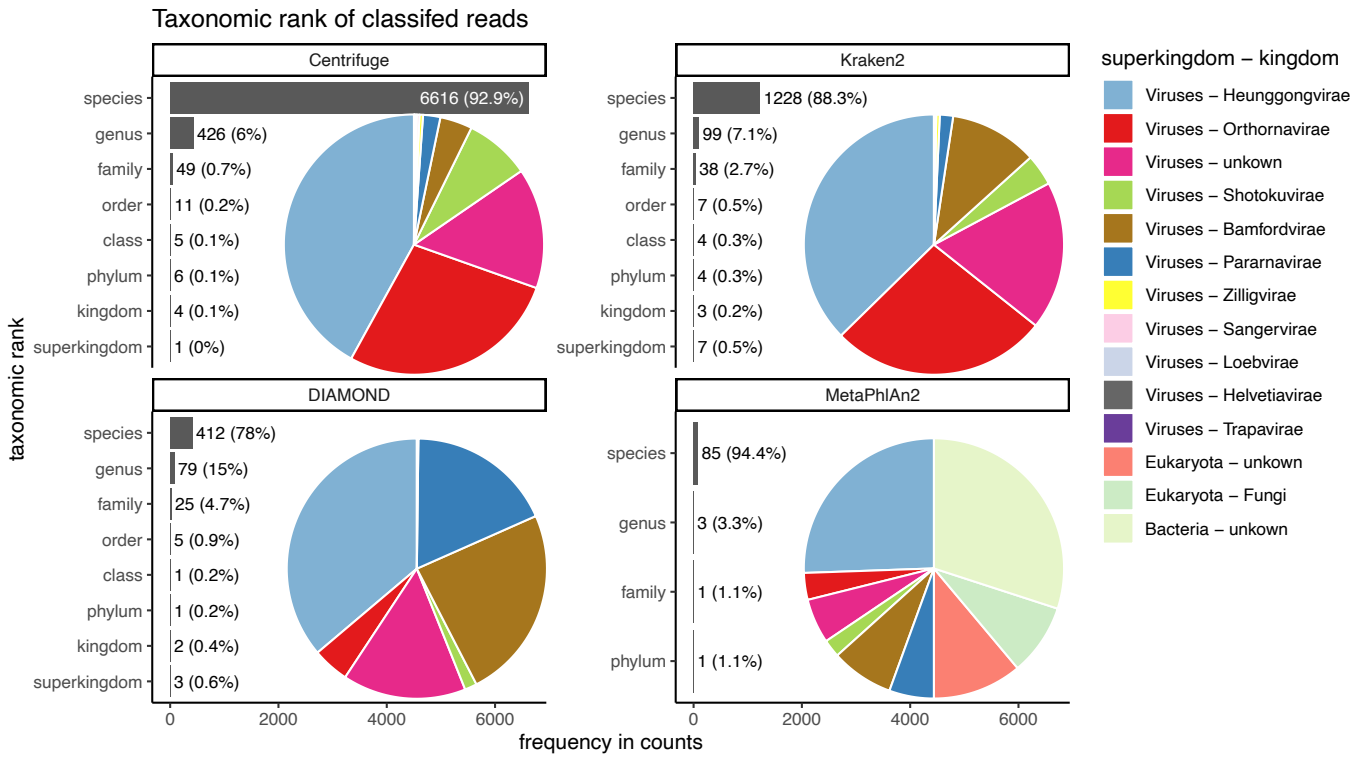


Figure S1. Classification results for the 60 bp read set.

A) Percentage of the reads classified in each of the four categories: “correct species”, “correct higher”, “incorrect” and “unclassified” (see Figure 1) per classifier for each viral sequence. Each bar corresponds to one of the 233 viruses selected for the simulations. Note that the viruses are not ordered the same way for each subplot. B) Means over the 233 viruses for each classification category.

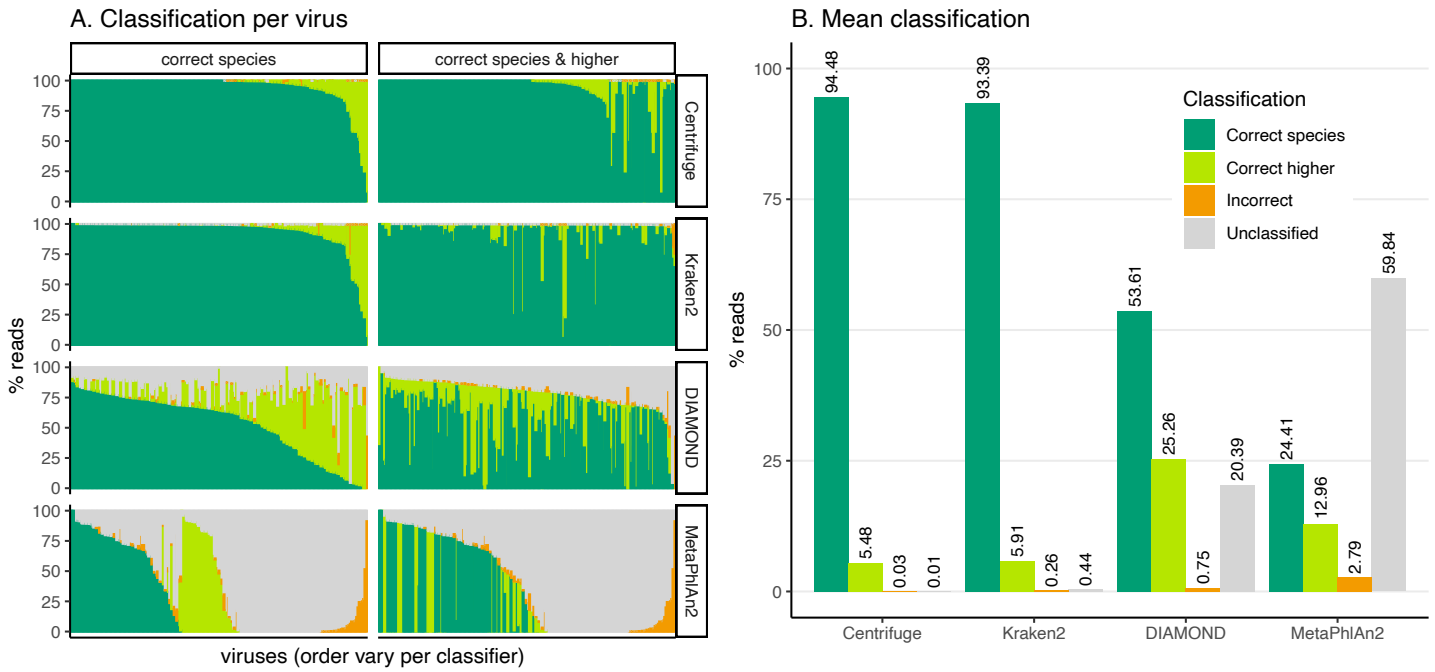


Figure S2. Average classification categories for all simulations (stacked bar plots).

Nine subplots showing the four classification categories (average across viral sequences) for Centrifuge (first row), Kraken2 (second row), DIAMOND (third row) and MetaPhlAn2 (fourth row) for all simulations; varying read length (first column), singe-stranded probability of deamination (second column), fold increase of substitution sequencing error (third column).

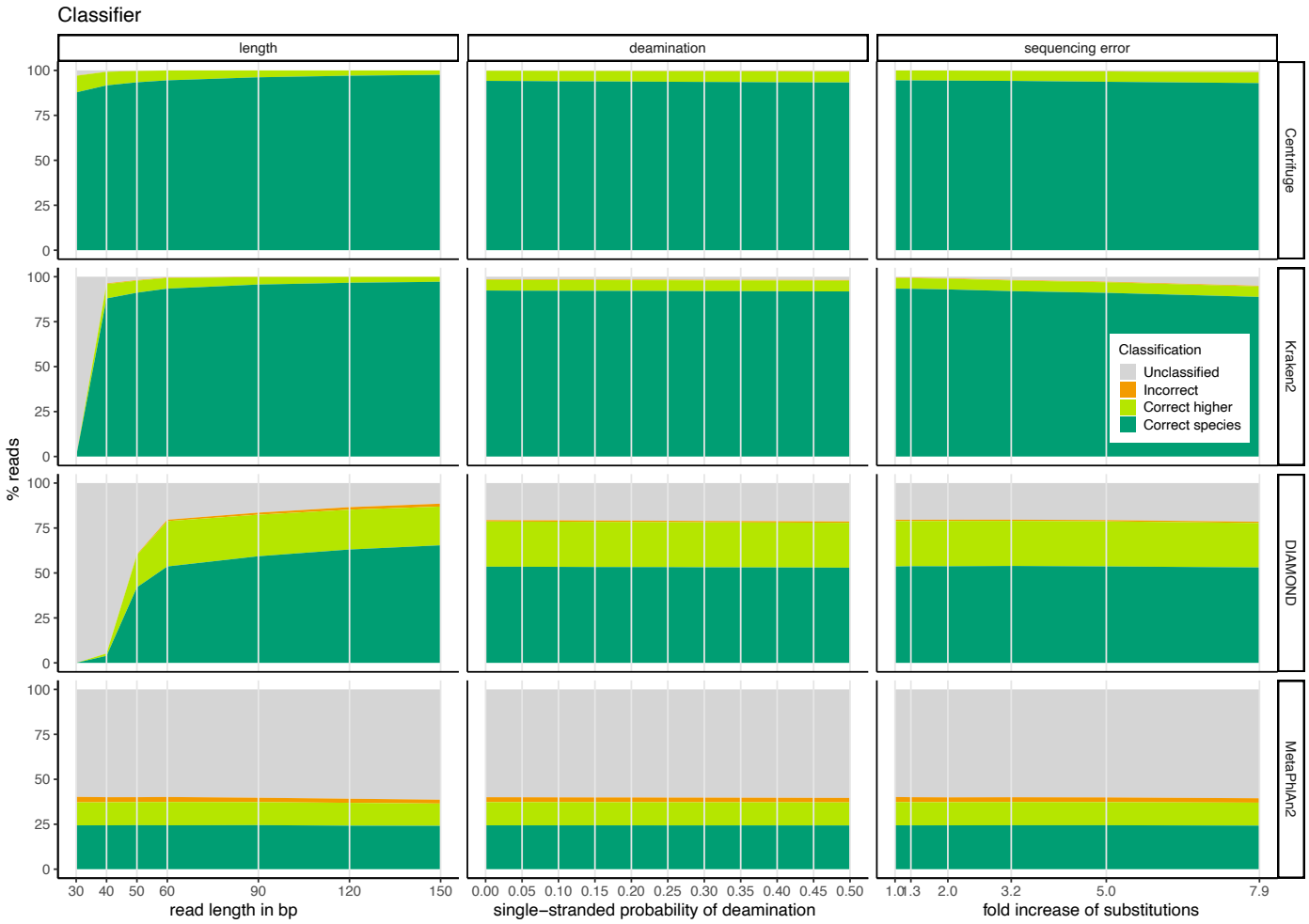
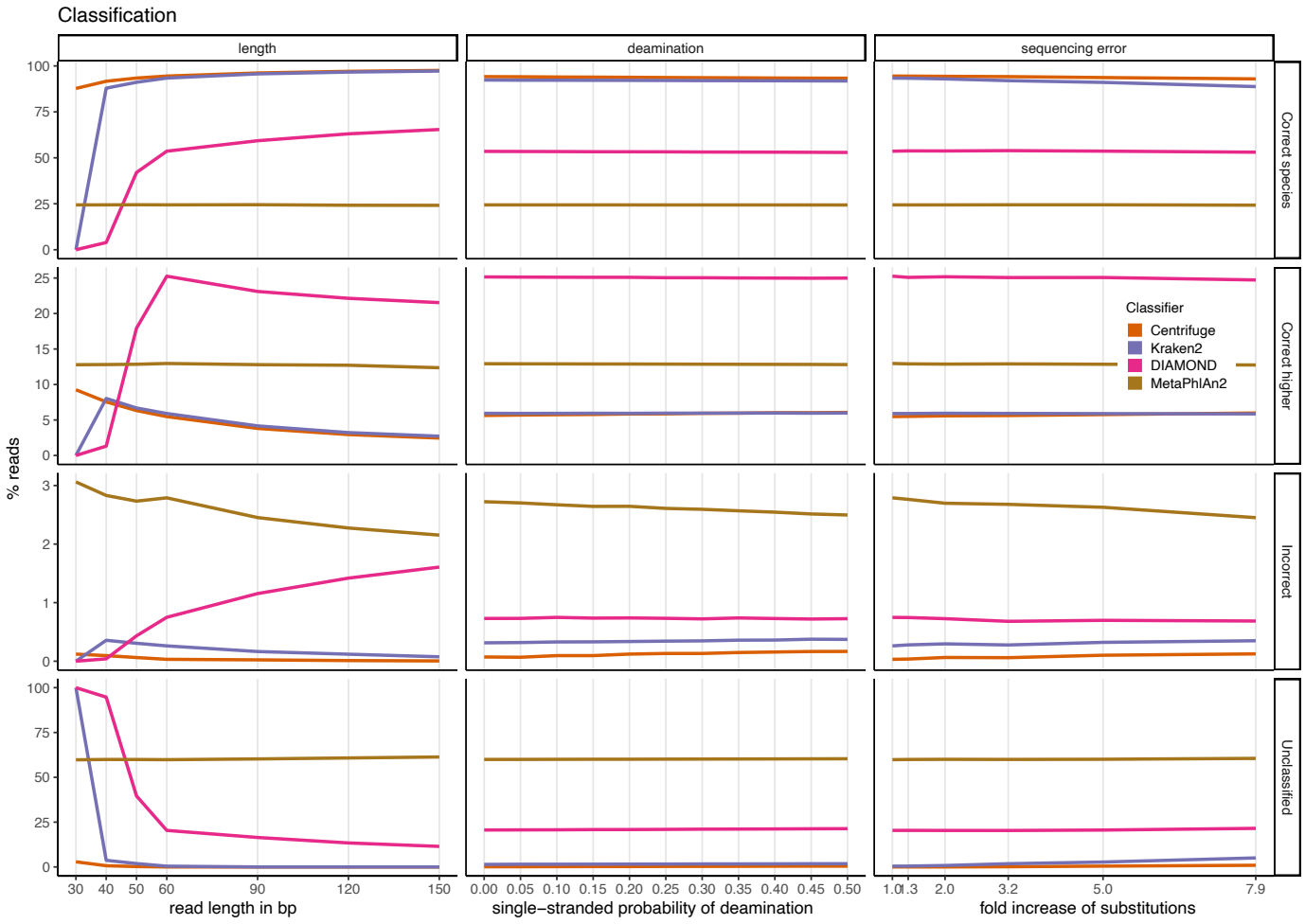


Figure S3. Average classification categories for all simulations (line plots).

Twelve subplots showing the four classification categories (average across simulated viral sequences) for the categories "correct species" (first row), "correct higher" (second row), "incorrect" (third row) and "unclassified" (fourth row) for all simulations; varying read length (first column), singe-stranded probability of deamination (second column), fold increase of substitution sequencing error (third column).



Supplementary Note 1: MALT

All the sets of simulated reads mentioned in the main text (with different lengths, deamination and sequencing error values; see Materials and Methods) were also classified with the software MALT (Vågene et al. 2018). In the following supplementary note we present our findings for this classifier.

Methodology

Classification
MALT version 0.5.3 (Vågene et al. 2018) was used in the nucleotide search mode (-m BlastN) as suggested by (Eisenhofer et al. 2019) for ancient DNA classifications.

Databases
As for Kraken2, Centrifuge and DIAMOND (see Supplementary Note 2), we tried to build two databases for MALT: a "full" database including RefSeq sequences for archaea, bacteria, viruses and human; and a viral database only including all the viral sequences from RefSeq. However, the creation of the full database required more than 1 TB of memory, which exceeds our resources. As a result, we only built the viral database. We presume that this extensive memory usage will also cause a problem to other researchers, limiting the usage of exhaustive MALT databases.

Reference sequences
To build the viral database we used all viral sequences from RefSeq, namely viral.1.1.genomic.fna.gz, viral.2.1.genomic.fna.gz and viral.3.1.genomic.fna.gz (downloaded from ftp.ncbi.nlm.nih.gov/refseq/release/viral/ on 27 June 2021).

Accession to taxonomy translation
Following the manual (https://software-ab.informatik.uni-tuebingen.de/download/malt/manual.pdf), we searched for the accession to taxon translation file within the installation folder of MALT, but there was no such file. We therefore downloaded the accession to taxon translation file *nucl_wgs.accession2taxid.gz* from NCBI (downloaded from ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/ on 27 June 2021) and built the database with the following command:

```
malt-build -i foo.fna -s DNA --index dbName --acc2taxa nucl_wgs.accession2taxid.gz
```

While the database creation worked well (in the sense that there were no errors reported) and while the overall results looked reasonable (see Figure SN1-1, NCBI-acc2taxa), inconsistent results were observed. In summary (but see also below), in some cases, the alignment done by MALT is correct but the reported taxon is wrong.

We assumed that there was an issue within the accession number to taxon ID translation. We therefore exchanged the translation file to megan-nucl-Jan2021.db.zip, a translation file provided by the authors available on their website for MEGAN, a related software (https://software-ab.informatik.uni-tuebingen.de/download/megan6/welcome.html; file downloaded on 17 February 2021).

Again, the database creation worked well (there was no error or warning reported). The classification performance was also similar to the previous classifications (see Figure SN1-1, MEGAN-acc2taxa below). But again, we detected incorrect translations of the sequence accession number to the taxonomic ID (taxID).

Using the command line manual ("`malt-build --help`"), we realized that (even though the software ran without errors and warnings) the command line manual and MALT's pdf manual are not in line with each other and that the parameter "`acc2taxa`" we used to pass the translation file was likely deprecated. We repeated the analyses following the command line manual and used then the parameter "`mapDB`" to pass the translation file:

```
malt-build -i foo.fna -s DNA --index dbName --mapDB megan-nucl-Jan201.db
```

With this new database the classification performance (sensitivity and precision) increased substantially (Figure SN1-1, MEGAN-mapDB) and we chose to analyse those results in more details (see below). However, issues with the undetected viruses persisted presumably also due to accession to translation file.

Results

Classifications for 60 bp simulations

The 60 bp simulations results for three combinations of translation files and parameters mentioned in the methodology above are shown hereafter.

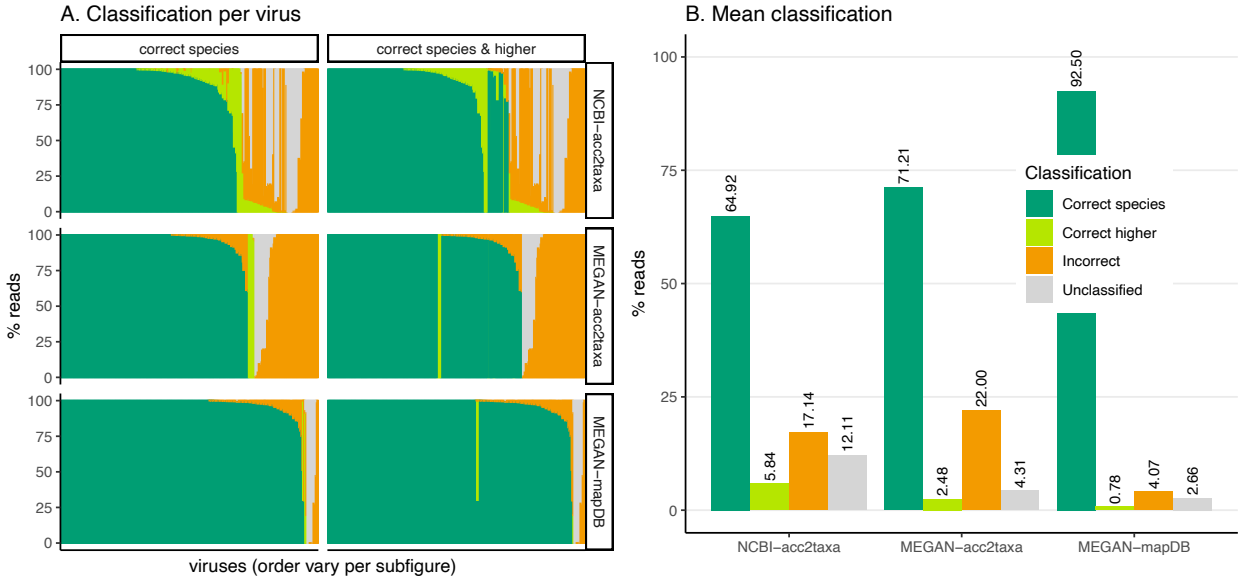


Figure SN1-1. A) Percentage of the reads classified in each of the four categories: "correct species", "correct higher", "incorrect" and "unclassified" (see Figure 1) per "translation file" for MALT. Each bar corresponds to one of the 233 viruses selected for the simulations. B) Means over the 233 viruses for each classification category. "NCBI-acc2taxa" corresponds to the run with nucl_wgs.accession2taxid.gz used as translation file; "MEGAN-acc2taxa" to the run with megan-nucl-Jan2021.db.zip used as translation file; and "MEGAN-mapDB" to the run with megan-nucl-Jan2021.db.zip used as translation file and the parameter mapDB to pass translation.

As can be seen in Figure SN1-1 A, most viruses are recovered and the number of correctly classified reads is quite high per virus, suggesting the software was run successfully. However, when we analysed

further the viruses that were not identified by MALT (see last section), we detected incorrect translations of the sequence accession number to the taxonomic ID (taxID). More specifically, there are instances where the alignment done by MALT is correct, but the reported taxon is wrong. This includes, surprisingly, classifications with reported taxa not present in the database. Note that although this issue is common to all our attempts, we compare MALT to the other four classifiers presented below for our best run, i.e. the case with the MEGAN translation file (`megan-nucl-Jan201.db`) and with the `--mapDB` parameter ("MEGAN-mapDB" in Figure SN1-1 A).

MEGAN accession to taxonomy translation file with command mapDB

For the 60 bp simulations, MALT shows sensitivity and precision values close to those for Centrifuge and Kraken2 and which outperform DIAMOND and MetaPhlAn2 (Figure SN1-2 A), both when considering correct species only and when including the higher taxa, with one exception: DIAMOND has a higher precision value when taking into account the higher taxa. A total of 221 viruses out of the 223 are identified correctly with MALT (less than Centrifuge, Kraken2 and DIAMOND; more than MetaPhlAn2) (Figure SN1-2 B). Among all the classifiers MALT reports the highest number of spurious extra taxa (Figure SN1-2 C).

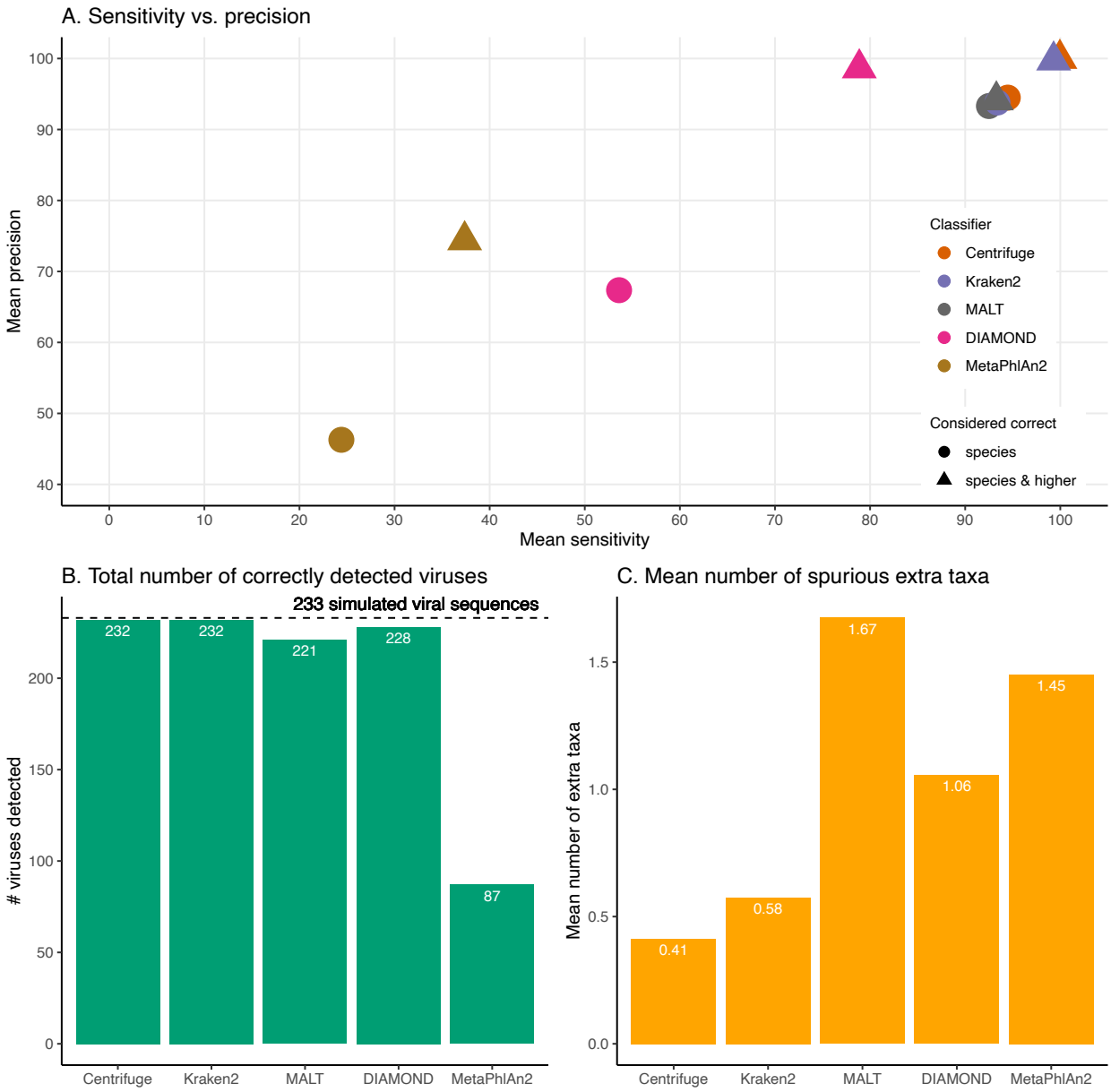


Figure SN1-2. A) Mean sensitivity versus mean precision. The mean sensitivities (Sensitivity_s & Sensitivity_s&h) are the means of the proportions of reads correctly classified over the total number of simulated reads across viruses. The mean precisions (Precision_s & Precision_s&h) are the means of the proportions of reads correctly classified over the number of classified reads across viruses. Circles denote the values if only "correct species" reads are considered as correctly classified reads; triangles denote the values if "correct species" and "correct higher" reads are considered as correctly classified reads (see Materials and Methods). The perfect classifier would have 100% sensitivity and 100% precision. B) Total number of viruses recovered for each classifier when correctly identifying at least 1 read per virus. The dashed line indicates the total number of tested viruses (233). C) Mean number of spurious extra taxa per classifier. In this plot, a taxon is assumed as identified by a classifier if at least 1 read is assigned to it.

When varying the read length MALT ranks mostly as the third classifier with the highest values of sensitivity and precision (Figure SN1-3 A & B). At 30 bp MALT outperforms Kraken2, and even Centrifuge when considering only species level; although Centrifuge remains the classifier with the best performance for the rest of the lengths (Figure SN1-3 A & B). DIAMOND outperforms MALT in precision when higher taxa are included (Figure SN1-3 B). The number of identified viruses for MALT

is the same across all the tested lengths (Figure SN1-3 C). For reads of length 30 bp and 40 bp MetaPhlAn2 reports the highest number of spurious extra taxa; MALT has the highest number for 50 bp, 60 bp and 150 bp; while DIAMOND has the highest number for 120 bp (Figure SN1-3 D).

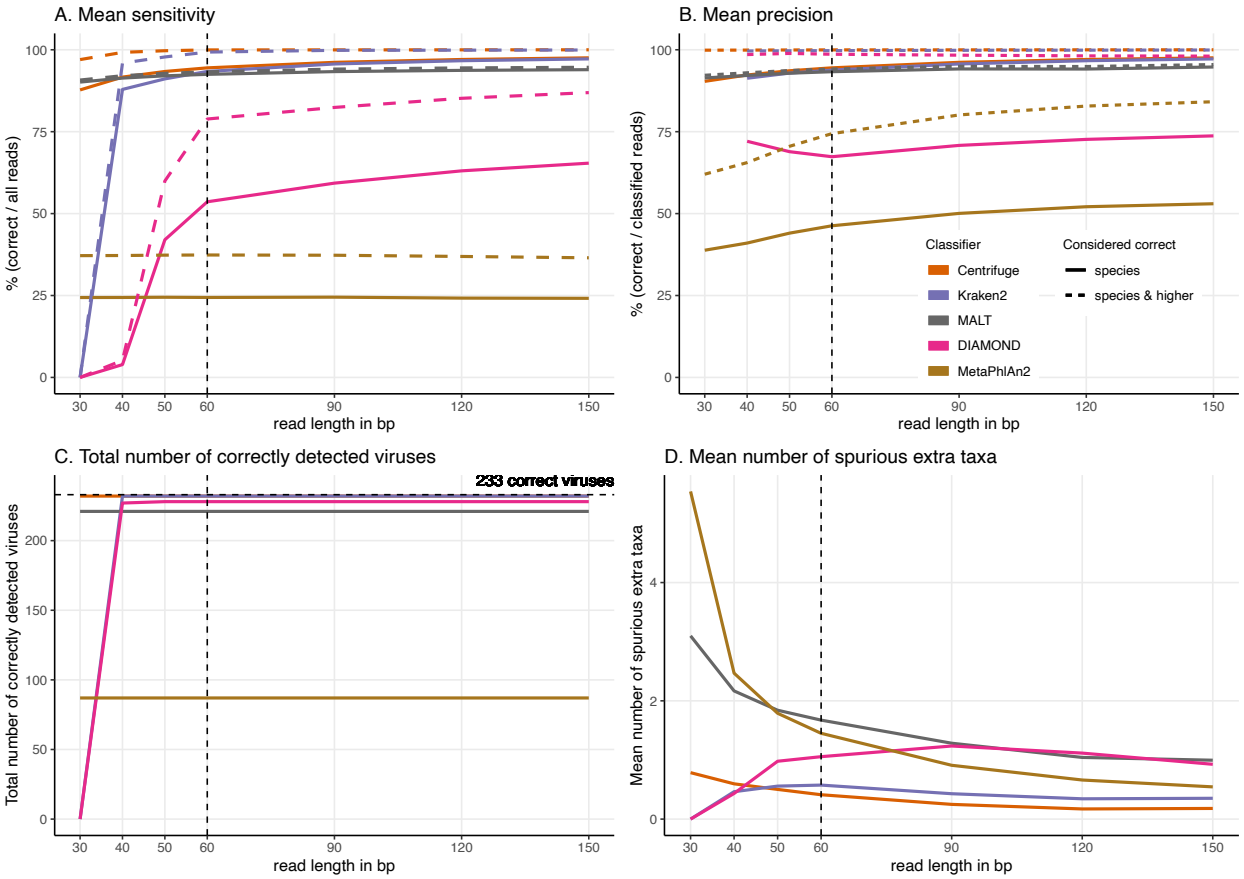


Figure SN1-3. For these simulations, read length was varied from 30 to 150 bp. A) Average Sensitivity_s (continuous lines) and Sensitivity_s&h (dashed lines) for each classifier. B) Average Precision_s (continuous lines) and Precision_s&h (dashed lines) for each classifier. C) Total number of viruses detected out of the 233 tested. The dashed line shows the maximum number of detectable viruses. D) Average number of spurious extra taxa across simulated viral sequences. The vertical dashed line indicates the initial 60 bp read set.

By adding deamination damage MALT has a slightly higher sensitivity (+~1%), only species level considered than Kraken2 (Figure SN1-4 A). However, if the correct higher taxa are considered, Kraken2 outperforms MALT both in sensitivity and precision (Figure SN1-4 B). Once more, DIAMOND has higher precision values than MALT by considering the correct high taxonomic classifications (Figure SN1-4 B); Centrifuge has the highest sensitivity and precision (Figure SN1-4 A & B); the number of identified viruses by MALT remains unaltered (Figure SN1-4 C); and MALT presents the highest number of spurious extra taxa (Figure SN1-4 D).

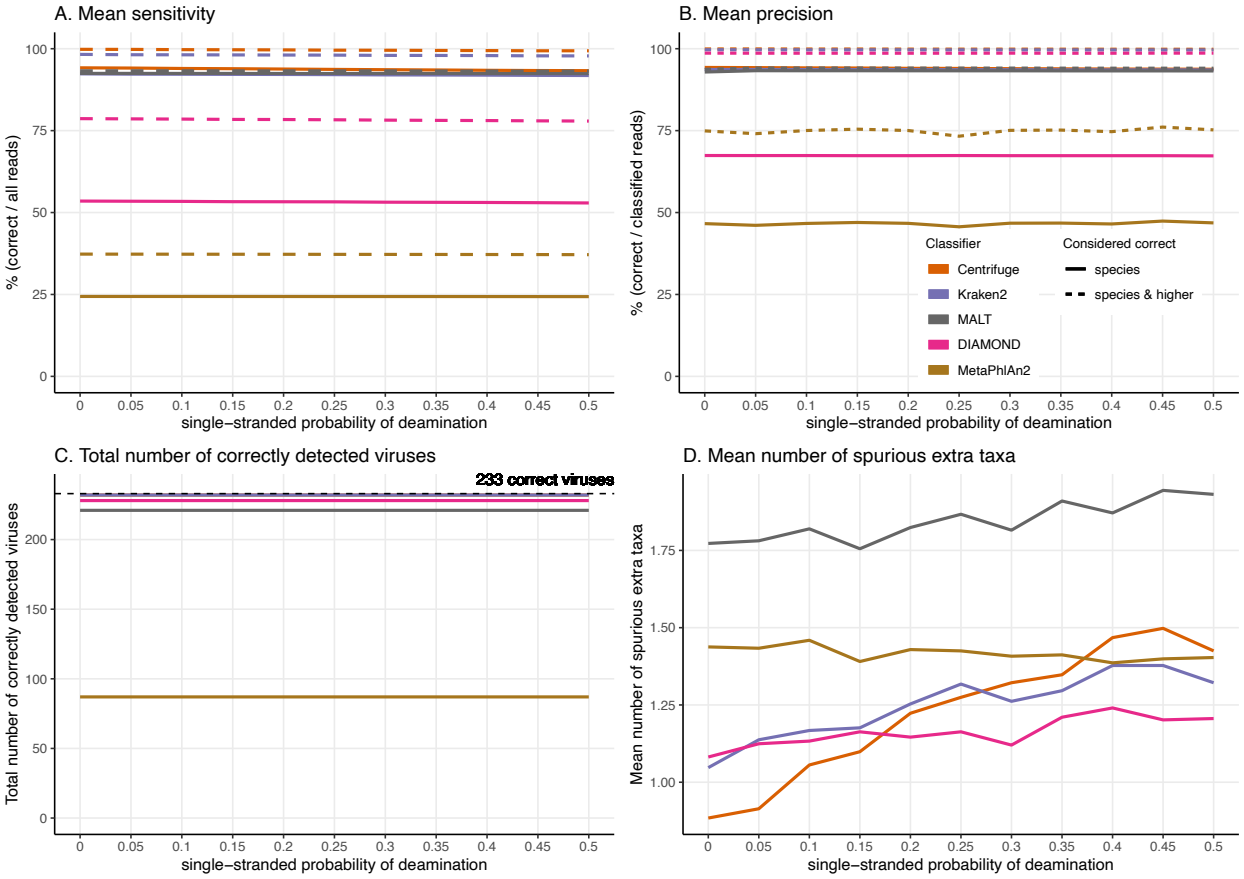


Figure SN1-4. For these simulations, errors were added using deamSim gargamel subprogram which assumes an ancient DNA deamination-like distribution and were added in addition to the ART Illumina like sequencing errors. The results shown here correspond to a single-stranded probability of deamination varying from 0 to 0.5. For all the results the nick frequency is set at 0.03, the average length of overhanging ends is set at 0.25, and the probability of deamination in the double-stranded portions of DNA is set at 0.01. A) Average Sensitivity_s (continuous lines) and Sensitivity_s&h (dashed lines) for each classifier. B) Average Precision_s (continuous lines) and Precision_s&h (dashed lines) for each classifier. C) Total number of viruses detected out of the 233 tested. The dashed line shows the maximum number of detectable viruses. D) Average number of spurious extra taxa across simulated viruses.

In the sequencing error simulations we observed that MALT has higher sensitivity (only considering species level) than Kraken2 for the lower qShift values (–5, –7 and –9, i.e. higher sequencing error) (Figure SN1-5 A). As with the deamination simulations Centrifuge outperforms all the tools (Figure SN1-5 A & B); DIAMOND has a higher precision (adding higher taxa) than MALT (Figure SN1-5 B); the number of identified viruses by MALT is the same across all qShift values (Figure SN1-5 C); and MALT remains as the tool with the highest number of spurious extra taxa (Figure SN1-5 D).

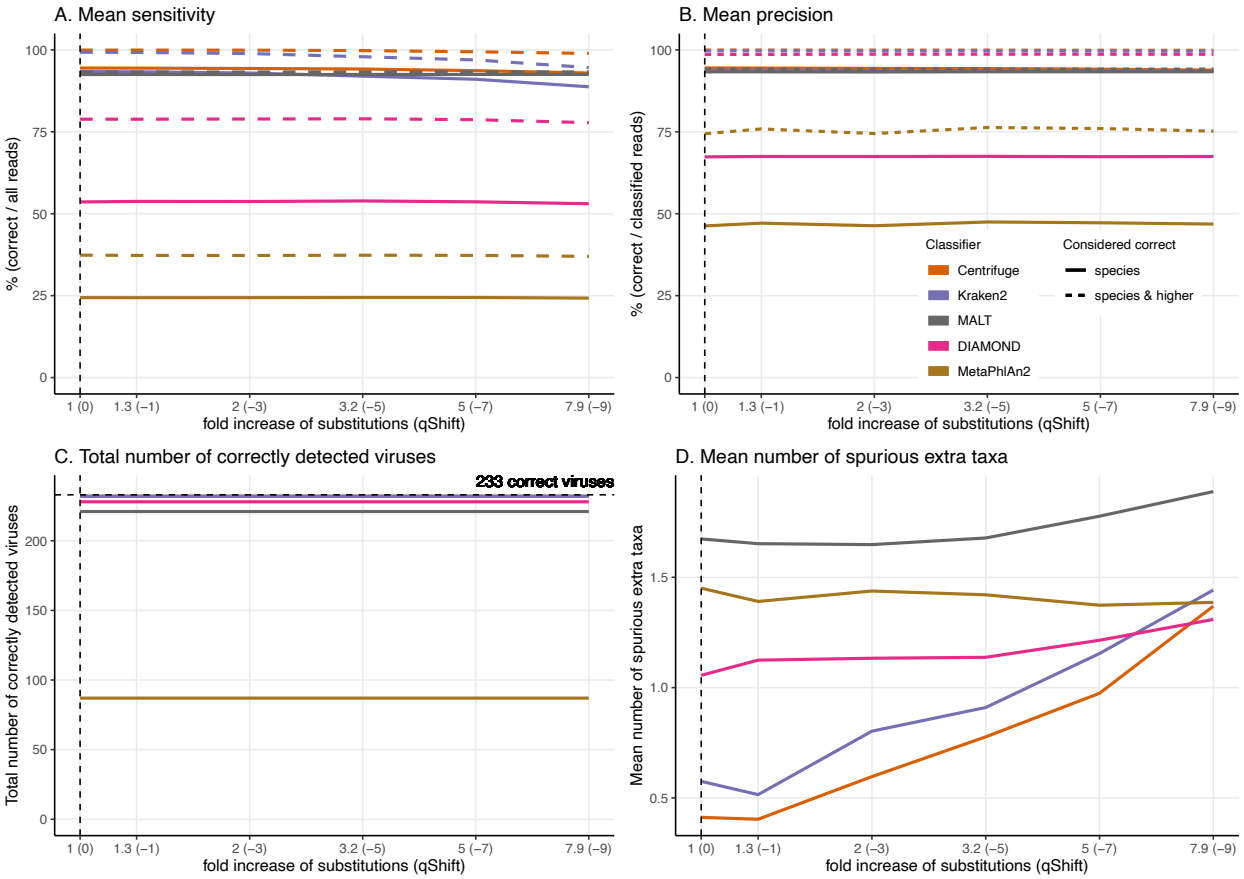


Figure SN1-5. For these simulations, errors were added using ART which assumes a profile similar to the ones observed for Illumina Sequencing machines (HiSeq 2500). The results shown here correspond to increasing the overall sequencing error rate ranging from 1 to 7.9-fold (qShift values from 0 to -9). On the x-axis, the first number correspond to the expected fold increase in error rate while the parameter that was varied, *qShift*, is shown in parenthesis. A) Average Sensitivity_s (continuous lines) and Sensitivity_s&h (dashed lines) for each classifier. B) Average Precision_s (continuous lines) and Precision_s&h (dashed lines) for each classifier. C) Total number of viruses detected out of the 233 tested. The dashed line shows the maximum number of detectable viruses. D) Average number of spurious extra taxa across simulated viral sequences. The vertical dashed line indicates the initial 60 bp read set.

Unidentified viruses and unexpected results

Finally, we analysed further the viruses that were not identified by MALT. By doing so, we detected incorrect translations of the sequence accession number to the taxonomic ID (taxID). Despite using different translation files (see above) and having changed the way we run the software, we still found discrepancies between MALT's alignments and MALT's taxonomic assignments. We found examples of three types of issues: 1) the read aligned to a specific taxon but is assigned incorrectly to a different one, 2) classifications that could be correct (the read aligned correctly to the virus tested, but did not have a taxonomic assignment), as well as, 3) classifications that could be incorrect (the read aligned incorrectly to another species rather than to the virus tested, but it did not have a taxonomic assignment). In the first case, and more specifically, it happens that MALT's alignment is to a specific taxon, but that the related reported taxon is a different one. This can lead to incorrect classifications with reported taxon not present in the database. For instance, the virus human papillomoavirus type 5 is aligning to Colobus guereza papillomavirus type 2, but the classification reports the primate *Colobus guereza,* a taxon not present in the database used. As an example of the second case, we identified reads generated from IAS

virus which were correctly aligned only to the IAS virus, but MALT did not report any reads assigned to this taxon. Finally, as an example of the third case, within the same set of IAS virus simulated reads, there were many reads uniquely aligned to crAssphage cr116_1, crAssphage cr128_1, crAssphage cr85_1 and to other members of the crAss-like viruses clade (all of them are taxa outside of the tested virus tree, i.e. they would be incorrect classifications), but none of these taxa are reported by MALT.

Note that some of the same issues has been already reported in 2018 (http://megan.informatik.uni-tuebingen.de/t/megan-assignation-after-malt/753/2) but there does not seem to be an answer. To try to resolve the issue we looked into the code, contacted the authors (http://megan.informatik.uni-tuebingen.de/t/malt-reports-classifications-outside-of-db-range/1909) and asked the "SPAAM" community (https://spaam-community.github.io/). We have not been able to resolve the issue so far.

To conclude, even though MALT looks very promising, given these results we cannot explain, it is unclear to us if the overall classification results we obtained are trustworthy.

Supplementary Note 2: archaeal, bacterial, human and viral databases for Centrifuge, Kraken2 and DIAMOND

In this section we present the classification results of the read length simulations (see Materials and methods) obtained with Kraken2 and Centrifuge's default databases (i.e. the databases that include archaeal, bacterial, human and viral data), as well as with an archaeal, bacterial, human and viral database for DIAMOND. Note that the results presented in the main text were obtained with viral databases for all classifiers except MetaPhlAn2.

Methodology
We built Centrifuge and Kraken2's databases following the developers' instructions. Centrifuge database includes RefSeq sequences from archaea, bacteria, human and viruses (built on 2 September 2021); Kraken2 database in addition has sequences from UniVec_Core (downloaded 17 May 2021). To build DIAMOND's database we used protein sequences from archaea, bacteria, human and viruses (so that all databases were comparable), downloaded from RefSeq on 6 September 2021. In the following lines we will refer to these databases as "full databases".

Results
In the initial 60 bp reads simulation we observed a better performance of the classifiers using the viral database (Figure SN2-1 & Figure SN2-2). The number of correct species and correct higher classification is greater when using viral databases for all three classifiers. In addition, the number of incorrect classifications and unclassified reads are reduced for the viral databases (compared to the full databases). This is likely explained by the greater number of species in the full databases, many of them can confound the taxonomic assignments as they could resemble the tested viruses (Figure SN2-1).

In the specific case of DIAMOND, the proportion of unclassified reads reaches 67.5% when using the full database. As the full database stores a greater number of sequences than the viral database the e-values of DIAMOND alignments increases; it therefore seems that many reads do not pass DIAMOND's threshold resulting in a greater proportion of unclassified viruses.

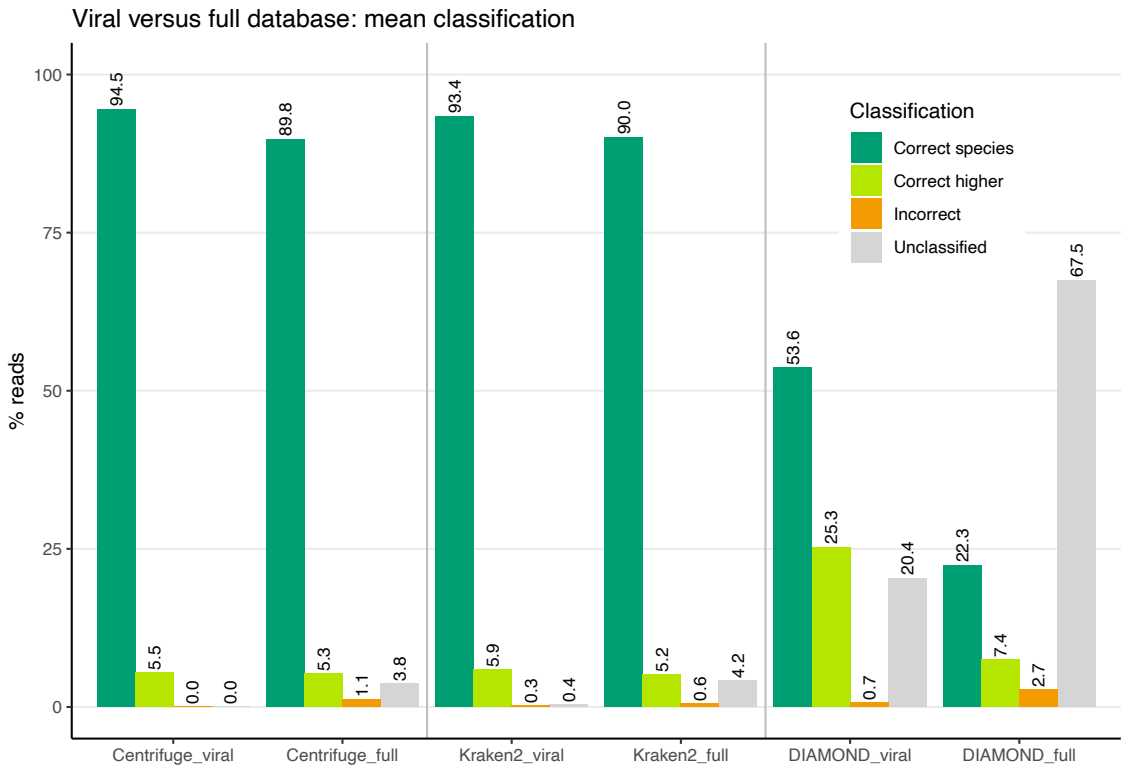


Figure SN2-1. Mean percentage over the 233 human DNA viruses of reads classified in each of the four categories: "correct species", "correct higher", "incorrect" and "unclassified" (see Figure 1). The bars with the suffix "_viral" show the means obtained with the "viral" databases. The bars with the suffix "_full" show the means obtained with the "full" databases, which contain archaeal, bacterial, human and viral sequences.

The aforementioned proportions are reflected in the mean sensitivity and precision values: the viral databases sensitivity and precision scores are higher than the full databases ones, both when considering only species level and when including higher taxa (Figure SN2-2 A & B). This difference is accentuated in the sensitivity, given that this measure considers all reads and the proportion of unclassified reads is higher with the full databases. The exception are Kraken2 and DIAMOND, which have slightly higher values for precision_s when using its full database (Figure SN2-2 B).

Regarding the number of identified viruses: Kraken2 and Centrifuge identify more viruses with the viral database, while DIAMOND detects the same number of viruses regardless the database used (Figure SN2-2 C). When checking Kraken2 and Centrifuge databases we found out that their default databases (full databases) did not include sequences for 1 virus in the case of Kraken2, and for 6 in the case of Centrifuge. We tried to re-build these databases several times but the viruses were missing, we do not know why the sequences are not automatically downloaded for those species during the steps to build a database for those two tools.

As discussed above, for all the classifiers we observed a higher amount of spurious extra taxa reported when they use their full databases compared to the viral database. This is expected as there are more possible confounding species in the full databases (Figure SN2-2 D).

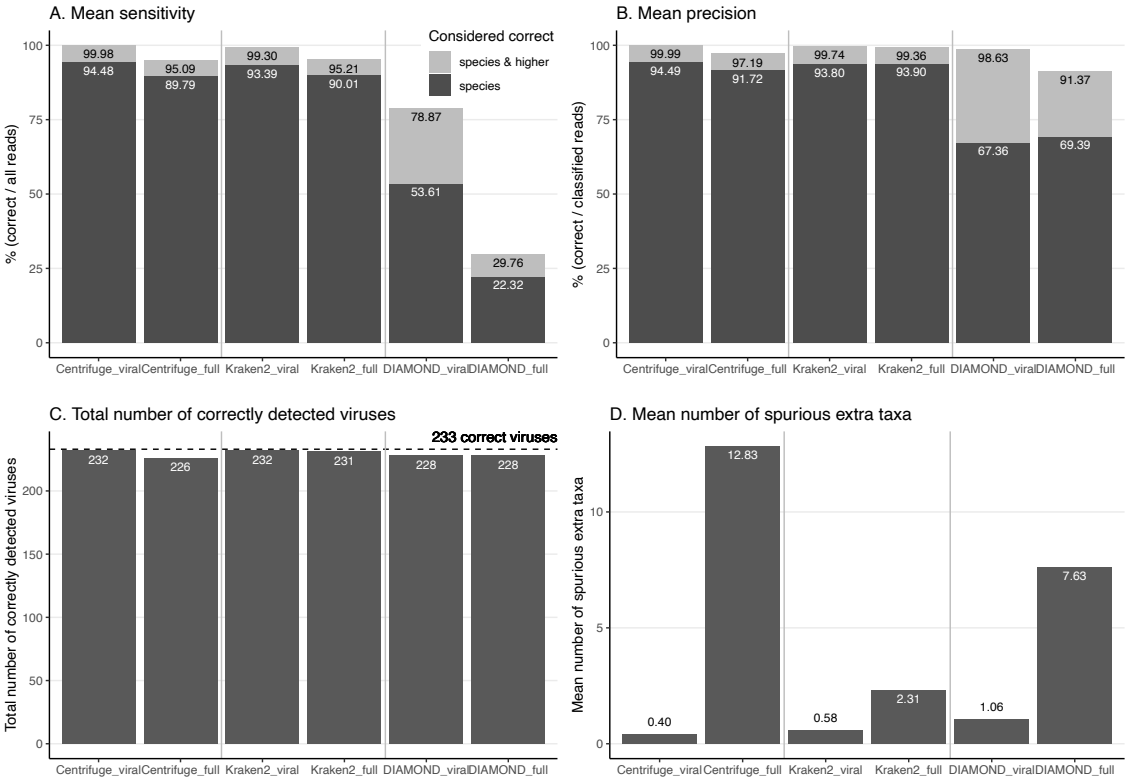


Figure SN2-2. A) Mean sensitivity over the 233 human DNA viruses. B) Mean precision over the 233 human DNA viruses. Darker grey bars show sensitivity and precision when considering species level only, light grey bars show sensitivity and precision when considering species level and higher taxonomic ranks together C) Total number of correctly detected viruses over the 233 tested. D) Mean number of reported spurious extra taxa over the 233 human DNA viruses. The bars with the suffix "_viral" show the results obtained with the "viral" databases. The bars with the suffix "_full" show the results obtained with the "full" databases, which contain archaeal, bacterial, human and viral sequences.

Similar trends to the 60 bp reads simulation are observed when varying the read length: higher sensitivity and precision with the viral databases; and a higher amount of spurious extra taxa with full databases (Figure SN2-3).

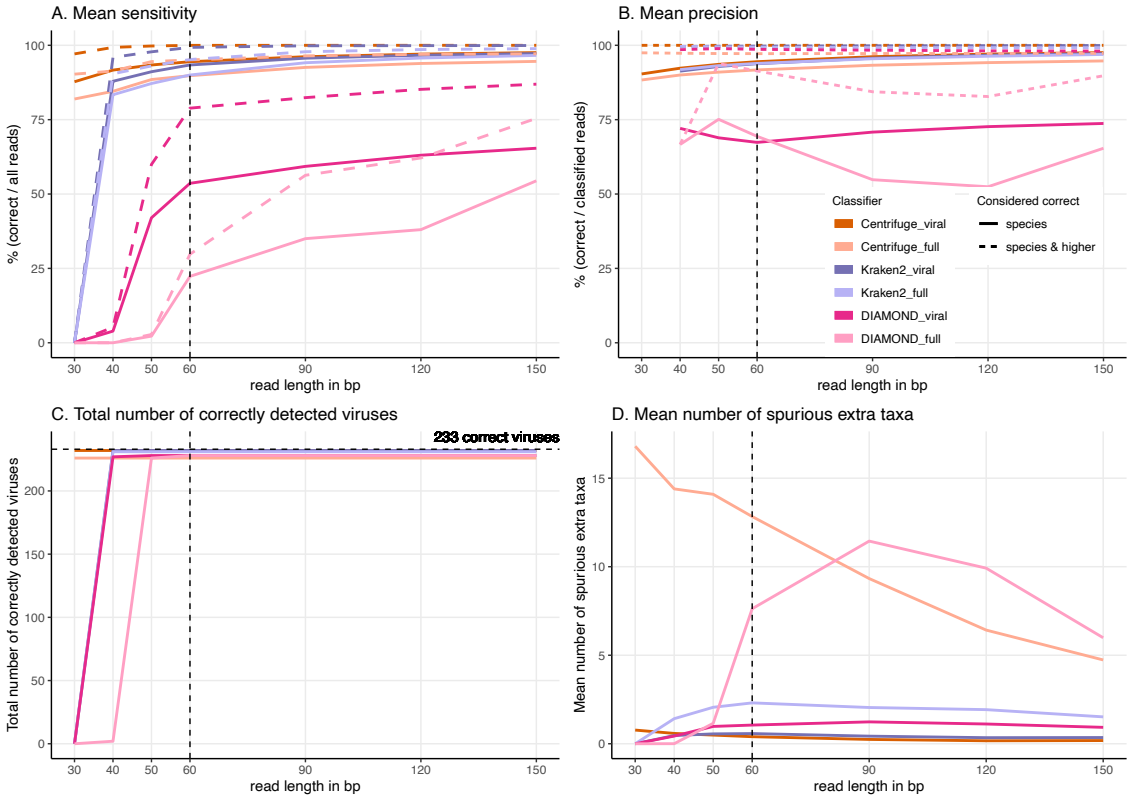


Figure SN2-3. Comparison of "viral" databases (darker colors) vs "full" databases (lighter colors) output. Simulated reads of different lengths (x axis) from the 233 viruses were used (see Materials & Methods in main). A) Average Sensitivity_s (continuous lines) and Sensitivity_s&h (dashed lines) for each classifier and for each database. B) Average Precision_s (continuous lines) and Precision_s&h (dashed lines) for each classifier and for each database. C) Total number of viruses detected out of the 233 tested. The dashed line shows the maximum number of detectable viruses. D) Average number of spurious extra taxa across simulated viral sequences. The vertical dashed line indicates the initial 60 bp read set.

Discussion

In summary, as the goal with the classification step in an ancient DNA study is to increase the number of true candidates that can be further explore, we believe that the viral database is the most appropriate choice of database to detect ancient (human) DNA viruses. We have made those databases available.

Supplementary Note 3: New parameters for DIAMOND and Kraken2 to handle short reads

As discussed in the main text, with default parameters, short reads cannot be "handled" by DIAMOND and Kraken2 (see Figure 5 in main, there are no classifications for neither of them). In this supplementary note, we implemented a set of simulations to test the performance of DIAMOND and Kraken2 with parameters that allow both of them to classify short reads (30 bp long).

Methodology

DIAMOND:
For this set of simulations, we reduced the length of DIAMOND's seed to 9 aminoacids with the parameter *--shape-mask 111111111*, which can align reads of length 27 bp or longer. The rest of the parameters and the database (viral sequences database) remained the same (default parameters).

Kraken2:
In the case of Kraken2 we built a new database (with the same sequences as the default one: archaeal, bacterial, human, viral and UniVec_Core sequences) with k-mers of length 29 using the parameter *--kmer-len 29* (i.e. in theory, reads of 29 bp or longer can be classified). The rest of the parameters and the database (viral sequences database) remained the same.

Results

With these new simulations, the reads were classified for different read lengths (30, 40, 50, 60, 90, 120 and 150 bp long) generated out of the 233 viruses (see main Materials and methods, Simulation of viral reads and read length); and the outputs of DIAMOND and Kraken2 were compared for default parameters versus a seed of 9 amino acids (DIAMOND) and 29 bp-kmer (Kraken2). The main results are shown below:

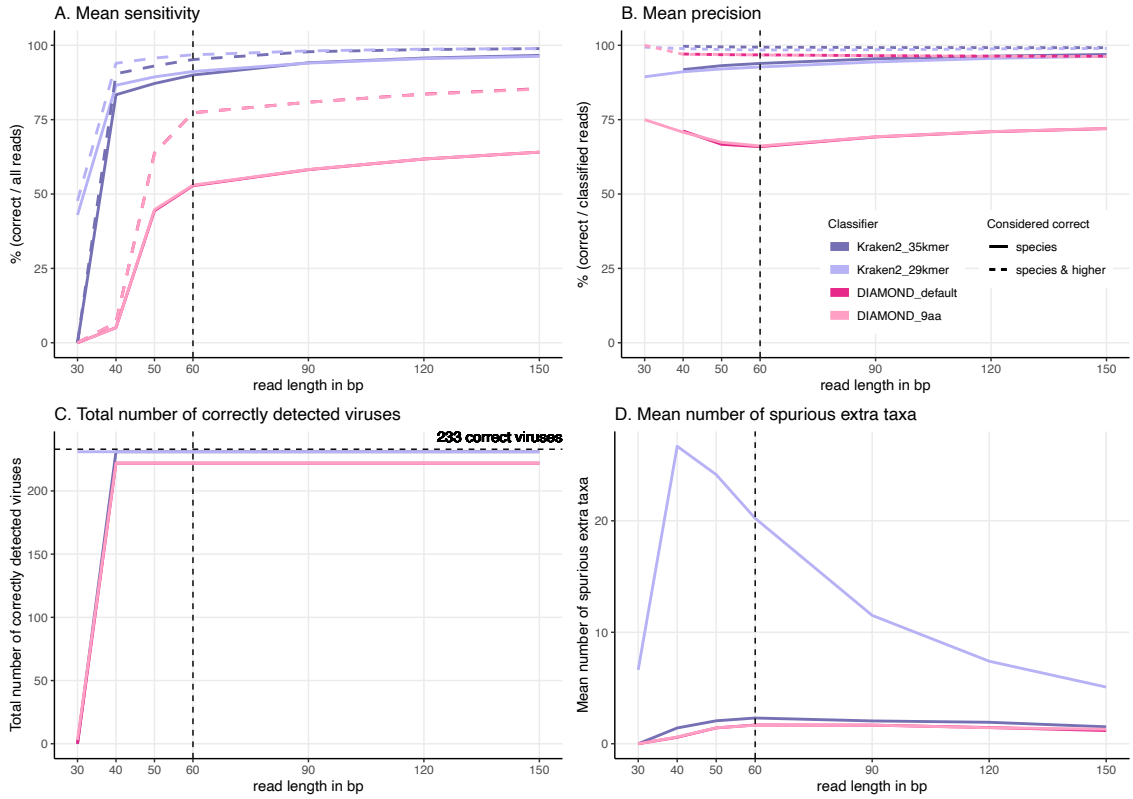


Figure SN3-1. DIAMOND and Kraken2's parameters for short reads. DIAMOND (pink colors) and Kraken2 (purple colors) comparison of default parameters (darker colors) versus parameters tailored for short reads (lighter colors). Simulated reads of different lengths (x axis) from the 233 viruses were used (see Materials & Methods in main). A) Average Sensitivity_s (continuous lines) and Sensitivity_s&h (dashed lines) for each classifier and for its different parameters. B) Average Precision_s (continuous lines) and Precision_s&h (dashed lines) for each classifier and for its different parameters. C) Total number of viruses detected out of the 233 tested. The horizontal dashed line shows the maximum number of detectable viruses. D) Average number of spurious extra taxa across simulated viral sequences. The vertical dashed line indicates the initial 60 bp read set.

By changing the parameters used to build the database (Kraken2) or to run the software (DIAMOND), both Kraken2 and DIAMOND are able to classify 30 bp reads (Figure SN3-1). Kraken2 with 29 bp k-mers has almost a sensitivity of 50% which contrasts with the 0% sensitivity of Kraken2 with 35 bp k-mers (default k-mer length), however as the read length increases, the sensitivity of Kraken2 with 35 bp k-mers and the sensitivity of Kraken2 with 29 bp k-mers reach a similar value (Figure SN3-1 A). Something similar happens with Kraken2's precision, although the precision is less affected (Figure SN3-1 B). Regarding the number of detected viruses both k-mers identify the same number across all lengths. Remarkably, Kraken2 with 29 bp k-mers also identifies the same viruses in the 30 bp reads simulations (Figure SN3-1 C). However, a considerable disadvantage observed when using 29 bp k-mers is the increase of spurious taxa reported by Kraken2; across all lengths the mean number of spurious extra taxa is considerably larger when using 29 bp k-mers compared to the default 35 bp-kmers (Figure SN3-1 D). In DIAMOND's case, the number of reads recovered in the 30 bp simulations with a shorter seed of 9 aminoacids is minimal (Figure SN3-1 A). In all cases tested (sensitivity, precision, detected viruses and mean number of spurious taxa) there is almost no difference of using a shorter seed of 9 aminoacids or using DIAMOND's default seeds (Figure SN3-1).

Discussion

Considering these mixed results, we present the default parameters in the main text. Considerably more work is needed to find out the best combination of parameters for those tools for ancient DNA short reads.

Supplementary Note 4: Unclassified viruses and viruses with high incorrect classifications

Unclassified viruses
None of the classifiers identified all the 233 viruses tested. The failure of identifying some viruses can be explained mainly by two reasons: either the virus sequence is very similar to sister species in the database or the virus was not represented on the database. In the following lines we give a detailed explanation of each of the missing viruses by each of the classifiers tested.

Kraken2
Kraken2 failed to identify human adenovirus 2 (AC_000007.1) because there is no minimizer specific to this serotype in the database, then the virus is mostly classified at a higher level (species) to human mastadenovirus C (5,943 reads). There were as well a couple of reads correctly classified at the genus level and few reads (6) incorrectly classified that were confounded with sister serotypes (human adenovirus 1 and human adenovirus 5).

DIAMOND
Within the 5 viruses not identified by DIAMOND, 3 of them have no proteins in RefSeq: porcine endogenous retrovirus E (NC_003059.1), senegalvirus marseillevirus (NC_04322[3-8]).1 and torque teno virus (NC_015783.1). Porcine endogenous retrovirus E and senegalvirus marseillevirus have correct classifications at higher taxonomic levels; while torque teno virus only have reads incorrectly classified given that most of the alignments were to species in the genus *Alphatorquevirus*, whereas torque teno virus is catalogued within the unclassified *Anelloviridae* rank. The other two viruses that DIAMOND was not able to identify were human adenovirus 2 (AC_000007.1) and human polyomavirus 12 (NC_020890.1), both of them had alignments but also to close sister species, then DIAMOND classified their reads as "correct higher" instead of "correct species".

Centrifuge
As Kraken2 and DIAMOND, Centrifuge failed to identify human adenovirus 2 (AC_000007.1). As Centrifuge compresses genomes from strains of the same species (Kim et al. 2016) almost all alignments were against the species level human mastadenovirus C, just two reads were misclassified as the sister serotype human adenovirus 1.

MetaPhlAn2
Finally, in MetaPhlAn2's case 146 viruses do not have any "correct species" reads. Out of the 151 viral sequences with no reads classified correctly at species level, 65 of them do not have any classifications, 33 have classifications only at a higher taxonomic level, 37 have only incorrect classifications and 16 have both high correct and incorrect classifications. Out of the 146 viruses, only one has a marker gene associated to its accession number in MetaPhlAn2's database: porcine endogenous retrovirus E (NC_003059.1), which is correctly classified but at a higher taxonomic level.

Incorrect classifications
Some of the tested viruses present high proportions of incorrectly classified reads for specific classifiers. As with the undetected viruses we observed that these misclassifications are due to the absence of the virus in the database, or because there are many sister species that confound the taxonomic assignment. Encouragingly, for almost all viruses across all classifiers, the incorrect proportion does not exceed 40%. In the following lines we discuss which were the viruses with the highest proportions of incorrect classifications per classifier.

Kraken2

Kraken2 has only four viruses with a proportion of incorrectly classified reads higher than 1%. adeno-associated virus - 8 (NC_006261.1) is the virus with the highest incorrect proportion with 24.5% incorrectly assigned reads, all of them assigned to close sister species; followed by adeno-associated virus - 7 (NC_006260.1) with 18.2%, adeno-associated virus - 3 (NC_001729.1) with 8.4% and adeno-associated virus - 4 (NC_001829.1) with 6.4%, in all the cases the incorrect assignments were to sister species within the adeno-associated dependoparvovirus A rank.

Centrifuge

For Centrifuge all the tested virus has less than 1% incorrect classified reads. The virus with the highest proportion of incorrect reads was cyclovirus VN (NC_039215.1) with 0.33%, the incorrect assignment was to the sister species cyclovirus TN25.

DIAMOND

In the case of DIAMOND, torque teno virus (NC_015783.1) has the highest incorrect proportion of classified reads with 42.7% as this virus has no proteins stored in RefSeq; all classified reads were incorrectly assigned to species in the genus *Alphatorquevirus*, whereas torque teno virus is catalogued within the unclassified *Anelloviridae* rank. The virus with the second highest incorrect proportion is gemycirculavirus C1c (NC_038496.1) with 29.7%, the incorrect assignments were to mongoose feces-associated gemycircularvirus c and porcine feces-associated gemycircularvirus, both of them belong to the same genus (*Gemykibivirus*) as the tested virus. The contigs from senegalvirus marseillevirus (NC_04322[3-8].1) also present high proportions of incorrect assignments ranging from 26.1% to 7.6%. As with torque teno virus this virus does not have any protein in RefSeq, the incorrect assignments are within the *Marseillevirus* genus.

MetaPhlAn2

In the case of MetaPhlAn2 the viruses with the highest proportions of incorrectly classified reads are not present in its database. Human papillomavirus 179 (NC_022095.1) has an incorrect proportion of 91.3%, with all reads classified as human papillomavirus which belongs to the clade human papillomavirus types, while human papillomavirus 179 belongs to the *Gammapapillomavirus* genus. It is followed then by human parvovirus 4 G1 (NC_007018.1) with 89%, all reads were classified as the sister species human parvovirus 4. Then we have human bocavirus 2c PK (NC_012042.1) with 51.6%, all the assignments were to the sister species human bocavirus 2 and to bocavirus gorilla/GBoV1/2009, a member of the same genus (*Bocaparvovirus*) as human bocavirus 2c PK.