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. Many of them assign sequenced reads to specific taxa 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 virome data remains unknown.

In this work, we conduct an extensive simulation study using public viral sequences to establish which tool is the most suitable for ancient virome studies. We compare 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 simulate reads by adding noise typical of ancient DNA to a randomly chosen set of publicly available viral sequences and to the human genome. We fragment the DNA into different lengths, add sequencing error and C to T and G to A deamination substitutions at the read termini. Then we measure the resulting precision and sensitivity for all classifiers.

Across most simulations, 119 out of the 120 simulated viruses are recovered by Centrifuge, Kraken2 and DIAMOND in contrast to MetaPhlAn2 which recovers only around one third. While deamination damage has little impact on the performance of the classifiers, DIAMOND and Kraken2 cannot classify very short reads. For data with longer fragments, if precision is strongly favoured over sensitivity, DIAMOND performs best. However, since Centrifuge can handle short reads and since it achieves the highest sensitivity and precision at the species level, it is our recommended tool overall. 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, if the goal is to detect a specific virus, given the high variability observed among tested viral sequences, a simulation study to determine if a given tool can recover the virus of interest should be conducted prior to analysing real data.

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 for instance to pinpoint the origin of a pathogen (Taubenberger et al. 2005; Mühlemann, Jones, et al. 2018; Rascovan et al. 2019), to characterize past epidemics and the prevalence of pathogens across time (Mühlemann, Jones, et al. 2018), and to inform about the evolutionary history (Duggan et al. 2016) and the changes in geographical distribution of given microbes (Worobey et al. 2016). For instance, among the recent discoveries related to ancient viruses, the recovery of the genome of the Spanish Influenza virus suggested that it had an avian origin (Taubenberger et al. 2005). Moreover, the detection and reconstruction of hepatitis B virus (HBV) genomes suggested that modern genotypes are the result of a recombination event across ancient strains (Mühlemann, Jones, et al. 2018). Additionally, HBV's prevalence across time was found to vary from the Bronze Age to Middle Ages across space (Mühlemann, Jones, et al. 2018). Furthermore, the reconstruction of a variola genome from a mummified tissue of a child that lived in the 17th century, allowed to date the gene disruption that is characteristic of the modern strains (Duggan et al. 2016). Finally, the retrieval of HIV virus genomes from serum samples permitted to track the geographical spread of the virus from the Caribbean to the USA, and from New York to San Francisco (Worobey et al. 2016).

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). Ancient DNA 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).

The biochemical reactions causing fragmentation and substitutions have been characterised. Fragmentation is the consequence of depurination. This post-mortem reaction consists in the excision of purine bases from the rest of the DNA molecule. As a result, abasic sites are formed, 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). A large fraction of the DNA fragments is considerably shorter than 100 bp resulting in sequenced reads that often read into the sequencing adapters. Substitutions at the end of the DNA fragments are caused by deamination, the reaction in which cytosine loses an amine group. Deamination also takes place post-mortem and results in the transformation of cytosines into uracils. After sequencing, deamination will lead, for double stranded libraries, to an increase of 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. As the DNA fragments are short, these C to T and G to A changes can be observed at the sequenced read termini (Briggs et al. 2007; Carøe et al. 2018). Damage patterns are useful for the authentication of ancient DNA, as ancient samples are usually contaminated and include DNA from different sources.

Sequenced ancient DNA is generally composed of DNA from the sampled organism (endogenous fraction), the microbiome of the organism at the time of death (the fraction of interest in the present work), but also of DNA from environmental microorganisms of the past and the present (including the DNA of microbes attacking the corpse or the DNA found in lab reagents), and possibly the DNA of the researchers involved in the sampling and sequencing processes (modern human contamination) (Hofreiter et al. 2001; Warinner et al. 2017).

One of the fundamental tasks when studying the microbiome is the identification of the species present in the sample. Computational tools have been developed for the taxonomic assignment of reads coming from metagenomic samples. Such tools, referred to as 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 present study, we compare four widely used classifiers which differ in their approaches: Centrifuge (Kim et al. 2016), Kraken2 (Wood et al. 2019), DIAMOND (Buchfink et al. 2015) and MetaPhlAn2 (Truong et al. 2015). 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).

In this study we investigate the performance of Centrifuge, Kraken2, DIAMOND and MetaPhlAn2 on ancient DNA-like reads. We simulate viral reads and focus on the effects of fragmentation, deamination, sequencing error and human contamination on the classification. We investigate the impact of varying the read length and artificially adding deamination damage or sequencing error on the taxonomic assignments for each classifier. Finally, as many studies focus on the ancient virome in humans, we evaluate the effect of human contamination by classifying simulated fragmented human reads with the four classifiers.

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, the assignments made by the classifiers on simulated short human reads were analysed.

Reference viral sequences

A total of 120 reference viral sequences were randomly selected from the viral genomic NCBI RefSeq database - which is curated and annotated and contains a total of 10,544 entries (Brister et al. 2015; O'Leary et al. 2016) (downloaded on 18 December 2018). The 120 selected viral reference sequences consisted of 67 complete genome sequences, 33 genome segments and 20 coding sequences (CDS). Note that segmented genomes belong to viruses whose genetic information is not contained in a single molecule of DNA or RNA, but rather is divided into several molecules of nucleic acids. In the RefSeq database, viruses with segmented genomes have one sequence per segment (Brister et al. 2015). CDS are included in RefSeq because whole genomes are not yet available for some viruses as some sequencing methods rely on primers that can be unavailable, especially for the termini (Alfson et al. 2014). The sequence lengths of the selected viruses varied between 932 and 2,077,288 bp, in line with the length distribution of the whole database (Figure S1). 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 120 randomly selected viral 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 resulting in typical Illumina HiSeq 2500 error profiles. In an initial simulation, for each of the 120 reference viral sequences, single-end reads of a fixed length were simulated emulating the short read length observed in 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 150 to 346,210 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×. Each set of reads from the 120 selected viral sequences 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

Centrifuge

Centrifuge version 1.0.3-beta was used with the RefSeq database including bacterial, archaeal, viral and human genomes (built on 7 February 2019 following the software instructions). 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. The "standard" Kraken2 database was chosen. This database includes the RefSeq of bacterial, archaeal, viral and human genomes plus UniVec_Core (built on 11 February 2019 following the software instructions).

DIAMOND

DIAMOND version 0.9.22 with the RefSeq viral protein database (downloaded on 9 November 2018) was used. 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.

MetaPhlAn2

MetaPhlAn2 version 2.6.0 was run with default parameters. The default microbe clade-specific marker genes database was used, 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:

$$Sensitivity_s = \frac{1}{120}\sum_{i=1}^{120}\frac{r_{cs}^{v_i}}{r_{cs}^{v_i}+r_{ch}^{v_i}+r_{i}^{v_i}+r_{u}^{v_i}}$$

$$Sensitivity_{s\&h} = \frac{1}{120}\sum_{i=1}^{120}\frac{r_{cs}^{v_i}+r_{ch}^{v_i}}{r_{cs}^{v_i}+r_{ch}^{v_i}+r_{i}^{v_i}+r_{u}^{v_i}}$$

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_007555.1 (Adult diarrheal rotavirus strain J19 complete genome) 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, -9, -11). 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, -11 to a 12.6-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_007555.1 (Adult diarrheal rotavirus strain J19 complete genome) 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 superkingdom 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 S2 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 S2). 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 22% (MetaPhlAn2), 47% (DIAMOND), 89% (Kraken2) and 90% (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. In the case of DIAMOND, no virus stands out and most of the simulated viruses have no "incorrect" reads. Centrifuge and Kraken2 show a similar small proportion for "incorrect" reads for most viruses, but in contrast to DIAMOND, there is also a few (five) viral sequences for which the proportion of "incorrect" reads exceeds 60% (Figure S2). MetaPhlAn2 also has a large proportion of "incorrect" reads for 13% of the viruses. The resulting averages of the proportions of "incorrect" reads are 0.2% (DIAMOND), 2.3% (MetaPhlAn2) and ~3.6% (Centrifuge and Kraken2). Finally, Centrifuge and Kraken2 classify most of the simulated reads for all viruses (with a mean "unclassified" below 1.33% for both), while both DIAMOND (mean of 35%) and MetaPhlAn2 (mean of 73%) 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). At the species level, both means are considerably higher for Centrifuge and Kraken2 with mean values above 90% for Sensitivity_s and Precision_s. In contrast, MetaPhlAn2's sensitivity and precision are the lowest with values of 22.4% (Sensitivity_s) and 57.7% (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 (Material and Methods). At the species level, DIAMOND has intermediate values: a Sensitivity_s of 47.4% and a Precision_s of 73.5%, 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 for Sensitivity_s&h but is changed for Precision_s&h. 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 64.8% and 47.4%, respectively. Similarly, its Precision_s&h (99.73%) is higher than its Precision_s (73.5%). As we observed for Sensitivity_s, DIAMOND's Sensitivity_s&h is lower than Centrifuge's and Kraken2's, and higher than MetaPhlAn2's. However, since DIAMOND has the lowest proportion of "incorrect" reads, it achieves the highest Precision_s&h, with a value of 99.7%, followed closely by Centrifuge and Kraken2 (~96.3%) then by MetaPhlAn2 (63.8%, 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 (or more) correctly classified read(s) (Figure 4B). Encouragingly, despite large differences observed at the read level, Centrifuge, Kraken2 and DIAMOND detected 119 out of the 120 investigated viruses by correctly classifying at least one read. Centrifuge and Kraken2 failed to detect at species level the viral sequence Squash leaf curl China virus – [B] (NC_007339.1). Nevertheless, this virus is classified correctly at higher taxonomic levels by both classifiers. DIAMOND failed to detect the virus *Trichoplusia ni cypovirus 15* (NC_002561.1). The only protein stored in the RefSeq viral protein database for this virus (named polyhedrin, NP_066381.1) is not included in the genome segment used to generate the simulated reads (NC_002561.1). In contrast, MethaPhlAn2 recovered only 39 viruses, that is ~ one third of the tested viruses (Figure 4B). Moreover, MetaPhlAn2 detect only four of the viruses at higher taxonomic levels. The majority of the viruses have either "incorrect" reads or remain unclassified.

In addition, we computed how many spurious extra taxa were reported, that is taxa that were neither the virus of interest nor in the lineage of that virus (Figure 4C). We found that Centrifuge is the classifier with the highest number of spurious extra taxa (on average 10.8 across tested viruses), followed by Kraken2 (7.2), MetaPhlAn2 (6.6) and, finally, DIAMOND (5.1). For Centrifuge, Kraken2 and MetaPhlAn2 the number of false positives can be reduced by using a more stringent threshold to consider a virus present without affecting the number of true viruses detected (119 for Centrifuge and Kraken2 and 39 for MetaPhlAn2). Here we observed that by increasing the threshold from 1 to 50 reads necessary to identify a virus, the average number of spurious extra decreases to 1.4 for Centrifuge and Kraken2, and it reduces to 2.1 for MetaPhlAn2. This strategy is somewhat less effective for DIAMOND: although the number of extra taxa is decreased when augmenting the threshold of necessary reads, fewer of the 120 tested viruses are then recovered (i.e. this strategy decreases the number of false positive but also the number of true positive). For instance, when requesting at least 50 reads, the average number of spurious extra taxa is reduced from 5.1 to 1.4 (Figure 4C) but only 113 true viruses are recovered instead of 119 (Figure 4B). Hence, for DIAMOND, for these simulations an intermediate value for the threshold is a good compromise; with a threshold of 10 extra reads, only one true virus is lost (118 detected) while the average number of spurious extra taxa is halved (2.4. extra taxa on average). Note however that the ideal threshold will be highly dependent on the viruses considered since viral genomes are very distinct, including in length (Figure S1), and differ in their abundance within the host. Hence, it is hard to determine a widely applicable threshold.

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 S3 & S4).

In general, by increasing the read length, the classifications improve. We observe higher proportions of "correct species" and "correct higher" reads (Figure S3 & S4), fewer "incorrect" reads, higher sensitivity and precision, and a decrease in the number of spurious extra taxa (Figures S3 & S4).

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. Similarly, DIAMOND classifies only very few reads for 40 bp and 50 bp compared to higher read length. This contrasts with the results for Centrifuge and MetaPhlAn2 whose performance do not significantly deteriorate with shorter reads (Figure 5).

From 60 bp onwards, most of the observations discussed above hold; Kraken2 and Centrifuge have the highest "correct species" mean proportions, followed by DIAMOND and MetaPhlAn2 (Figure S3). DIAMOND has the highest "correct higher" and the lowest "incorrect" mean proportions. Consequently, from 60 bp onwards, the highest sensitivities (Sensitivity_s and Sensitivity_s&h) and Precision_s are observed for Kraken2 and Centrifuge followed by DIAMOND and MetaPhlAn2. In contrast, and as above, DIAMOND has the highest Precision_s&h as it classifies correctly many reads at higher taxonomic level and makes few classification mistakes but has a Sensitivity_s&h that is lower than Kraken2 and Centrifuge's. Moreover, as above, regardless of the simulated read length, most viruses are detected by Centrifuge, Kraken2 and DIAMOND (>117/120) while most viruses (81/120) are not detected with MetaPhlAn2 (Figure 5C). As for 60 bp, for Centrifuge and Kraken2, the only undetected viral sequence was the same across all read lengths, namely the Squash leaf curl China virus – [B] which is only classified correctly at higher taxonomic levels. For DIAMOND, *Trichoplusia ni cypovirus 15* (NC_002561.1) remained undetected regardless of the read length. Surprisingly, DIAMOND fails to detect two extra viruses at read lengths higher than 60 bp (i.e. it detects 117/120 for reads > 60 bp), namely *Melaka orthoreovirus* (NC_020439.1) and *Orbivirus SX-2017a* (NC_033788.1). The reads for those viruses are only classified correctly at higher taxonomic ranks.

Interestingly, the average number of spurious extra taxa varies greatly between classifiers and across read lengths (Figure 5D). A sharp decrease in the number of extra taxa for increased read lengths is observed for Centrifuge and MetaPhlAn2, the two most constant classifiers across read lengths for other statistics. In comparison, the average number of spurious extra taxa is rather constant for Kraken2 and DIAMOND for read lengths above 40 bp and 60 bp, respectively. Up to 120 bp the highest number of spurious extra taxa is found for Centrifuge, while Kraken2 has the worst value at 150bp. DIAMOND has the lowest value up to 60 bp. Above 60 bp, MetaPhlAn2 becomes the classifier with the lowest number of extra taxa reported.

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 S3 & S4). 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 S3 & S4). 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 S3 & S4). As a result, the sensitivity improves substantially for longer reads for DIAMOND, especially for the Sensitivity_s&h while the precision decreases slightly (Figure 5).

MetaPhlAn2 is – with Centrifuge – the most stable classifier when varying read length with the classification categories “correct species” and “correct higher” remaining almost constant (Figure S3 & S4). The only summary statistics that changes is the proportion of “incorrect” reads which 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-strand 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 at 0.01 for all simulations, while we varied the different probabilities of single stranded deamination from 0 to 0.5 (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. Simulated deamination damage caused only a small reduction in the proportion of classified reads ("correct species", "correct higher" and "incorrect reads") in DIAMOND (Figure S3), almost no effect on Centrifuge and Kraken2 and no effect on MetaPhlAn2. The slight reduction of "correct species" and "correct higher" reads leads to a hardly noticeable decrease of the sensitivities, DIAMOND being the most and MetaPhlAn2 being the least affected (Figure 6A). Similarly, the precision of all classifiers remains constant for all values of the single-stranded deamination probabilities (Figure 6B). In this case, for DIAMOND, despite the reduction in the number of classifications, the number of "incorrect" reads is small relative to the number of correct classifications, thus the precision remains stable.

The deamination damage has no effect, as well, in the number of correctly detected viruses (among the 120 viral sequences tested) (Figure 6C). As above, Centrifuge, Kraken2 and DIAMOND detect 119 viral sequences and MetaPhlAn2 detects 39 regardless of the amount of deamination.

Deamination only impacts the average number of spurious extra taxa for Centrifuge and Kraken2 increasing from 11 to 13 for Centrifuge and 9 to 11 for Kraken2 (Figure 6D). In contrast, for DIAMOND and MetaPhlAn2, the average number of spurious extra taxa seems essentially constant for all deamination probability values.

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 of the errors 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 12.6-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, MetaPhlAn2 is essentially not affected by the sequencing error across simulations with results for a 12.6-fold increase in error rate essentially identical to the ones at 1-fold across all statistics (Figure 7, Figures S3 & S4).

The trends observed for Centrifuge, Kraken2 and DIAMOND are exacerbated with Illumina-like sequencing error compared to damage. This suggests that when errors are mostly concentrated at the read termini (such as the ones observed for ancient DNA), they have a smaller impact than if more substitutions of any kind (not just C to T and G to A) are distributed more evenly across the read. The most striking effect overall is that much fewer reads are classified with increasing error rate (Figures S3 & S4). The proportion of “correct species”, “correct higher” and “incorrect” reads decrease with increased error rate (Figure 7).

This results in an almost linear decrease in sensitivity when compared to the fold increase in error for Centrifuge, Kraken2 and DIAMOND (Figure 7). The reason for this is that the absolute number of correctly classified ("correct species” and “correct higher”) is high relative to “incorrect” reads such that the precision remains stable with 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, but for DIAMOND, which does not detect the virus Orbivirus SX-2017a (NC_033788.1) at species level for qShift values of -1 and -3. The virus is however detected at higher level.

Moreover, as for deamination, the number of spurious extra taxa increases substantially for Centrifuge and Kraken2 with increased number of substitutions from 10.3 (1-fold increase in substitutions) to 16.9 (12.6-fold increase) for Centrifuge and from 6.8 to 13.8 for Kraken2. The effect on DIAMOND is also remarkable in this case, as the average number of spurious extra taxa is much lower (~4.6) and does not increase with substitutions. For MetaPhlAn2 the increased number of substitutions has also no effect on the mean number of spurious extra taxa with a mean number of 6.3, a bit higher than for DIAMOND.

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.

Centrifuge and Kraken2 classify the vast majority of the reads as human (>99.1%) while most reads are unclassified for DIAMOND and MetaPhlAn2 (Figure 8A). The large proportions of unclassified reads in DIAMOND and MetaPhlAn2 are expected since neither of them include human sequences in their databases. Nevertheless, a small proportion of reads (0.005% for Centrifuge, 0.006% for Kraken2, 0.004% for DIAMOND and 0.015% for MetaPhlan2 (Figures 8A & 8B) were not classified as human (or as a taxon within the human lineage) and thus mainly consist of microbial classifications, including viruses. For Centrifuge and Kraken2 a large proportion (~95%) of the misclassified reads is assigned to the root (the highest taxonomic rank possible). Interestingly, the number of spurious extra taxa varied greatly across classifiers and did not correlate with the number of reads classified as non-human (Figure 8C). Hence, Kraken2 had the largest number of spurious extra taxa with 1,246 identified by at least 1 read, followed by Centrifuge (669), DIAMOND (168) and MetaPhlan2 (91). When analysed in more detail, we found that the misclassified reads are assigned to different superkingdoms present in the classifiers' database (Table 1) with only viruses for DIAMOND; viruses, archaea and eukaryotes for MetaPhlan2; and bacteria, archaea and viruses for Centrifuge and Kraken2 (Figure 9). Moreover, most reads (over 75% across classifiers) are given a very specific assignment since they are classified at the species level. Table S1 contains a list of each taxon identified by at least one read for each classifier. We observed little overlap in assignment between classifiers in the identified taxa. Moreover, the number of reads per taxa varies greatly across classifiers. For Centrifuge, up to five reads per taxa are identified. For Kraken2, it is up to 15 reads. DIAMOND's and MetaPhlan2's lists contain a few taxa with large number of reads (several thousand for DIAMOND and several hundred for MetaPhlAn2, Figure 11).

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 for ancient DNA studies, we compared state-of-the art classifiers under controlled conditions, i.e. in silico simulations. To do so, we randomly selected 120 published viral sequences, 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. 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.

119 out of 120 viral sequences recovered by Centrifuge, Kraken2 and DIAMOND

In most simulations with reads over 50 bp long, three of the classifiers, namely Centrifuge, Kraken2 and DIAMOND, successfully detect at the species level 119 out of the 120 simulated viral sequences. 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. Note also that the one virus missed by Centrifuge and Kraken2 is correctly identified at a higher taxonomic level for those classifiers. For DIAMOND, the missed viral sequence does not encode for a protein included in the RefSeq viral protein database.

In contrast, in all simulations, MetaPhlAn2 did not detect two thirds (81/120) of the tested viruses. This could be explained by MetaPhlAn2's database, which may lack clade specific markers for a large proportion of the tested viruses.

Kraken2 and DIAMOND do not classify 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. 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.

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 47% 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). It is unclear to us why MetaPhlAn2's precision is considerably lower than the other classifiers even when considering the "correct higher" category.

A high Precision_s&h 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 database 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 the highest Precision_s&h. 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 and Kraken2 (see below). 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 -11), deamination has little effect, while sequencing error reduces substantially the sensitivities of Centrifuge and Kraken2. This suggests having errors mostly concentrated at the end of the reads (ancient DNA-like errors) likely decrease their impact compared to higher numbers of any kind of substitution errors more evenly distributed (Illumina-like errors) (Pfeiffer et al. 2018).

Centrifuge and Kraken2 have the largest number of spurious taxa

Across all simulations with reads above 30 bp, DIAMOND achieved the lowest fraction of misclassified reads across viruses and Centrifuge and Kraken2 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. Centrifuge has the highest number of spurious extra taxa for most simulations, followed by Kraken2 and, depending on the read length, DIAMOND or MetaPhlAn2. This could be explained by the databases of each classifier. Centrifuge and Kraken2 depend on whole genome DNA databases that include several superkingdoms and noise could in principle make a simulated read look like another taxon by chance only. DIAMOND could be more robust to noise as it depends on a protein database and the database only contains viruses. This hypothesis is supported by the observation that deamination (or sequencing error) leads to an increase in the number of spurious extra taxa in an almost linear way for Centrifuge and Kraken2. Finally, longer reads considerably reduce the number of spurious extra taxa for Centrifuge. This could be explained by Centrifuge's taxonomic binning strategy with scores favouring longer hits.

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 - even for classifiers whose databases include human DNA - 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 the retroviruses identified by DIAMOND. In this case, retroviruses are among the top hits (i.e. the taxa with the highest number of reads Table S1). For such viruses, their genomes in their proviral phase insert in their host genome (endogenous retroviruses are the result of this kind of events, (Belshaw et al. 2004)). As a result, some of the human genomes assemblies have proviruses inserted making the identification of genuine retrovirus infections very difficult.

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 is generally followed by extensive work. These steps generally include the mapping of sequenced data to individual candidate genomes identified by classifiers (Mühlemann, Jones, et al. 2018; Mühlemann, Margaryan, et al. 2018). 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 has a slightly higher precision but has a considerably lower sensitivity. Moreover, Centrifuge, Kraken2 and DIAMOND recover essentially all the simulated viruses for most simulations. Nevertheless, short reads cannot be handled by either Kraken2 or DIAMOND. Considering all those results together, Centrifuge is 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 most simulations. Hence, for studies with longer fragment lengths, DIAMOND and Kraken2 could be a good choice to reduce the number of candidates for further downstream analyses. MetaPhlAn2 did not perform well in our analyses. While MetaPhlAn2 has advantages, such as reporting a limited number of spurious extra taxa (Figure 5), the large number of viruses that would be missed in ancient virome studies suggest it is not well-suited for such analyses. If researchers are after specific viruses, the users should first verify that clade specific markers for the virus of interest are included in MetaPhlAn2's database.

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 the tree of life that 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).

Our simulations show that the results vary a lot across viruses in a way that also depends on the classifier. Some viruses are easy to identify for some classifiers but impossible or hard for others. The reasons for this are likely multifactorial and could include the characteristics of the classifier and associated database, the structure of viral genomes or their evolutionary history. Hence, if the intention is to answer a specific biological question and to identify a given set of viruses, we recommend performing a simulation study to identify the best possible tool prior to the analyses.

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 outperforms the other classifiers as it can handle short fragments and as it has a high sensitivity and precision 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 the sequence search algorithms (e.g. by reducing the length of the k-mer or seed) so the classifiers can handle shorter reads while maintaining their performance for longer 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

Alfson KJ, Beadles MW, Griffiths A. 2014. A new approach to determining whole viral genomic sequences including termini using a single deep sequencing run. *J. Virol. Methods* 208:1–5.

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.

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

Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M. 2004. Long-term reinfection of the human genome by endogenous retroviruses. *Proc. Natl. Acad. Sci. U. S. A.* 101:4894–4899.

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.

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.

Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S. 2001. Ancient DNA. *Nat. Rev. Genet.* 2:353–359.

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

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9:357–359.

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.

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.

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.

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.

Classifier	Database					Query algorithm	Taxonomic binning strategy
	Name	Organisms included	Loci	Molecule	Size		
Centrifuge	RefSeq	bacteria, archaea, viruses, and human	Whole genomes	DNA	24G	Exact alignment (no mismatches or gaps)	Score that favours longer hits
Kraken2	Refseq + UniVec_Core	bacteria, archaea, viruses, human, and (vectors, adapters, linkers, and primers)	Whole genomes	DNA	35G	Exact k-mer matching	Highest number of k-matches considering a root-to-leaf path
DIAMOND	RefSeq of viral proteins	viruses	Protein coding regions	Proteins	117M	Alignment using spaced seeds	Lowest common ancestor
MetaPhlAn2	Custom clade-specific markers	bacteria, archaea, viruses, and eukaryotes	Clade specific markers	DNA	1.2G	Alignment with BowTie2: ungapped seed match followed by extension	Clade specific marker

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 at higher taxonomic rank 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.

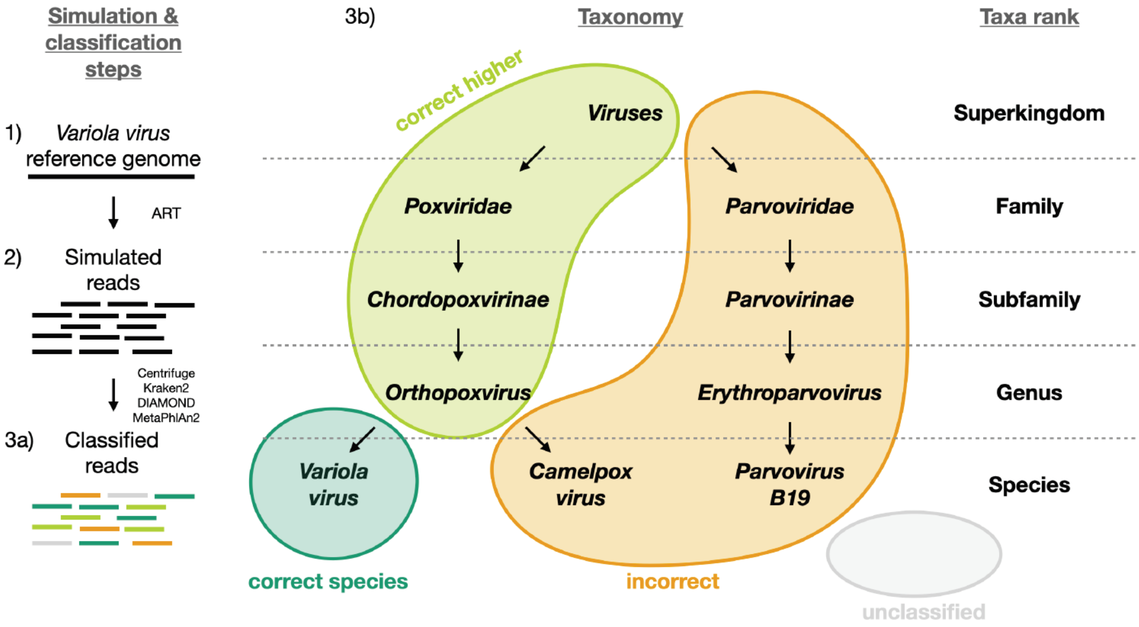


Figure 2. Observed deamination damage across the reads.

Observed frequency of substitutions across the reads simulated from the viral sequence NC_001338.1 (*Sulfolobus virus*) 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).

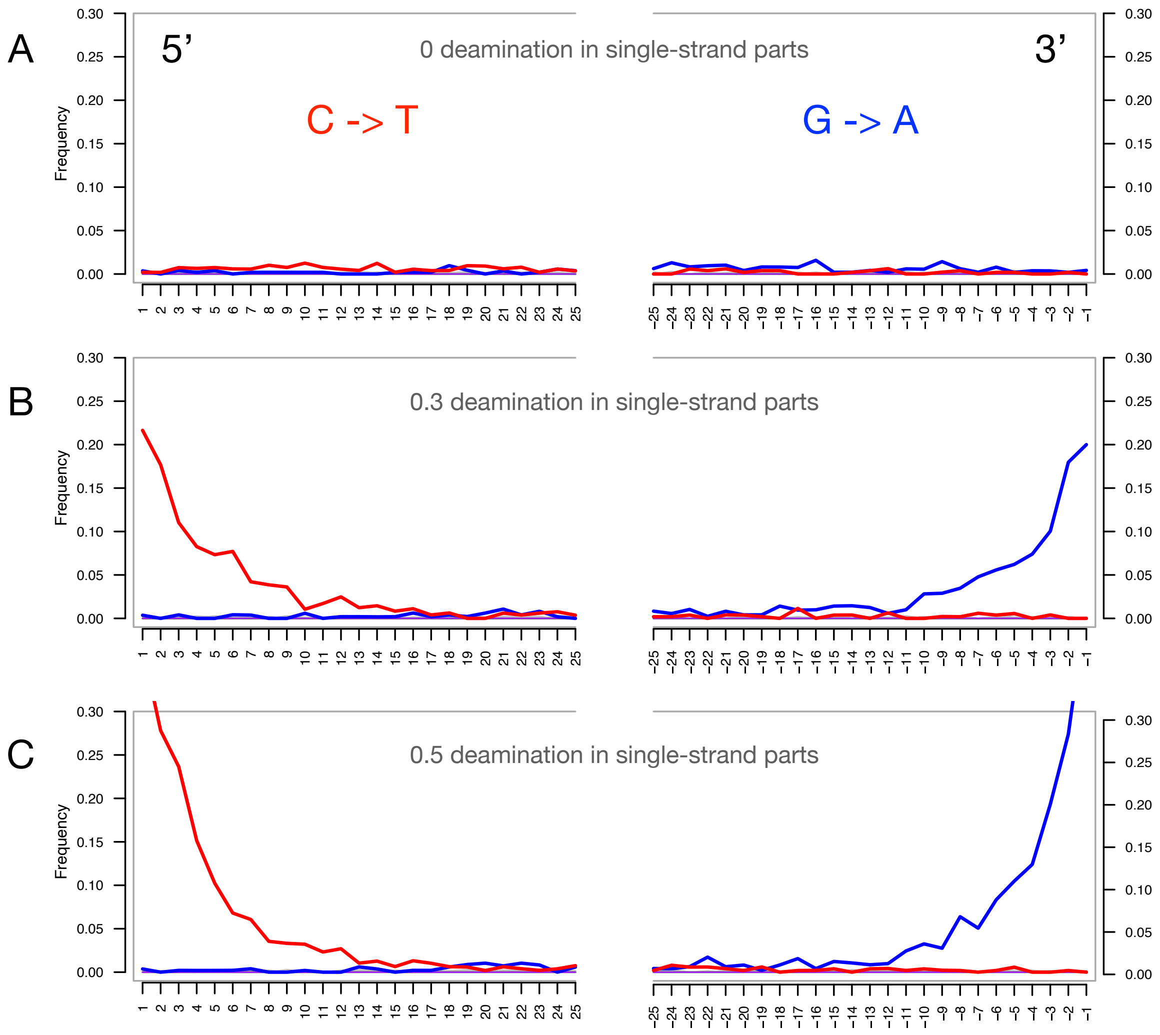


Figure 3. Average number of substitutions for the deamination and sequencing error simulations.

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_007555.1 (Adult diarrheal rotavirus strain J19 complete genome) with 150 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 -11 which corresponds to a 1 fold (*qShift* of 0) to 12.6 fold (*qShift* of -11) increase compared to standard Illumina 2500 sequencing (Huang et al. 2012).

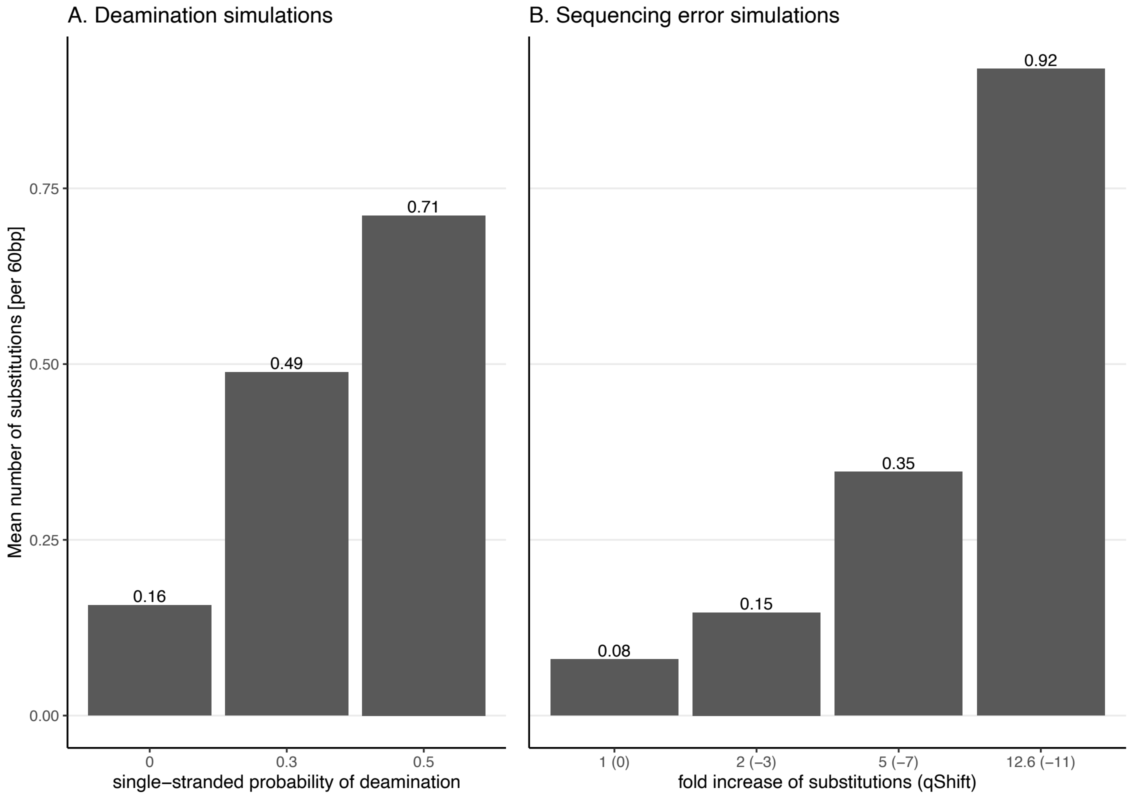


Figure 4. Classification results for the 60 bp read set.

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 (Materials and Methods). An ideal classifier would have 100% sensitivity and 100% precision. B) Total number of viral sequences recovered for each classifier when correctly identifying at least 1, 5, 10, 20 and 50 reads per simulated viral sequence. The dashed line indicates the total number of tested viral sequences (120). C) Mean number of spurious extra taxa per classifier. In this plot, a taxon is assumed identified by a classifier if at least 1, 5, 10, 20 or 50 reads are assigned to this taxon.

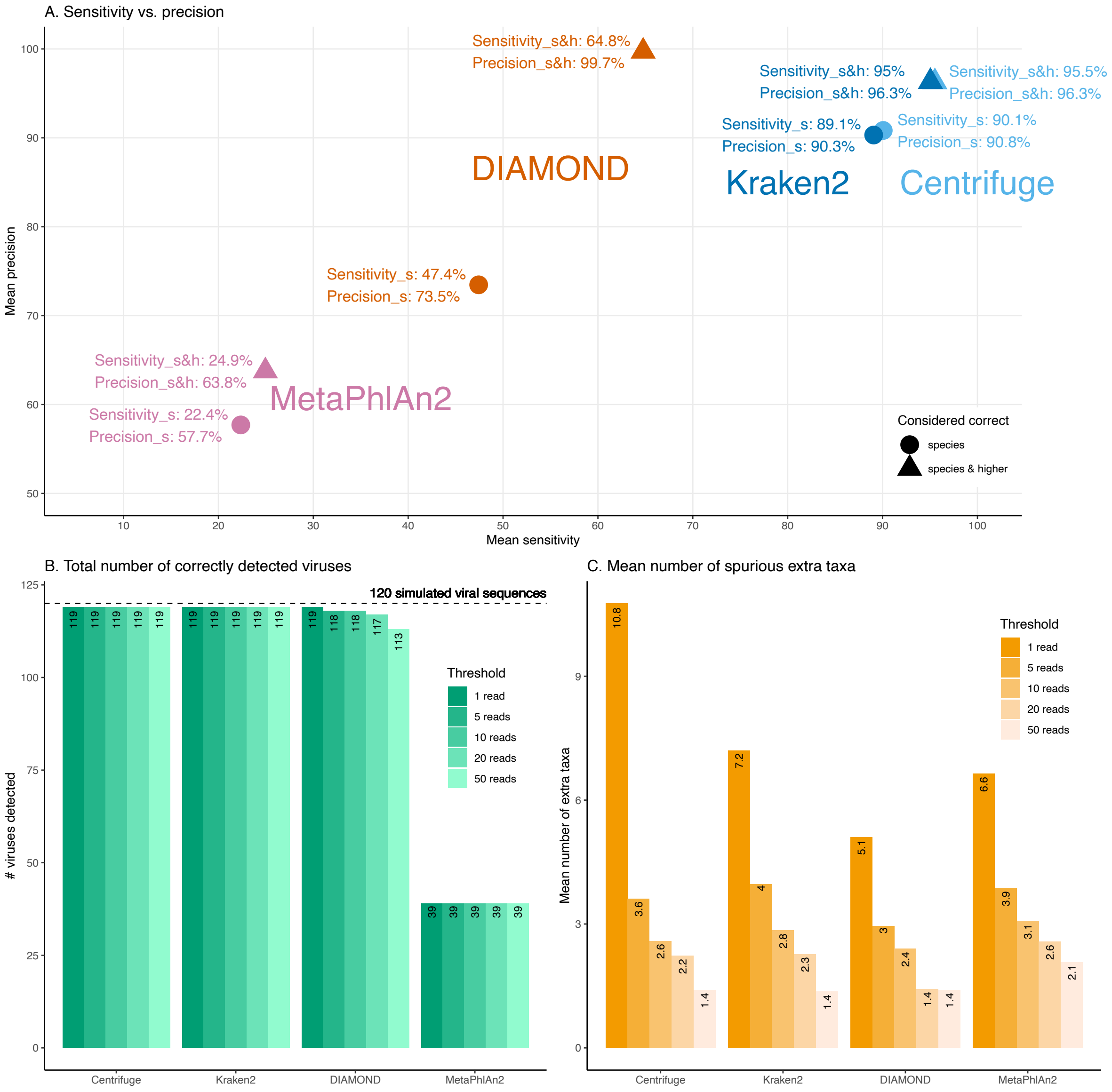

A. Sensitivity vs. precision
Mean precision
Mean sensitivity
100
90
80
70
60
50
10
20
30
40
50
60
70
80
90
100
Sensitivity_s&h: 64.8%
Precision_s&h: 99.7%
DIAMOND
Sensitivity_s: 47.4%
Precision_s: 73.5%
Sensitivity_s&h: 95%
Precision_s&h: 96.3%
Sensitivity_s: 89.1%
Precision_s: 90.3%
Kraken2
Sensitivity_s&h: 95.5%
Precision_s&h: 96.3%
Sensitivity_s: 90.1%
Precision_s: 90.8%
Centrifuge
Sensitivity_s&h: 24.9%
Precision_s&h: 63.8%
MetaPhlAn2
Sensitivity_s: 22.4%
Precision_s: 57.7%
Considered correct
● species
▲ species & higher
B. Total number of correctly detected viruses
# viruses detected
125
100
75
50
25
0
120 simulated viral sequences
119
119
119
119
119
119
119
119
119
119
119
118
118
117
113
39
39
39
39
39
Threshold
1 read
5 reads
10 reads
20 reads
50 reads
Centrifuge
Kraken2
DIAMOND
MetaPhlAn2
C. Mean number of spurious extra taxa
Mean number of extra taxa
9
6
3
0
10.8
3.6
2.6
2.2
1.4
7.2
4
2.8
2.3
1.4
5.1
3
2.4
1.4
1.4
6.6
3.9
3.1
2.6
2.1
Threshold
1 read
5 reads
10 reads
20 reads
50 reads
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 120 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.

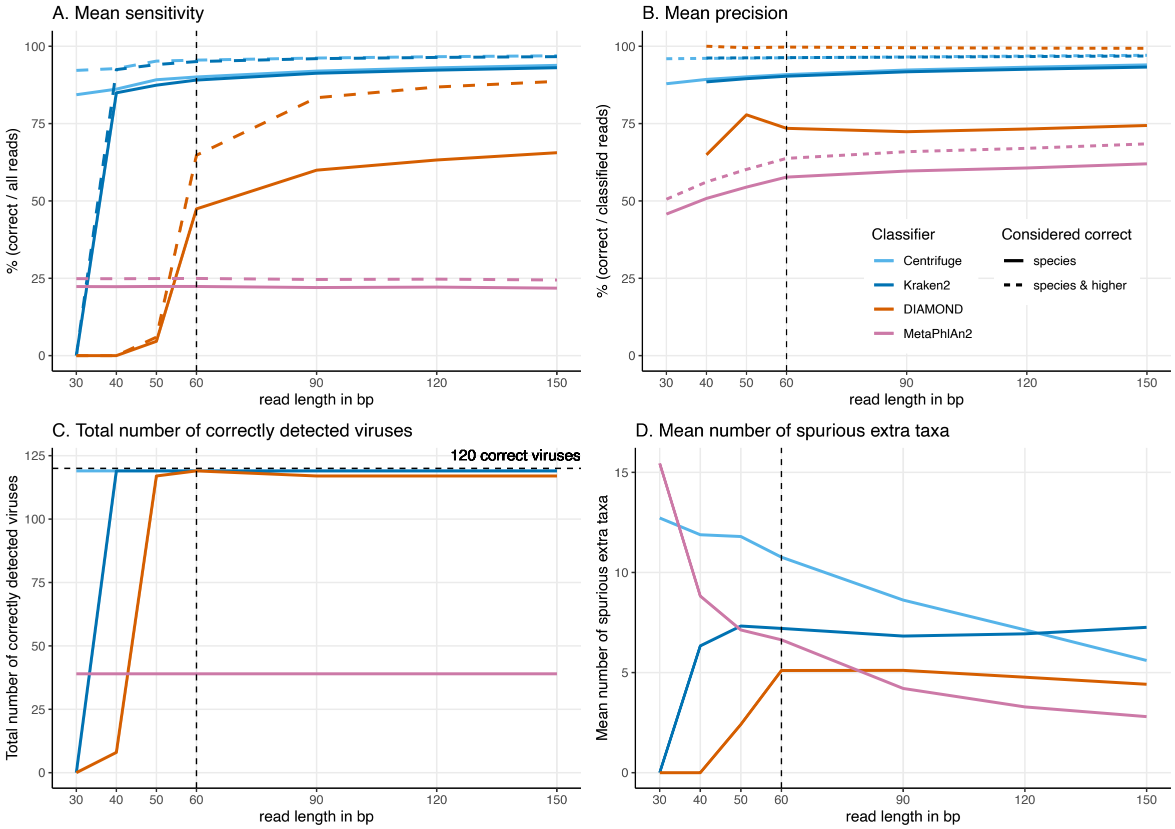


Figure 6. Effect of the deamination damage on the classification performance.

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 120 tested. The dashed line shows the maximum number of detectable viruses. D) Average number of spurious extra taxa across simulated viral sequences.

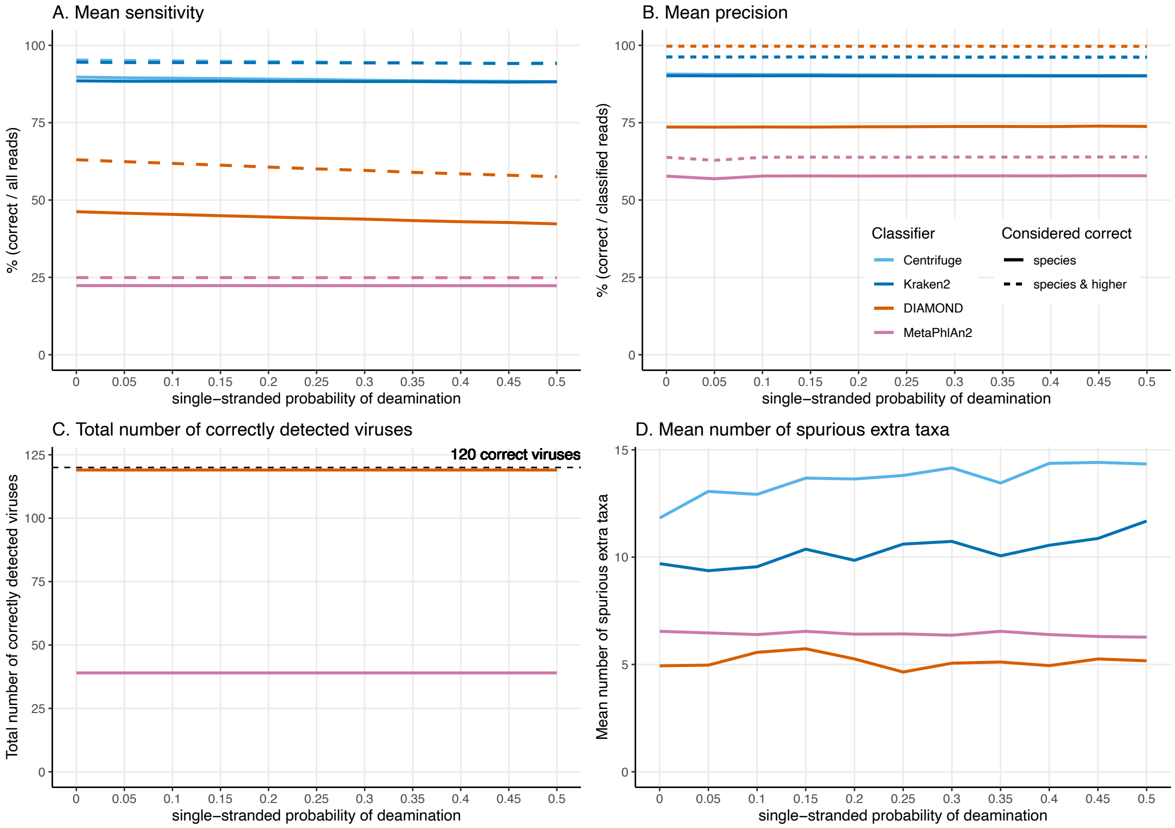


Figure 7. Effect of the substitution sequencing error on the classification performance.

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 12.6-fold (qShift values from 0 to -11). 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 120 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.

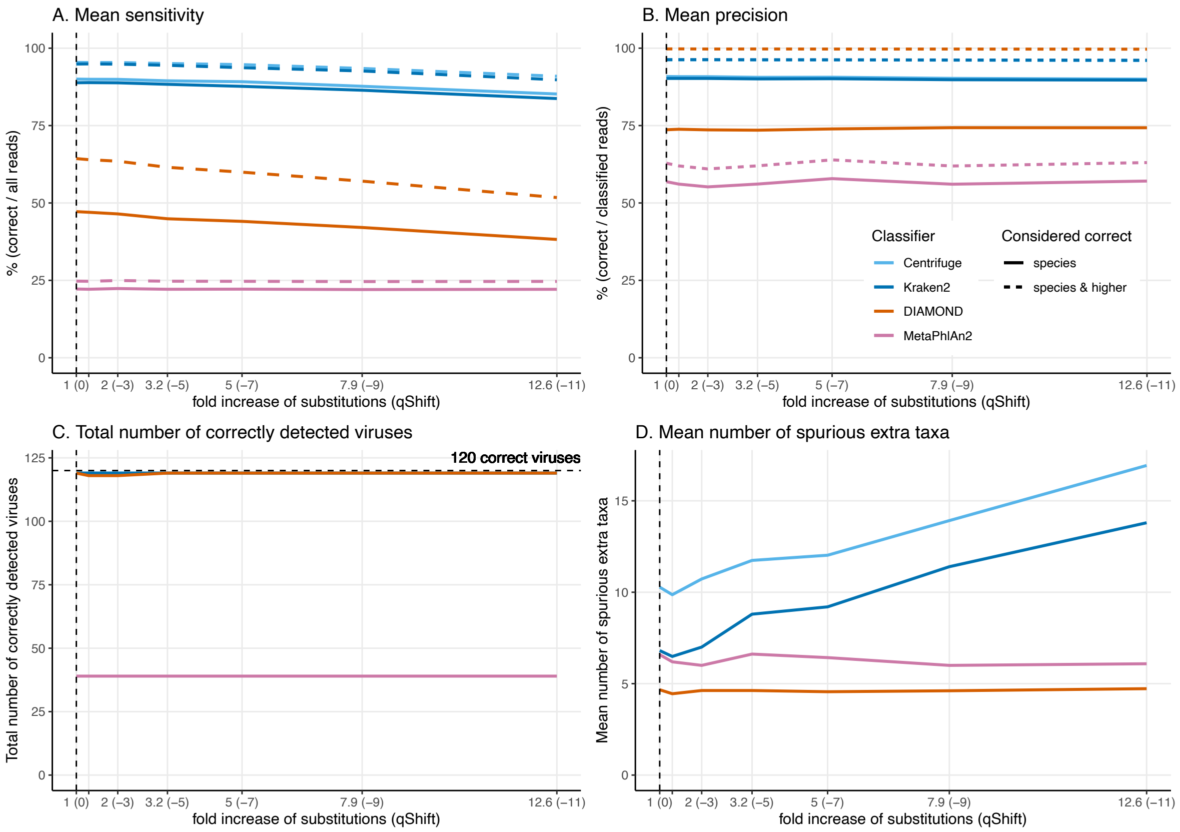


Figure 8. Classification of human reads.

A) Proportion of reads for each classification category (Figure 1). Note that the databases used for DIAMOND and MetaPhan2 do not include human in contrast to Centrifuge's and Kraken2's databases (see Table 1). B) Total number of incorrectly classified reads. C) Number of spurious extra taxa reported by each classifier.

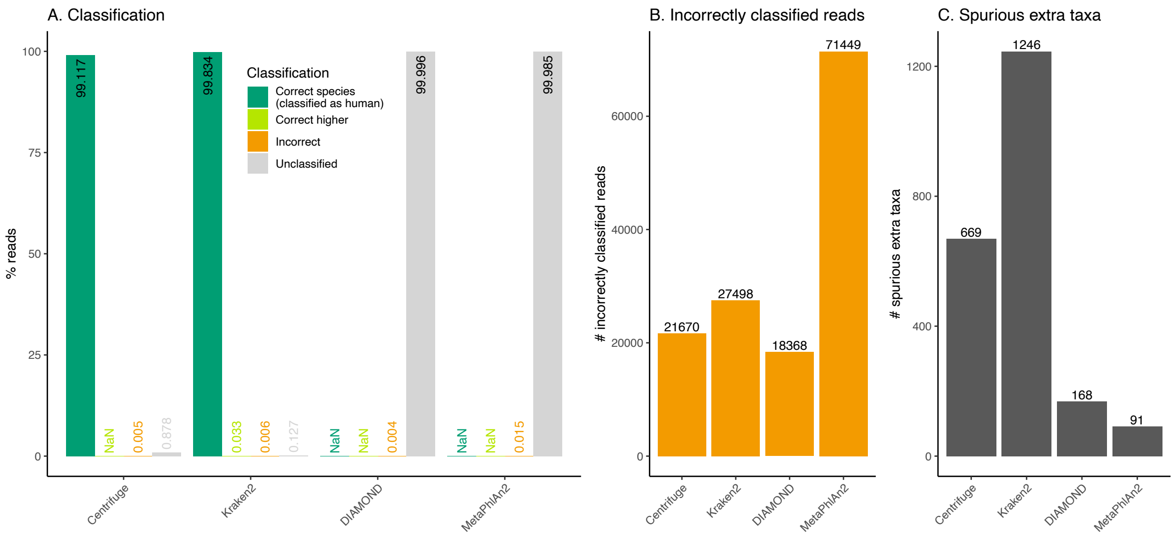


Figure 9. Taxonomic ranks represented in the classification of human reads.

Number of reads and the proportion of reads per classifier (bar plots). Proportion of taxa classified as bacteria, archaea, viruses and eukaryotes (pie charts).

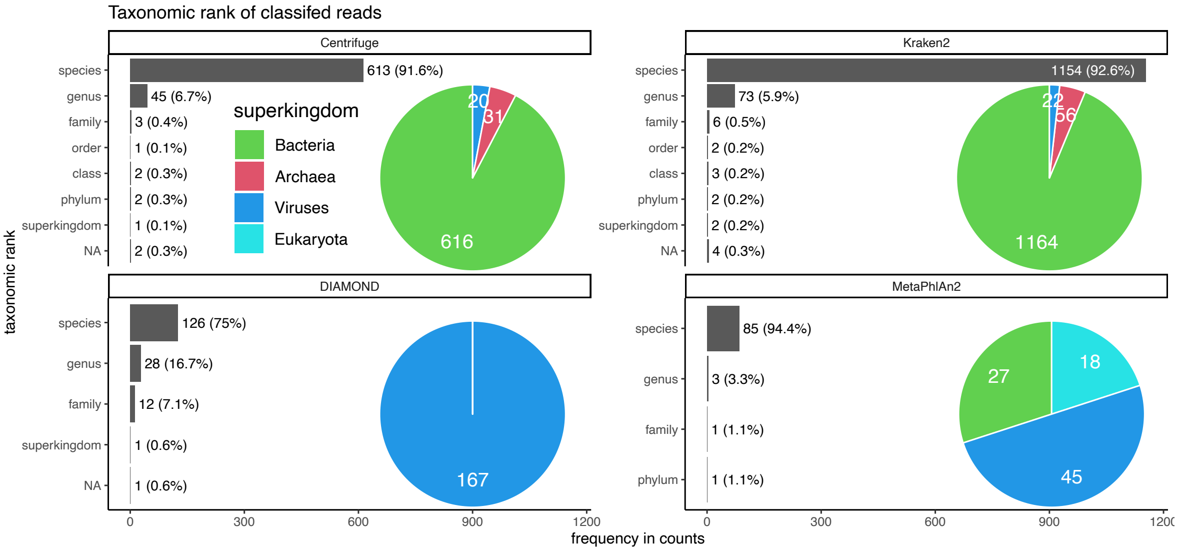


Table S1 Extra taxa identified in the human simulations.

See excel spread sheet.

Figure S1. Sequence length distribution of the selected viral sequences.

Length distribution of the 120 viral sequences randomly selected for simulations (blue histogram) and for all the 10,544 viral sequences in the RefSeq database (yellow histogram).

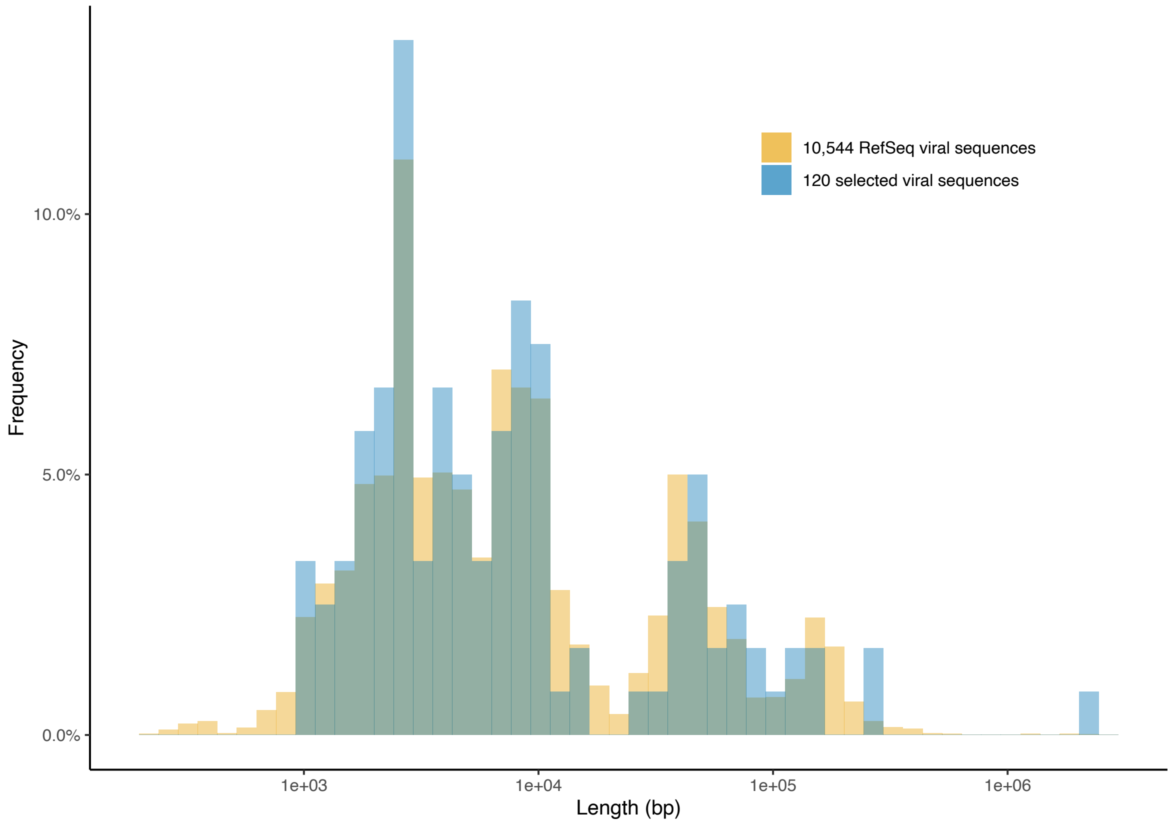


Figure S2. Classification results for the 60 bp read set.

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 120 viral sequences selected for the simulations. Note that the viral sequences are not ordered the same way for eac h subplot. B) Means over the 120 viral sequences for each classification category.

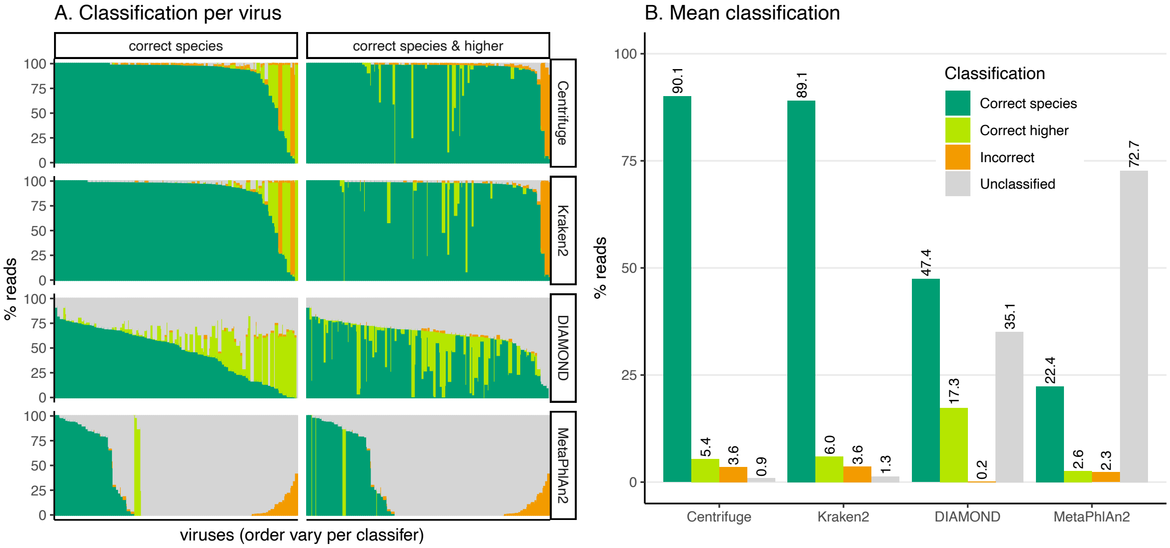


Figure S3. Average classification categories for all simulations (stacked bar plots).

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).

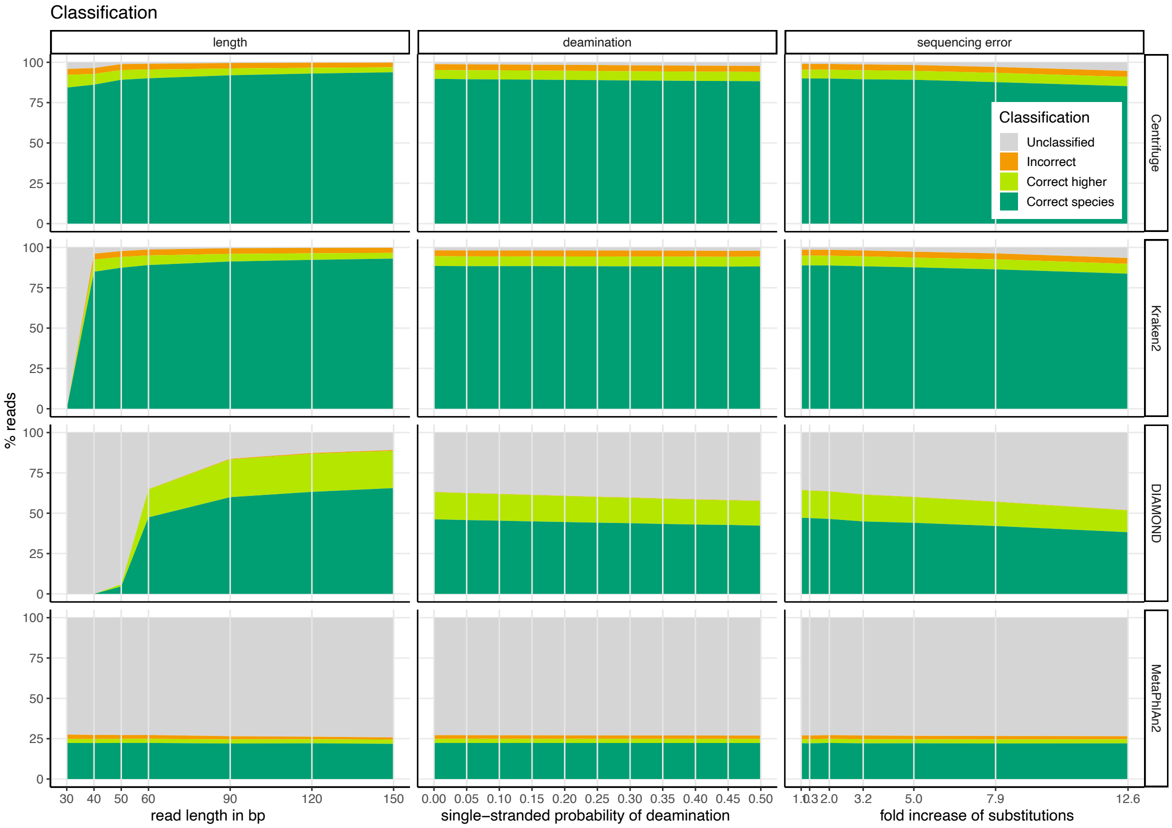
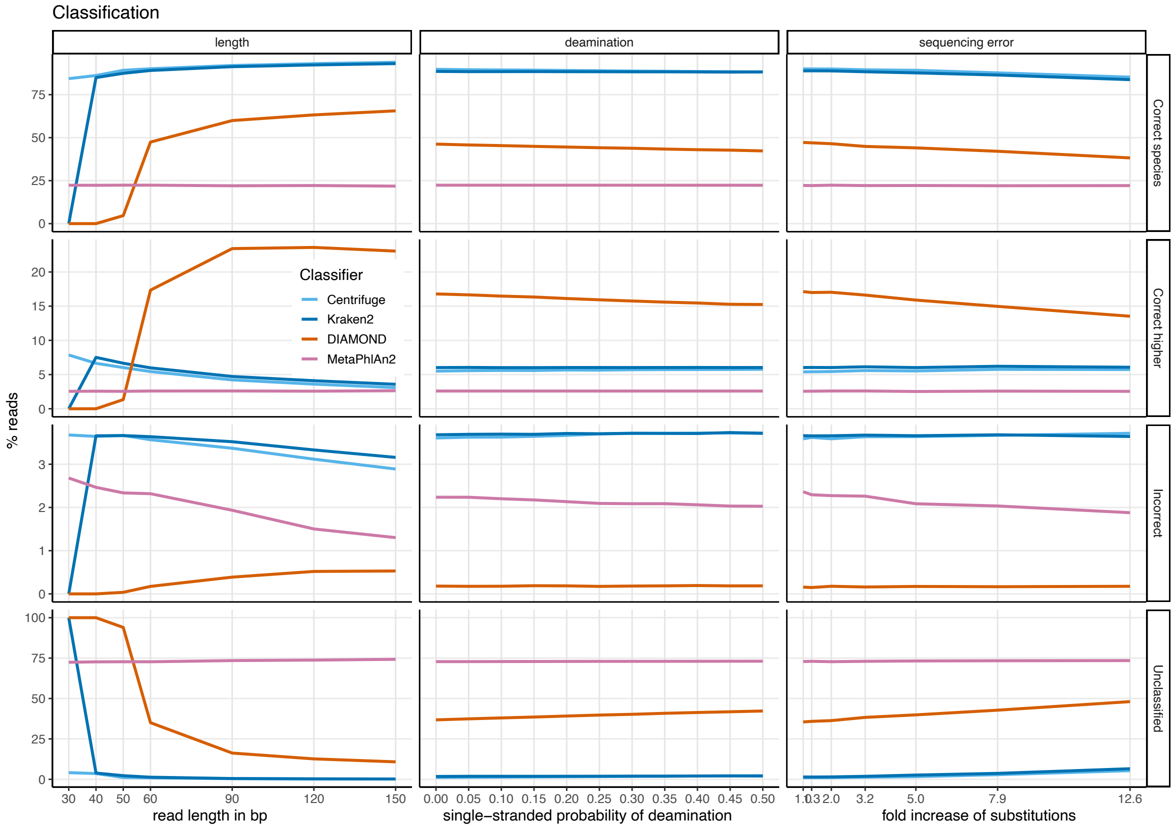


Figure S4. Average classification categories for all simulations (line plots).

Nine 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).


Classification
length
deamination
sequencing error
Correct species
Correct higher
Incorrect
Unclassified
% reads
Classifier
Centrifuge
Kraken2
DIAMOND
MetaPhlAn2
read length in bp
single−stranded probability of deamination
fold increase of substitutions