Benchmarking metagenomics classifiers on ancient viral DNA: a simulation study

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

[1] Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

[2] Swiss institute of Bioinformatics, Lausanne, Switzerland

[3] Vital-IT, Swiss Institute of Bioinformatics, Lausanne, Switzerland

Corresponding Authors: Samuel Neuenschwander, Anna-Sapfo Malaspinas
Email address: samuel.neuenschwander@unil.ch, annasapfo.malaspinas@unil.ch

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.

1 # Benchmarking metagenomics classifiers on ancient

2 # viral DNA: a simulation study

4 Yami Ommar Arizmendi Cárdenas[1,2], Samuel Neuenschwander[1,3]*,

5 Anna-Sapfo Malaspinas[1,2]*

9 [1] Department of Computational Biology, University of Lausanne, 1015, Lausanne,

10 Switzerland

11 [2] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

12 [3] Vital-IT, Swiss institute of Bioinformatics, 1015 Lausanne, Switzerland

24 *Corresponding authors

25 Samuel Neuenschwander

26 Email: samuel.neuenschwander@unil.ch

28 Anna-Sapfo Malaspinas

29 Email: annasapfo.malaspinas@unil.ch

Abstract

31   Owing to technological advances in ancient DNA, it is now possible to sequence

32   viruses from the past to track down their origin and evolution. However, ancient DNA data

33   is considerably more degraded and contaminated than modern data making the

34   identification of ancient viral genomes particularly challenging. Several methods to

35   characterise the modern microbiome (and, within this, the virome) have been developed; in

36   particular, tools that assign sequenced reads to specific taxa in order to characterise the

37   organisms present in a sample of interest. While these existing tools are routinely used in

38   modern data, their performance when applied to ancient microbiome data to screen for

39   ancient viruses remains unknown.

40   In this work, we conducted an extensive simulation study using public viral

41   sequences to establish which tool is the most suitable to screen ancient samples for human

42   DNA viruses. We compared the performance of four widely used classifiers, namely

43   Centrifuge, Kraken2, DIAMOND and MetaPhlAn2, in correctly assigning sequencing reads

44   to the corresponding viruses. To do so, we simulated reads by adding noise typical of

45   ancient DNA to a set of publicly available human DNA viral sequences and to the human

46   genome. We fragmented the DNA into different lengths, added sequencing error and C to T

47   and G to A deamination substitutions at the read termini. Then we measured the resulting

48   sensitivity and precision for all classifiers.

49   Across most simulations, more than 228 out of the 233 simulated viruses are

50   recovered by Centrifuge, Kraken2 and DIAMOND, in contrast to MetaPhlAn2 which

51   recovers only around one third. Overall, Centrifuge and Kraken2 have the best performance

52   with the highest values of sensitivity and precision. We found that deamination damage has

53   little impact on the performance of the classifiers, less than the sequencing error and the

54   length of the reads. Since Centrifuge can handle short reads (in contrast to DIAMOND and

55   Kraken2 with default settings) and since it achieves the highest sensitivity and precision at

56   the species level across all the simulations performed, it is our recommended tool.

57   Regardless of the tool used, our simulations indicate that, for ancient human studies, users

58   should use strict filters to remove all reads of potential human origin. Finally, we

59   recommend to verify which species are present in the database used, as it might happen that

60   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'

91   end of the fragments and guanines to adenines (G to A) substitutions at the 3' ends of the

92   fragments for double stranded libraries (Briggs et al. 2007; Carøe et al. 2018).

93   In most microbiome aDNA studies the first step conducted is to assign a taxon to

94   each of the sequenced reads using a classifier (i.e. classify the reads). The classification of

95   the reads results in a list of microbes that are potentially ancient having infected the host

96   prior to the death of the sampled organism. From this list of potential candidates some

97   microbes (usually pathogens) are selected for follow up analyses to corroborate if they are

98   truly ancient microbes and "endogenous". The classification of reads and selection of

99   candidates is known as the screening or scanning step (Warinner et al. 2017). The steps

100   following the screening are computationally intensive, for instance, they generally include a

101   mapping step (BWA (Li and Durbin 2009), BowTie2 (Langmead and Salzberg 2012)) and

102   a blasting step (Altschul et al. 1990) to determine if a particular genome is the most likely

103   source of the sequenced reads (Krause-Kyora et al. 2018; Mühlemann, Jones, et al. 2018;

104   Vågene et al. 2018; Mühlemann et al. 2020).

105   For the screening step there are many options of classifiers available. The classifiers

106   are diverse and differ from each other in the underlying database, sequence search

107   algorithm, and taxonomic binning (the process of assigning a taxonomic rank to each of the

108   reads sequenced from a sample). In the specific case of ancient microbiome studies,

109   benchmarks have been conducted to choose an appropriate classifier (Velsko et al. 2018) or

110   to determine the most adequate parameters to enhance the performance of specific

111   classifiers (Eisenhofer and Weyrich 2019). Nevertheless, these benchmark studies only

112   focus on the bacterial fraction of the microbiomes. Considering this, we have decided to

113   perform a benchmark study to assess the performance of different classifiers in screening

114   ancient microbiome samples for viruses; specifically, our study focuses on DNA viruses

115   with human as their host (human DNA viruses), since most of the ancient viral DNA

116   studies published so far target such viruses. In short, we classified simulated samples with

117   four classifiers: Centrifuge, Kraken2, DIAMOND and MetaPhlAn2 (for a fifth classifier:

118   MALT, see Supplementary Note 1). The classified samples consisted of simulated ancient-

119   like DNA reads, with different length, deamination damage and sequencing error values.

120   These samples represent an ideal case scenario as they come from a single human DNA

121   viral species and have a high coverage. As we knew the true species in the sample, we

122   scored the classifications per sample; we evaluated the effect of read length, deamination

123   and sequencing error; and we compared the overall performance of all classifiers. In

124   addition, we evaluate the effect of human contamination by classifying simulated human

125   reads with the four classifiers.

126   The four classifiers assessed in the present study: Centrifuge (Kim et al. 2016),

127   Kraken2 (Wood et al. 2019), DIAMOND (Buchfink et al. 2015) and MetaPhlAn2 (Truong

128   et al. 2015) differ on their approaches. These classifiers were not specifically developed for

129   aDNA; fragmentation and cytosine deamination may present a challenge for the taxonomic

130   assignment since reads with post-mortem alterations could lead to misclassifications. In the

131   subsequent paragraphs we will describe the characteristics in terms of database, algorithms

132   and taxonomic binning of these four classifiers (see Table 1 for a summary).

133   The four classifiers databases are characterised by the type of molecule and the

134   genomic region(s) that are included. Among the four classifiers, two different types of

135   molecules are used: DNA and protein sequences. In terms of genomic regions, the

136   databases either include multiple loci from each organism (such as a set of proteins or

137   marker genes) or whole genomes. Centrifuge uses whole genome DNA databases. Kraken2

138   is versatile in database usage as it can use protein or DNA, whole genome or single locus

139   databases. DIAMOND uses a protein database and can use as query both DNA or amino

140   acid sequences. MetaPhlAn2 relies on a multiple loci DNA database consisting of core

141   genes that are shared within a clade but not outside of it (clade specific marker genes). In

142   contrast to its previous version, MetaPhlAn2's database includes marker genes for viruses

143   and eukaryotic microbes.

144   For the sequenced search algorithm, classifiers either rely on alignments or on exact

145   k-mer matches. Centrifuge, DIAMOND and MetaPhlAn2 depend on alignment algorithms.

146   Centrifuge uses the Burrows-Wheeler Transform (Burrows and Wheeler 1994) and

147   Ferragina-Manzini index (Ferragina and Manzini 2000). These algorithms allow to create a

148   data structure which facilitates fast alignments with efficient memory usage. Centrifuge

149   works by searching for exact matches (i.e. no mismatches or gaps are allowed). DIAMOND

150   implements double index alignment (both query and reference are indexed), looking for

151   matches of seeds (short subsequences of fixed length, with default lengths of 15-24 bp) and

152   then extending the alignment. DIAMOND's seeds are spaced seeds, meaning that some

153   positions in the seeds are treated as wildcards. Finally, MetaPhlAn2 performs alignments

154   using BowTie2 (Langmead and Salzberg 2012). BowTie2 is an index-assisted aligner that

155   allows gaps in the alignment. Its algorithm has two phases: ungapped seed match, and an

156   extension that permits gaps. Kraken2 is in contrast an alignment-free classifier which is

157   based on exact matches of k-mers (sequence substrings of length k) (Wood and Salzberg

158   2014). Kraken2 builds an index of k-mers (of length 35 by default for nucleotide

159   sequences) from the database. K-mers are first obtained from the query read and these k-

160   mers are then looked up in the database (an exact k-mer match is performed).

161   Finally, the classifiers use different strategies for taxonomic binning. DIAMOND

162   uses the lowest common ancestor (LCA) algorithm: that is, if a read matches several

163   species, it will be assigned to the most specific ("lowest") taxonomic rank shared by all the

164   matching species. If the read only matches one species, the taxonomic assignation will

165   remain at the level of that species. As a consequence, conserved sequences will be assigned

166   to higher taxonomic levels, and specific sequences will be assigned to species level (Huson

167   et al. 2007). Centrifuge and Kraken2 perform taxonomic binning using scoring schemes.

168   Centrifuge ranks the alignments done with a score that favours longer hits. The query read

169   is attributed to the taxon in the database with the highest score. Similarly, Kraken2 assigns

170   a read to the taxon which has the most k-mer matches in common. In this case, a root-to-

171   leaf path is used to sum k-mers coming from higher taxonomic ranks. MetaPhlAn2 relies

172   on a clade specific marker catalogue. If there is an alignment to a marker, the read is

173   assigned to the taxonomic clade associated with that marker. A clade can be as specific as a

174   strain, or as broad as a phylum (Segata et al. 2012).

Materials & Methods

176   In short, the analyses consisted in simulating reads from reference viral sequences,

177   classifying them with four widely used classification tools and quantifying the performance

178   of the classifiers under different conditions. In particular, we investigated how varying the

179   read lengths, adding DNA damage (features typical of ancient DNA) and sequencing error

180   impact the classifications made by the tools. Finally, we analysed the assignments made by

181   the classifiers on simulated short human reads.

Reference viral sequences

182
183   A total of 238 reference viral sequences from 233 different human DNA viruses were used
184   in this study. To get the aforementioned sequences we first queried the Viral-Host DB
185   (downloaded on 25 September 2019) (Mihara et al. 2016) for viruses that infect *Homo*
186   *sapiens*, obtaining a list of 1,315 viruses. We downloaded the reference sequences of such
187   viruses in genbank format from NCBI RefSeq (Brister et al. 2015; O'Leary et al. 2016) and
188   kept the sequences if their "molecule type" record was "DNA" or "ss-DNA", obtaining a
189   total of 250 DNA viral sequences. Finally, we removed duplicated sequences with the same
190   taxID, representing viruses with different isolates or alternative genomes assemblies; we
191   also removed the Human endogenous retrovirus K113 (HERV-K113) sequence
192   (NC_022518.1), as this virus is a provirus in the human reference genome (Turner et al.
193   2001). This gave us a total of 233 viruses with 238 different sequences (one of the viruses
194   is stored in RefSeq in six different contigs). A table with information about the viruses used
195   in this study is available on Table S1; it includes the following: virus name, sequence name,
196   sequence length, type of molecule (DNA or ss-DNA), type of genome (linear or circular),
197   sequence accession number, virus taxID and Baltimore group. For each selected viral
198   sequence, we generated a set of simulated reads as described below.

Simulation of viral reads and read length

199
200   ART (ART-MountRainier-2016-06-05, art_Illumina Q Version 2.5.8) (Huang et al.
201   2012) was used to simulate sequencing reads starting from each of the 233 human DNA
202   viruses' sequences as reference. By default, ART simulates reads with sequencing error
203   typical for the specified sequencer. The sequencing machine was set to Illumina HiSeq
204   2500 (*art_illumina -ss HS25*) for all simulations, and the parameter *qShift* (that controls the
205   amount of sequencing error, see below) was set to 0 for most of the simulations (but also to
206   different values when assessing the effect of sequencing error) resulting in typical Illumina
207   HiSeq 2500 error profiles. In an initial simulation, for each of the 233 viruses, single-end
208   reads of a fixed length were simulated emulating the short read length observed in typical
209   ancient DNA data, see e.g. (Green et al. 2008). First, the read length was set at 60 bp (*-l 60*)
210   and a coverage of 10× was requested (*-f 10*) resulting in 280 to 39,270 reads per virus,
211   depending on the viral sequence length. Second, the read length was varied from 30 to 150
212   bp while keeping the coverage at 10×. We set the coverage of 10× (an unrealistically high

213   number for ancient DNA data) as we wanted to compare the classifiers under ideal

214   conditions. As the reads are classified independently of each other, we do not expect that

215   the relative performance of the classifiers (point estimates) would change with a lower

216   depth of coverage.

217   Each set of reads from the 233 selected human DNA viruses was classified using

218   four different classifiers: Centrifuge, Kraken2, DIAMOND and MetaPhlAn2 (Table 1). The

219   parameters for each of the classifiers are specified below.

220   **Classification of reads**

221   As much as possible, we used the default parameters suggested by the developers of each

222   tool to build the databases and perform the classification as the performance of each tool is

223   likely optimal for those parameters. Otherwise the parameters are stated.

224   **Centrifuge**

225   Centrifuge version 1.0.3-beta was used. To build a custom viral database we downloaded

226   all the viral sequences stored in RefSeq on 12 September 2021 from

227   https://ftp.ncbi.nlm.nih.gov/refseq/release/viral/viral.${i}.1.genomic.fna.gz (${i} has values

228   from 1 to 3) and NCBI's taxonomy nodes information (files nodes.dmp and names.dmp)

229   from https://ftp.ncbi.nih.gov/pub/taxonomy/taxdmp.zip on 14 October 2021. We created

230   our own "conversion table" file required by Centrifuge which links the accession number to

231   its taxID using the following command of NCBI's E-utilities (Anon):

232   `esearch –db nuccore –query "${accession}[Accession]" | elink –target`

233   `taxonomy | efetch –format xml | xtract –pattern Taxon –element TaxId`,

234   where "${accession}" represents the accession number. We then created the database with

235   the following command: `centrifuge-build --conversion-table`

236   `my_viral_accession2taxid.map --taxonomy-tree nodes.dmp --name-table`

237   `names.dmp all_viral.genomic.fna Centrifuge_viral`, where

238   "my_viral_accession2taxid.map" is our "conversion-table" file; "nodes.dmp" and

239   "names.dmp" are the taxonomy files downloaded from NCBI Taxonomy;

240   "all_viral.genomic.fna" contains all RefSeq viral sequences in fasta format; and

241   "Centrifuge_viral" is the name of the created database.

242   We also created Centrifuge's default database which includes all RefSeq sequences from

243   bacterial, archaeal, viral and human genomes (built on 2 September 2021) following the

244 software instructions). The results obtained with the default database are shown in

245 Supplementary Note 2.

246 When running Centrifuge to classify the reads a maximum of one taxonomic assignment

247 per read was retrieved (*-k 1*), to obtain an output comparable to the other classifiers. All

248 other parameters were left with default values.

249 **Kraken2**

250 Kraken2 version 2.0.7-beta was run with default parameters. To build our viral custom

251 database first we downloaded the taxonomy using the following command `kraken2-build`

252 `--download-taxonomy --db Kraken2_viral_db`, where Kraken2_viral_db is the name

253 of the database. Then we replace all the internal files (except

254 "Kraken2_viral_db/taxonomy") to have the following directory structure:

255 `Kraken2_viral_db/`

256 `|`

257 `|-- library`

258 `|   \-- viral`

259 `|       \-- library.fna`

260 `|`

261 `|-- seqid2taxid.map`

262 `|`

263 `\-- taxonomy`

264 Where "library.fna" is the same as "all_viral.genomic.fna" and "seqid2taxid.map" is the

265 same as "my_viral_accession2taxid.map"; both files used to build Centrifuge's custom

266 viral database. To build the database the following command was used: `kraken2-build -`

267 `-build --db Kraken2_viral_db`.

268 We also downloaded Kraken2 the prebuilt "standard" database from

269 https://benlangmead.github.io/aws-indexes/k2 dated to 17 May 2021. This database

270 includes the RefSeq of bacterial, archaeal, viral and human genomes plus UniVec_Core.

271 The results obtained with this default database are shown in Supplementary Note 2.

272 We also performed analyses building a database with a shorter k-mer length in order to

273 classify short (30 bp) reads, for these results see Supplementary Note 3.

274 **DIAMOND**

275   DIAMOND version 0.9.22 was used with the RefSeq viral protein database. The protein

276   sequences were downloaded from

277   https://ftp.ncbi.nlm.nih.gov/refseq/release/viral/viral.${i}.protein.faa.gz (${i} has values

278   from 1 to 3); the taxonomy nodes where downloaded from

279   https://ftp.ncbi.nih.gov/pub/taxonomy/taxdmp.zip; and the file which links accession

280   number to taxID was downloaded from

281   https://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/prot.accession2taxid.gz. All files

282   were downloaded on 1 September 2021. The database was built following the manual's

283   instructions.

284   In addition, we built a "full" database with archaeal, bacterial, human and viral protein

285   sequences. The viral sequences and taxonomy files are identical as the previous database.

286   We downloaded the archaeal sequences from

287   https://ftp.ncbi.nlm.nih.gov/refseq/release/archaea/archaea.nonredundant_protein.${i}.prote

288   in.faa.gz (${i} has values from 1 to 12); the bacterial from

289   https://ftp.ncbi.nlm.nih.gov/refseq/release/bacteria/bacteria.nonredundant_protein.${i}.prot

290   ein.faa.gz (${i} has values from 1 to 1396); and human from

291   https://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/human.${i}.protein.faa.gz (${i} has

292   values from 1 to 11). Archaeal and bacterial protein sequences were downloaded on 1

293   September 2021; human protein sequences were downloaded on 6 September 2021. The

294   results obtained with this "full" database are shown in Supplementary Note 2.

295   DIAMOND was run in the taxonomic classification mode (*-f 102* option) in order to

296   perform the taxonomic assignments; all other parameters were set as default. As the input

297   data is DNA, the query is translated into protein for each of its six reading frames and

298   aligned against the protein database.

299   We also performed analyses changing the default seed length used by DIAMOND in order

300   to classifiy short (30 bp) reads, for these results see Supplementary Note 3.

301 **MetaPhlAn2**

302   MetaPhlAn2 version 2.6.0 was run with default parameters. The default microbe clade-

303   specific marker genes database was used (mpa_v20_m200 version), it includes marker

304   genes from bacteria, archaea, viruses and eukaryotic microbes.

305   Summary statistics

306   To evaluate the performance of the classifiers, each read was assigned to one of the

307   following categories: i) correctly classified read at the species level ("correct species"), ii)

308   correctly classified read at a higher taxa level ("correct higher"), iii) misclassified read at

309   any taxa level ("incorrect") and iv) unclassified read ("unclassified"). Figure 1 shows a

310   schematic example of each of the four categories for *Variola virus*. The correctly classified

311   reads are divided into two sets hereafter: a set including only the reads classified correctly

312   at the species level ("correct species", abbreviated "s" below) and a set including all the

313   reads classified correctly at the species and at higher taxa ("correct species" and "correct

314   higher", abbreviated "s&h" below). Further, we computed the widely used statistics

315   sensitivity and precision to compare the classifiers (Wood and Salzberg 2014). Sensitivity

316   is the proportion of correctly classified reads over all simulated reads; precision is the

317   proportion of correctly classified reads over all classified reads. We computed two sets of

318   sensitivity and precision measures depending on whether "correct species" is assumed to be

319   correct or "correct species" and "correct higher" are assumed to be correct. More

320   specifically, sensitivity and precision are defined as follows for each viral sequence v_i:

321
$$Sensitivity_s^{v_i} = \frac{r_{cs}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i} + r_u^{v_i}}$$

322
$$Sensitivity_{s\&h}^{v_i} = \frac{r_{cs}^{v_i} + r_{ch}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i} + r_u^{v_i}}$$

323
$$Precision_s^{v_i} = \frac{r_{cs}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i}}$$

324
$$Precision_{s\&h}^{v_i} = \frac{r_{cs}^{v_i} + r_{ch}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i}}$$

325   where $r_{cs}^{v_i}$ are the number of "correct species" reads for the viral sequence v_i, $r_{ch}^{v_i}$ are the

326   number of "correct higher" reads for the viral sequence v_i, $r_i^{v_i}$ are the number of

327   "incorrect" reads for the viral sequence v_i, $r_u^{v_i}$ are the number of "unclassified" reads for

328   the viral sequence v_i. For some of the viral sequences v_j, all reads are "unclassified" and

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

330   We summarize the results by computing the mean sensitivities and precisions across the

331   simulated viral sequences. For the sensitivities, we have:

332
$$Sensitivity_s = \frac{1}{233}\sum_{i=1}^{233}\frac{r_{cs}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i} + r_u^{v_i}}$$

333
$$Sensitivity_{s\&h} = \frac{1}{233}\sum_{i=1}^{233}\frac{r_{cs}^{v_i} + r_{ch}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i} + r_u^{v_i}}$$

334   For the precisions, we have:

335
$$Precision_s = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{r_{cs}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i}}$$

336
$$Precision_{s\&h} = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{r_{cs}^{v_i} + r_{ch}^{v_i}}{r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i}}$$

337   where n_c is the number of viral sequences for which there is at least one read classified

338   (correctly or not), i.e. for which $r_{cs}^{v_i} + r_{ch}^{v_i} + r_i^{v_i} > 0$.

339   In addition, we counted and characterised the number of spurious extra taxa

340   identified by the classifiers. These taxa are associated with one or more "incorrect" reads.

341   They include any taxa that were not used to simulate the sequencing reads or taxa not

342   included in the lineages of the simulated viruses.

Effect of adding deamination damage

344   Besides DNA fragmentation, another molecular characteristic of ancient DNA is the

345   deamination process which, for double stranded libraries, results in an increase of C to T

346   substitutions at the 5' ends and G to A substitutions at the 3' ends (Briggs et al. 2007;

347   Carøe et al. 2018). We simulated reads of 60 bp length and added different levels of

348   deamination using deamSim (June 06 2016 version) gargammel sub-program (Renaud et al.

349   2017). The following parameters were applied: a nick frequency of 0.03, a geometric

350   parameter of 0.25 for the average length of overhanging ends, a probability of deamination

351   in the double-stranded portions of DNA of 0.01, and 11 different values for the probability

352   of deamination in the single stranded portions of the DNA (0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3,

353   0.35, 0.4, 0.45 and 0.5). Note that the resulting damage pattern is similar to what has been

354   observed for real data (Figure 2) with higher deamination fractions at read termini. The

355   resulting observed average number of substitutions was computed for the case of

356   NC_006273.2 (*Human betaherpes virus 5*) for all tested single-stranded probabilities of

357   deamination (Figure 3).

Effect of increasing substitution sequencing error

359   To better understand the effect of the distribution of errors across the reads,

360   substitution errors were also simulated following an Illumina-like error profiles. To do so,

361   sequencing errors were added to the simulated reads using ART (Huang et al. 2012)

362   indirectly varying the overall error rate by changing the parameter *qShift* of ART.

363   Specifically, seven different values for the parameter *qShift* were tested (0, -1, -3, -5, -7 and

364   -9). This parameter changes the quality score of the simulated reads, leading to more

365   substitutions in the simulated reads with lower values. Each *qShift* value corresponds to an

366   overall fold increase of substitution error (Huang et al. 2012):

367
$$fold\_increase = \frac{1}{10^{-\frac{qShift}{10}}}$$

368   For instance, a *qShift* of 0 corresponds to a one-fold increase (i.e. no increase or in

369   other words typical sequencing error), -3 to a ~two-fold increase and, finally, -9 to a 7.9-

370   fold increase. The effect of decreasing *qShift* on the observed number of substitutions in

371   simulated reads is shown in Figure 3 for the case of NC_006273.2 (*Human betaherpes*

372   *virus 5*) values. For this analysis also the simulated read length was set at 60 bp.

Human simulations

374   When human remains are sequenced, the resulting data contains microbial and

375   human DNA. To study the microbiome, ideally, reads mapping to the human genomes

376   should be first removed. However, such cleaning steps are imperfect and can lead to false

377   classifications. To evaluate how the tools classify potential human sequences that were not

378   properly filtered out, human reads were simulated using ART (ART-MountRainier-2016-

379   06-05, art_Illumina Q Version 2.5.8) with the human genome (GRCh37) as reference. As

380   above, the read length used was 60 bp, the coverage 10×, and the sequencing technology

381   was set to Illumina HiSeq 2500. Centrifuge, Kraken2, DIAMOND and MetaPhlAn2 were

382   then used to classify the reads as above. The proportions and the number of correctly (i.e.

383   classified as human) and incorrectly (i.e. not classified as human or at as a taxon within the

384   human lineage) classified reads were calculated; the number of extra taxa were counted and

385 characterised by determining the assigned kingdom and the number of assigned reads per

386 taxon.

Results

388   Overall, the performance of the four classifiers is surprisingly distinct across

389 simulations (Figure 4-9). The results for Centrifuge and Kraken2 are the most similar to

390 each other, MetaPhlAn2 is the most distinct compared to Centrifuge and Kraken2, while

391 DIAMOND occupies an intermediate position.

Classification results for the 60 bp read set

393   Figure 4 and Figure S1 summarise the results for the simulated 60 bp read set. As

394 discussed above, the simulated reads could be either correctly classified at the species level

395 ("correct species"), correctly classified at a higher taxa level ("correct higher"),

396 misclassified ("incorrect"), or "unclassified" (Figure 1).

397   When considering each virus separately, most reads are classified correctly at the

398 species level ("correct species") for most of the simulated viral sequences for Centrifuge

399 and Kraken2 (Figure S1). In contrast, for DIAMOND, the proportion of "correct species"

400 varies substantially with many viruses with low, intermediate and high "correct species"

401 proportions while the remaining reads are mostly "unclassified" or "correct higher".

402 MetaPhlAn2 stands out by having roughly two thirds of the viruses with no correctly

403 classified reads (neither as "correct species" or "correct higher"). As a result, the means

404 across viruses of the proportion of "correct species" are 24.41% (MetaPhlAn2), 53.61%

405 (DIAMOND), 93.39% (Kraken2) and 94.48% (Centrifuge). Unlike the other classifiers,

406 DIAMOND classifies a large proportion of reads correctly but at a higher taxonomic rank

407 ("correct higher") for many of the viruses. The other three classifiers have a handful of

408 viruses with a large proportion of "correct higher". Overall, most of the simulated viral

409 sequences have zero or a small proportion of "incorrect" reads, especially in the case of

410 Centrifuge and Kraken2 (Figure S1, for detailed list of the viruses with the highest incorrect

411 proportions per classifiers see Supplementary Note 4; Table S2 contains the raw numbers

412 of reads assigned to each taxa per classifier). The resulting averages of the proportions of

413 "incorrect" reads are 0.01% (Centrifuge), 0.26% (Kraken2), 0.75% (DIAMOND) and

414 2.79% (MetaPhlAn2). Finally, Centrifuge and Kraken2 classify most of the simulated reads

415   for all viruses (with a mean "unclassified" of 0.01% and 0.44%), while both DIAMOND

416   (mean of 20.4%) and MetaPhlAn2 (mean of 59.8%) do not classify a large proportion of

417   reads for most viruses (Figure S2).

418   The sensitivity and precision reflect the previous "correct species", "correct higher",

419   "incorrect" and "unclassified" overall values (Figure 4A circles and triangles). At the

420   species level, both means are considerably higher for Centrifuge and Kraken2 with mean

421   values above 93% for Sensitivity_s and Precision_s. In contrast, MetaPhlAn2's sensitivity

422   and precision are the lowest with values of 24.4% (Sensitivity_s) and 46.3% (Precision_s).

423   MetaPhlAn2's low Sensitivity_s value is a consequence of the high proportions of

424   "unclassified" reads for this classifier. Note that for this classifier, the precision is

425   computed by excluding the viruses without any classified reads, i.e. by excluding most of

426   the data (see Material and Methods). At the species level, DIAMOND has intermediate

427   values, with a Sensitivity_s of 53.6% and a Precision_s of 67.4%; both lower than

428   Centrifuge and Kraken2's, but higher than MetaPhlAn2's (Figure 4A circles).

429   When considering the "correct species" and "correct higher" reads to compute the

430   precision and the sensitivity (Figure 4A triangles), the order of the classifiers stays the

431   same, however the performance of DIAMOND and MetaPhlAn2 are greatly improved. In

432   particular, DIAMOND has many viruses with a large fraction of "correct higher" reads. As

433   a result, DIAMOND's Sensitivity_s&h is considerably higher than its Sensitivity_s with

434   values of 78.9% and 53.6%, respectively. Similarly, its Precision_s&h (98.6%) is higher

435   than its Precision_s (67.4%). DIAMOND's Sensitivity_s&h is lower than Centrifuge's and

436   Kraken2's, and higher than MetaPhlAn2's. Centrifuge and Kraken2 achieve the highest

437   Precision_s&h values with 99.9% and 99.7%, respectively; followed closely by

438   DIAMOND (98.6%), and lastly by MetaPhlAn2 (74.4%) (Figure 4A triangles).

439   Beyond the proportion of correctly classified reads per virus, we also considered

440   which viruses used in the simulations were recovered with at least one correctly classified

441   read (Figure 4B). Encouragingly, despite large differences observed at the read level, three

442   classifiers detected almost all the 233 viruses tested: Centrifuge and Kraken2 identified 232

443   viruses, and DIAMOND detected 228. In contrast MetaPhlAn2 reported only 87 viruses.

444   In addition, we computed how many spurious extra taxa were reported, these are

445   taxa that were neither the virus of interest nor in the lineage of that virus (Figure 4C).

446   Encouragingly, no classifier reports more than 2 extra taxa per virus on average. We found

447   that MetaPhlAn2 is the classifier with the highest number of spurious extra taxa (on

448   average 1.45 across tested viruses), followed by DIAMOND (1.06), Kraken2 (0.58) and,

449   finally, Centrifuge (0.41).

450   **Effect of increasing read length on the classification performance**

451   To better understand the effect of the length of the sequenced reads on the

452   classification results, we then investigated the performance of the classifiers for read

453   lengths ranging from 30 to 150 bp (Figure 5, Figures S2 & S3).

454   In general, by increasing the read length, the classifications improve. We observe

455   higher proportions of "correct species" reads (Figure S2 & S3), fewer "incorrect" reads,

456   higher sensitivity and precision, and a decrease in the number of spurious extra taxa

457   (Figures S2 & S3).

458   The most drastic result across classifiers for the read length simulations is that

459   DIAMOND and Kraken2 do not classify very short reads. Specifically, no 30 bp reads are

460   classified for DIAMOND and for Kraken2 given the default parameters used. We have

461   changed their default settings to classify these short 30 bp reads, however we found that the

462   number of classified reads was minimal for DIAMOND. In the case of Kraken2, it

463   classifies less than half of the reads and presents an increase of spurious extra taxa reported,

464   the results are shown in Supplementary Note 3. Similarly, DIAMOND classifies only very

465   few reads for 40 bp and 50 bp compared to higher read lengths. This contrasts with the

466   results for Centrifuge and MetaPhlAn2 whose performance do not significantly deteriorate

467   with shorter reads in the case of Centrifuge, or it remains stable across all lengths in the

468   case of MetaPhlAn2 (Figure 5).

469   From 60 bp onwards, most of the observations discussed above hold; Centrifuge and

470   Kraken2 have the highest "correct species" mean proportions, followed by DIAMOND and

471   MetaPhlAn2 (Figure S3). DIAMOND has the highest "correct higher" proportions; and

472   Centrifuge has the lowest "incorrect" mean proportions. Consequently, from 60 bp

473   onwards, the highest sensitivities (Sensitivity_s) and precisions (Precision_s) are observed

474   for Centrifuge and Kraken2 followed by DIAMOND and MetaPhlAn2. Regarding the

475   "correct higher" classifications, Centrifuge, Kraken2 and DIAMOND have the highest

476   Precision_s&h; as DIAMOND classifies correctly many reads at higher taxonomic level its

477   precision is greatly benefited. DIAMOND's has a Sensitivity_s&h that is lower than

478   Centrifuge's and Kraken2's but higher than MetaPhlAn2. Moreover, as above, regardless of

479   the simulated read length, the number of detected viruses remains the same. Most viruses

480   are detected by Centrifuge, Kraken2 and DIAMOND while an important fraction of the

481   tested viruses are missed by MetaPhlAn2 (Figure 5C).

482   A sharp decrease in the number of extra taxa for increased read lengths is observed

483   for MetaPhlAn2 (Figure 5D), the most constant classifier across read lengths for sensitivity.

484   A slighter decrease of spurious extra taxa is observed for Centrifuge and Kraken2. In the

485   case of DIAMOND, we see an increase up to 90 bp, followed by a decrease. The highest

486   number of spurious extra taxa is found in MetaPhlAn2 up to 60 bp; for read lengths above

487   60 bp DIAMOND present the highest number (Figure 5D).

488   When looking more in detail per classifier, for read lengths of 40 bp and longer, the

489   proportion of "correct species" reads increases for Centrifuge and Kraken2, while the

490   proportions of "correct higher" and "incorrect" reads decreases (Figure S2 & S3). This

491   results in a slight increase in the sensitivity and precision with longer reads (Figure 5A &

492   B). For DIAMOND, the fraction of "unclassified" reads decreases dramatically from 100%

493   (30 bp read lengths) to less than 25% (150 bp read lengths). Among the reads that are

494   classified, the "correct species" and "correct higher" ones increase substantially up to 90 bp

495   while the proportion of "correct higher" slightly decreases above 90 bp (Figure S2 & S3).

496   Unlike other classifiers, substantially more reads are classified with longer read lengths,

497   and a slight increase in the proportion of "incorrect" reads is observed as well (Figure S2 &

498   S3). As a result, the sensitivity improves substantially for longer reads for DIAMOND,

499   especially for the Sensitivity_s&h while the precision's increase is slighter (Figure 5).

500   MetaPhlAn2 is the most stable classifier when varying read length with the

501   classification categories "correct species" and "correct higher" remaining almost constant

502   (Figure S2 & S3). The only summary statistics that changes is the proportion of "incorrect"

503   reads which slightly decrease when increasing read length, resulting also in a decrease in

504   the number of spurious extra taxa. Consequently, the sensitivity remains constant and the

505   precision increases with longer reads (Figure 5A & B).

Effect of the deamination damage on the classification performance

506

507   To further assess the performance of the classifiers for ancient DNA like data, we

508 added deamination damage characteristic of ancient double-stranded DNA libraries

509 (increase of C to T substitutions at the 5' read termini and G to A substitutions at the 3'

510 read termini, see Figure 2) to the simulated 60 bp reads and evaluated the impact of

511 deamination on the taxonomic assignments (Figure 6). The probability of deamination in

512 the double-stranded portions of DNA was set at 0.01 for all simulations, while we varied

513 the different probabilities of single stranded deamination from 0 to 0.5 (see Material and

514 Methods). Those parameters result in realistic damage profiles, with 3' values of up to 0.45

515 being on the higher end of what can be observed (Figure 3), see e.g. (Malaspinas et al.

516 2014; Allentoft et al. 2015).

517   Overall, reassuringly, deamination has little to no effect on the classifiers'

518 performance. The slight reduction of "correct species" reads leads to a hardly noticeable

519 decrease of the sensitivities (Figure 6A). Similarly, the precision of all classifiers remains

520 constant for all values of the single-stranded deamination probabilities (Figure 6B). The

521 deamination damage has no effect, as well, in the number of correctly detected viruses

522 (among the 233 viruses tested) (Figure 6C).

523   Deamination only impacts the average number of spurious extra taxa for Centrifuge,

524 Kraken2 and DIAMOND, which increases (Figure 6D). In contrast, for MetaPhlAn2 the

525 average number of spurious extra taxa presents a small reduction. For single-stranded

526 probability of deamination values ranging from 0 to 0.15 Centrifuge reports the lowest

527 number of spurious extra taxa; from 0.2 and above DIAMOND has the lowest numbers.

528 MetaPhlAn2 reports the highest number of spurious extra taxa up to 0.35; above that value

529 Centrifuge presents higher numbers.

Effect of the substitution sequencing error on the classification performance

530

531   As the effects of ancient DNA deamination were minor, we indirectly investigated

532 whether this was the result of the unusual ancient DNA-like distribution and type (C to T in

533 5' ends and G to A in 3' ends) of the substitutions compared to standard sequencing error.

534 For these simulations, errors were added using ART which assumes a profile similar to the

535 ones observed for Illumina Sequencing machines (HiSeq 2500) by decreasing the *qShift*

536 parameter resulting in an increase of 1 to 7.9-fold of the overall error rate. As with damage,

537   a higher number of errors leads to divergent reads (reads with different sequences

538   compared to the reference from which they were generated, Figure 3).

539   As for damage, DIAMOND and MetaPhlAn2 are essentially not affected by the

540   sequencing error across simulations with results for a 7.9-fold increase in error rate

541   essentially identical to the ones at 1-fold across all statistics (Figure 7, Figures S2 & S3).

542   The trends observed for Centrifuge and Kraken2 (increase of "unclassified" reads

543   and small decrease of the rest of the categories) are exacerbated with Illumina-like

544   sequencing error compared to damage (Figure S2 & S3). This result is an almost linear

545   decrease in sensitivity when compared to the fold increase in error for Centrifuge and

546   Kraken2 (Figure 7). However, Centrifuge and Kraken2 precision values remain constant,

547   the reason for this is that the absolute number of correctly classified ("correct species"and

548   "correct higher") is high relative to "incorrect" along the increased substitutions (Figure

549   7B).

550   When considering not only the proportion of reads but also the viruses identified,

551   we observe that the number of recovered true viruses is constant regardless of the error rate

552   substitutions. Moreover, as for deamination, the number of spurious extra taxa increases

553   substantially for Centrifuge and Kraken2 with increased number of substitutions, while

554   DIAMOND's increase is less sharp and MetaPhlAn2 presents an overall small reduction

555   (Figure 7 D).

Classification of human reads

557   In recent years, many ancient microbes have been detected in ancient human

558   remains. In those studies, one of the challenges is to disentangle human from microbial

559   DNA. To assess the effect of not filtering out effectively human DNA, simulated reads

560   were generated from the reference human genome and used as input to the classifiers.

561   Almost the totality of reads (more than 98%) remains unclassified for all four

562   classifiers (Figure 8A), which is expected as none of them have human sequences in the

563   databases used. MetaPhlAn2 and DIAMOND have the highest proportion of unclassified

564   reads (99.98%), followed by Kraken2 (99.5%) and Centrifuge (98.4%). Among all

565   classifiers, Centrifuge presents the highest number of incorrectly assigned reads (Figure

566   8B). Above 78% of the classifications were made at the species level for all classifiers; and

567   about a quarter of the assignments were within the Heunggongvirae kingdom (Figure 9).

568   Interestingly, the number of spurious extra taxa varied greatly across classifiers

569   (Figure 8C), being Centrifuge the tool with the largest number of spurious extra taxa with

570   7,118 identified by at least 1 read, followed by Kraken2 (1,390), DIAMOND (528) and

571   MetaPhlan2 (91). Table S3 contains a list of each taxon identified by at least one read for

572   each classifier, we observed that several taxa have a large number of assigned reads

573   (several thousands) for all classifiers.

574 Discussion

575   By providing a direct window into the past, ancient virome studies have the

576   potential to shed light into the pathogens responsible for historical epidemics, to uncover

577   prehistorical epidemics, but also to provide clues about the molecular biology and the

578   evolutionary history of ancient viruses. Yet, finding ancient viruses is akin "finding broken

579   needles in noisy haystacks". Ancient genomes are fragmented, affected by post-mortem

580   damage, incomplete and contaminated by ancient and present organisms. Furthermore, the

581   DNA of viruses represent only a tiny fraction of the DNA extracted from the host. Thus, the

582   recovery of ancient viruses' genomes is an experimental and computational challenge and,

583   given how rare ancient samples are, it is crucial to recover as much ancient DNA as

584   possible.

585 Our approach and its limitations

586   To get a sense of the best suited classifier to screen for viruses in ancient

587   microbiome samples, we compared state-of-the art classifiers under controlled conditions,

588   i.e. in silico simulations. To do so, we selected 238 viral sequences from 233 human DNA

589   viruses, fragmented them and added ancient DNA-like noise, and classified the resulting

590   short and damaged reads with Centrifuge, Kraken2, DIAMOND and MetaPhlAn2. These

591   classifiers (that were not specifically developed for ancient DNA) differ in essentially all

592   possible ways including the underlying databases, the sequence search algorithms and the

593   taxonomic binning strategies (Table 1). Hereafter we provide some clues to explain the

594   main results that were obtained for each classifier and conclude with some

595   recommendations. However, it is important to note that our results and recommendations

596   are limited to the four tested classifiers and to the databases available at the time of the

597   analyses. Moreover, the simulations are a proxy of a real-life situations (i.e. a screening

598   step in a microbiome study). In other words, our recommendations are a good starting point

599   for anyone wishing to study ancient viromes. Yet future work should allow to update and to

600   further refine them.

601   ## Unidentified and incorrectly classified viruses

602   In most simulations with reads over 50 bp long, three of the classifiers, namely Centrifuge,

603   Kraken2 and DIAMOND, successfully detect at the species level more than 228 out of the

604   233 tested viruses. These encouraging results suggests that if viral genomes were present in

605   a sample with sufficient coverage, they would be detected by Centrifuge, Kraken2 and

606   DIAMOND even if the reads are fragmented and damaged. In contrast, MetaPhlAn2

607   missed 146 viruses. Regarding the incorrect assignments, encouragingly, for almost all

608   viruses across all classifiers (except for MetaPhlAn2), the incorrect proportion does not

609   exceed 43%. For a list as well as a discussion of the unidentified viruses and viruses with

610   the highest incorrect proportions see Supplementary Note 4.

611   ## Classification of very short (30 bp) reads

612   One of the main characteristics of ancient DNA is the highly fragmented nature of

613   the molecules, generally shorter than the sequencing read length (resulting in the

614   sequencing of the adapters). The classifiers we tested were not developed to tackle such

615   short reads and as expected, longer reads, which contain more information, positively

616   impact the overall classification results. Our simulations suggest that DIAMOND and

617   Kraken2's performance is substantially affected by a reduction in read length. This

618   contrasts with the result for Centrifuge and especially MetaPhlAn2 showing these two

619   classifiers can still handle shorter reads. In particular, we observed that 30 bp long reads are

620   not classified at all by DIAMOND and Kraken2, using their default parameters.

621   DIAMOND's performance is still substantially reduced at 40 bp and 50 bp while Kraken2's

622   sensitivity is close to its highest value at 40 bp. In other words, a significant number of

623   reads shorter than 40 bp would be lost when using Kraken2; and in the case of DIAMOND,

624   shorter than 60 bp. These results can be explained by the underlying algorithms and

625   databases. Kraken2's default k-mer length (35 bp) was used to build its database, making

626   an exact match impossible for 30 bp, as the k-mers are longer than the reads. Similarly,

627   DIAMOND uses as seeds 4 shapes of length 15-24 and weight 12 by default which

628   translates into 45-72 bp and weight 36. This explains why it failed to classify any read of

629 30 bp long, and why it has issues classifying 40 bp reads (the reads being only 4 bp longer

630 than the default seed weight). Relatedly, DIAMOND is the only classifier depending on

631 protein alignments and longer DNA reads are required to align the same number of amino

632 acids as nucleotides. Even when changing the default settings of DIAMOND and Kraken2

633 the number of short reads classified is minimal (DIAMOND) or it comes with a trade-off:

634 higher number of spurious taxa reported (Kraken2), see Supplementary Note 3.

635 **A high sensitivity and precision for Centrifuge and Kraken2**

636 As aDNA studies are usually limited in the amount of starting biological material, it

637 is crucial to assign correctly as many reads as possible and to minimise the number of

638 errors, i.e. to use a classifier with a high sensitivity and precision. With sensitivity and

639 precision values at the species level around 90% and above, Centrifuge and Kraken2

640 outperform DIAMOND and MetaPhlAn2. In other words, Centrifuge and Kraken2 classify

641 correctly more reads among the simulated reads but also classify correctly more reads

642 among the classified ones. In comparison, DIAMOND and MetaPhlAn2 detect correctly

643 less than 65% of the reads (sensitivity) on average across viruses (Figure 4A). For the

644 sensitivity, this can likely be explained in part by the differences in database; Centrifuge

645 and Kraken2 include whole genomes, while DIAMOND's database contains proteins and

646 MetaPhlAn2's contains custom clade specific markers, i.e. only a fraction of the genomes

647 in both cases. The latter two classifiers have therefore a large proportion of unclassified

648 reads. DIAMOND's precision is not higher at the species level as it classifies correctly a

649 large fraction of reads at higher taxonomic ranks ("correct higher", see below).

650 MetaPhlAn2's mean precision is considerably lower than the other classifiers because it has

651 high proportions of "incorrect" reads for some viruses, and because the number of viruses

652 used to compute it is smaller (only the viruses with classifications are considered, and many

653 viruses had 0 classified reads). As mentioned in Supplementary Note 4, three and 145 out

654 of the 233 tested viruses do not have sequences stored in DIAMOND's and MetaPhlAn2's

655 databases, respectively. As a result, both sensitivity and precision are reduced for both

656 tools. When computing the statistics considering only the viruses that are present in their

657 databases their performance improves: with 60 bp long reads DIAMOND reaches a

658 Sensitivity_s of 54.31%, Sensitivity_s&h of 79.58%, Precision_s of 68.24% and

659 Precision_s&h of 99.24%; while MetaPhlAn2 reaches a Sensitivity_s of 65.37%, a

660   Sensitivity_s&h of 69.67%, a Precision_s of 89.9% and a Precision_s&h of 96.92%.

661   Despite this improvement both of them are still outperformed by Centrifuge and Kraken2.

High proportions of "correct higher" reads for DIAMOND

663   DIAMOND exhibits a markedly higher number of reads correctly classified at a

664   higher taxonomic rank compared to the other classifiers. Protein changes have the potential

665   to directly impact the phenotype of an organisms and are therefore more conserved (Li

666   1997) and DIAMOND is the only classifier based on a protein database. Thus, the protein

667   database (which benefits from the redundancy of the genetic code and the high degree of

668   conservation of proteins) but also the LCA binning algorithm implemented in DIAMOND

669   may explain why this classifier has a much larger fraction of reads classified as "correct

670   higher", and as a result a great improvement of its precision when considering the higher

671   taxa. As an example, the virus with the highest proportion of reads assigned to higher levels

672   in DIAMOND, human erythrovirus V9 (NC_004295.1) has 84.8% of its reads correctly

673   assigned to the *Erythroparvovirus* genus. All these reads aligned to proteins of the correct

674   species human erythrovirus V9, but also to proteins from species within the genus. Given

675   the use of the LCA algorithm, DIAMOND reported the reads at genus level. Consequently,

676   when using DIAMOND, one should consider jointly the classifications at the species level

677   and above.

Sensitivity and precision are generally robust to deamination damage

679   Besides fragmentation, deamination damage is another key feature of ancient DNA

680   that could negatively impact classifications. Encouragingly, our simulations suggest that

681   ancient DNA-like damage has little to no effect on the classifiers' performance aside for an

682   increased number of spurious extra taxa for Centrifuge, MetaPhlAn2 and Kraken2 (see

683   below). To avoid a high number of spurious extra taxa, the use of DIAMOND would be

684   convenient, as long as the reads are long enough. Interestingly, when comparing

685   deamination and sequencing error simulations resulting in similar number of mismatches

686   (i.e. when the single stranded probability of deamination is set at 0.5 and the *qShift* value at

687   -9), deamination has little effect, while sequencing error reduces the sensitivities of

688   Centrifuge and Kraken2. This suggests having errors mostly concentrated at the end of the

689   reads (ancient DNA-like errors) do not hinder their classification as much as Illumina-like

690   errors, which are any kind of substitutions more evenly distributed across the read (Pfeiffer
691   et al. 2018).

692   **Extra spurious taxa**

693   Across all simulations with reads above 30 bp, Centrifuge and Kraken2 achieved the
694   lowest fraction of misclassified reads across viruses and MetaPhlAn2 the highest. Besides
695   the proportion of "incorrect" reads, we also investigated the number of spurious extra taxa
696   reported since they provide a clue of the amount of follow-up work to confirm candidate
697   microbes. In length simulations MetaPhlAn2 and DIAMOND have the highest number of
698   spurious extra taxa. In contrast, Centrifuge is the classifier with the lowest number of
699   spurious taxa for the same simulations; as the length increases the number of spurious extra
700   taxa decreases. This could be explained by Centrifuge's taxonomic binning strategy with
701   scores favouring longer hits. The contrasting high numbers of MetaPhlAn2 could be
702   explained by its database: with MetaPhlAn2's database missing markers for 145 of the
703   tested viruses and having markers for diverse organisms (archaea, bacteria and eukaryotes,
704   besides the viruses). However, it is important to mention that in the human simulations,
705   with the true species absent in all databases, Centrifuge is the tool with the highest number
706   of spurious taxa.

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

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

713   Ancient human reads mistakenly classified as viruses can be very problematic in
714   studies of human remains as both viral and human reads would exhibit characteristic
715   ancient DNA features such as fragmentation and deamination damage. To assess whether
716   human reads had any chance of being classified as viruses if they were not properly cleaned
717   out, we classified simulated human reads. The results show that encouragingly, most reads
718   remain as unclassified (the species is not in the database). Nevertheless, thousands of reads
719   are misclassified. Most of those incorrectly classified reads are assigned to a wrong species
720   (false positive), giving the user a sense of false confidence in the classification as it is so

721   specific. The assignments are generally hard to interpret except for a retrovirus identified

722   by Centrifuge, Kraken2 and DIAMOND. In this case, one of the top hits (i.e. the taxa with

723   the highest numbers of reads, see Table S3) was human endogenous retrovirus K113

724   (HERV-K113). As its name suggests this viral genome is integrated in the human genome

725   (Turner et al. 2001). As a result, and as we did not filter any of the simulated human data

726   prior to classification, some simulated human reads from the reference assembly are

727   assigned to the HERV-K113 present in the databases. The simulated reads coming from the

728   location in the human genome where the endogenous retrovirus was integrated aligned to

729   the HERV-K113 in the database. Note that if there had been a filtering step for human

730   reads (such as the removal of reads that map to the reference human genome), this virus

731   would not have been recovered (the reads would have mapped to the proviral form of

732   HERV-K113 in the reference human genome, hence would have been filtered out).

733   ## Recommendations and future directions

734   In the context of ancient virome studies, the ideal classification tool is one that

735   allows us to recover all viruses present in the studied sample but also that does not identify

736   any spurious extra viruses. In other words, we hope to find the right trade-off between

737   having a high number of true positives and a low number of false positives. In this study,

738   we considered both the fraction of reads per virus that were correctly classified, as well as

739   the number of recovered viruses and the number of spurious extra taxa with at least one

740   read assignment.

741   Sequenced ancient genomes are generally incomplete and being able to correctly

742   classify most regions in the viral genomes will be an advantage to recover ancient viruses.

743   Given the difficulty in finding ancient viruses and in being certain they are actually ancient,

744   every lead in the screening phase is generally followed by extensive work. These steps

745   generally include the mapping of sequenced data to individual candidate genomes

746   identified by classifiers (Krause-Kyora et al. 2018; Mühlemann, Jones, et al. 2018; Vågene

747   et al. 2018; Mühlemann et al. 2020). Any classification is treated as a candidate and

748   reducing the number of false positives is key to achieve a manageable workload, to

749   decrease the computational resources needed and to minimize false claims.

750   We find that, when considering the fraction of correctly classified reads, Centrifuge and

751   Kraken2 exhibit the highest sensitivity and precision at the species level. When including

752   higher taxonomic ranks, DIAMOND reaches a precision close to those of Centrifuge and

753   Kraken2; although the sensitivity remains considerably lower. Moreover, Centrifuge,

754   Kraken2 and DIAMOND recover essentially all the simulated viruses for most simulations.

755   Regarding Kraken2 and DIAMOND, very short fragments cannot be classified with these

756   two classifiers when using default parameters (shorter than 35 bp in the case of Kraken2,

757   and shorter than 45 bp for DIAMOND). We performed additional analyses by adjusting

758   some of the parameters (using a seed of 9 amino acids for DIAMOND and reducing the k-

759   mer size to 29 bp when building the database for Kraken2) so these tools can handle shorter

760   sequences (see Supplementary Note 3). In summary, DIAMOND classifies very few

761   additional reads. Moreover, for Kraken2, there is a trade-off: the number of spurious extra

762   taxa increases. Considering all those results together, Centrifuge and Kraken2 are likely the

763   better choice among the four tools to increase the likelihood of recovering ancient viruses.

764   One caveat with Centrifuge is that it outputs the largest numbers of spurious extra

765   taxa across when the species tested is not present in its database (see Figure 8 and

766   Supplementary Note 2) or when the reads are highly deaminated. Hence, if the list of

767   candidates is very wide (probably due to the extra taxa) and option to optimise

768   computational resources for downstream analyses would be to focus on the candidates with

769   the highest number of assignments (as we saw in the human simulations, an important

770   proportion of extra taxa has less than 10 assigned reads). For studies with high

771   deamination, DIAMOND could be a good choice to reduce the number of candidates for

772   further downstream analyses (as long as the reads have the appropriate length) as its

773   sensitivity and precision are stable (Figure 6). We also recommend DIAMOND for fast

774   screenings given its high Precision_s&h and as its database is smaller than Kraken2's or

775   Centrifuge's and is easier to build (see Table 1 and Materials and Methods).

776   MetaPhlAn2 did not perform well in our analyses. While MetaPhlAn2 has

777   advantages, such as having stable sensitivity and precision values in different conditions

778   (Figures 5-7), the large number of viruses that would be missed in ancient virome studies

779   suggest it is not well-suited for such analyses. However, its small database makes this tool

780   convenient if researchers are after specific viruses, as long as the virus of interest has a

781   marker gene in the database or the researchers define its marker genes and add them to the

782   database following the developer's instructions.

783   For human studies, our simulations suggest that, even when the human genome is
784   included in the database, cleaning as many human reads as possible would be necessary
785   prior to classification to minimize the number of false positives. This would come at the
786   expense of losing true candidates. However, we found hundreds of detected spurious extra
787   taxa across different kingdoms within the viruses, and this could lead to false claims. To
788   avoid having to handle so many false positives, our suggestion would be to map the data to
789   the human genome and remove all reads that get assigned coordinates including those with
790   low mapping qualities prior to classification using e.g. Sunbeam (Clarke et al. 2019).
791   The results presented here are focused on DNA viruses that infect humans.
792   However, we expect they likely hold for viruses from other hosts as well. The reasons for
793   the difficulty to classify some viruses for specific classifiers are mainly the specific
794   databases used and the viral taxonomy of the tested viruses (see Supplementary Note 4).
795   Hence, if the intention is to answer a specific biological question, i.e. to identify a given set
796   of viruses with any of the classifiers tested here, we recommend to first make sure the
797   viruses of interest are present in the classifier's database. To help users in that regard, we
798   have also included Table S2, that lists (for 60 bp) the raw number of reads per taxon for
799   each classifier for all human DNA viruses included in this study.
800   To conclude, the four classifiers that we compared performed remarkably well
801   considering they were not specifically developed to handle aDNA data and are robust to
802   ancient DNA deamination damage. In our simulations, Centrifuge and Kraken2
803   outperforms the other classifiers as Centrifuge can handle short fragments, and both of
804   them have high sensitivity and precision values across read lengths. Moreover, for human
805   studies, human ancient DNA contamination could lead to a very large number of false
806   positives if human reads are not filtered properly. Finally, an area of future research would
807   be to adapt the databases and default parameters of the classifiers to improve their precision
808   and sensitivity values, and to increase the number of identified viruses, especially when
809   handling short reads.
810

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.

840   *Methods Ecol. Evol.* 9:410–419.

841   Clarke EL, Taylor LJ, Zhao C, Connell A, Lee J-J, Fett B, Bushman FD, Bittinger K. 2019.

842   Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments.

843   *Microbiome* 7:46.

844   Dabney J, Meyer M, Pääbo S. 2013. Ancient DNA Damage. *Cold Spring Harb. Perspect.*

845   *Biol.* 5:a012567.

846   Duggan AT, Perdomo MF, Piombino-Mascali D, Marciniak S, Poinar D, Emery MV,

847   Buchmann JP, Duchêne S, Jankauskas R, Humphreys M, et al. 2016. 17th Century Variola

848   Virus Reveals the Recent History of Smallpox. *Curr. Biol.* 26:3407–3412.

849   Eisenhofer R, Weyrich LS. 2019. Assessing alignment-based taxonomic classification of

850   ancient microbial DNA. *PeerJ* 7:e6594.

851   Ferragina P, Manzini G. 2000. Opportunistic data structures with applications. In:

852   Proceedings 41st Annual Symposium on Foundations of Computer Science. p. 390–398.

853   Green RE, Malaspinas A-S, Krause J, Briggs AW, Johnson PLF, Uhler C, Meyer M, Good

854   JM, Maricic T, Stenzel U, et al. 2008. A complete Neandertal mitochondrial genome

855   sequence determined by high-throughput sequencing. *Cell* 134:416–426.

856   Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read

857   simulator. *Bioinformatics* 28:593–594.

858   Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data.

859   *Genome Res.* 17:377–386.

860   Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and sensitive

861   classification of metagenomic sequences. *Genome Res.* [Internet]. Available from:

862   http://genome.cshlp.org/content/early/2016/11/16/gr.210641.116

863   Krause-Kyora B, Susat J, Key FM, Kühnert D, Bosse E, Immel A, Rinne C, Kornell S-C,

864   Yepes D, Franzenburg S, et al. 2018. Neolithic and medieval virus genomes reveal complex

865   evolution of hepatitis B. *eLife* 7:e36666.

866   Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*

867   9:357–359.

868   Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler

869   transform. *Bioinformatics* 25:1754–1760.

870   Li W-H. 1997. Molecular Evolution. Sunderland, Mass

871   Malaspinas A-S, Lao O, Schroeder H, Rasmussen M, Raghavan M, Moltke I, Campos PF,

872   Sagredo FS, Rasmussen S, Gonçalves VF, et al. 2014. Two ancient human genomes reveal

873   Polynesian ancestry among the indigenous Botocudos of Brazil. *Curr. Biol.* 24:R1035–

874   R1037.

875   Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, Hingamp P,

876   Goto S, Ogata H. 2016. Linking Virus Genomes with Host Taxonomy. *Viruses* 8:66.

877   Mühlemann B, Jones TC, Damgaard P de B, Allentoft ME, Shevnina I, Logvin A,

878   Usmanova E, Panyushkina IP, Boldgiv B, Bazartseren T, et al. 2018. Ancient hepatitis B

879   viruses from the Bronze Age to the Medieval period. *Nature* 557:418–423.

880   Mühlemann B, Margaryan A, Damgaard P de B, Allentoft ME, Vinner L, Hansen AJ,

881   Weber A, Bazaliiskii VI, Molak M, Arneborg J, et al. 2018. Ancient human parvovirus B19

882   in Eurasia reveals its long-term association with humans. *Proc. Natl. Acad. Sci.* 115:7557–

883   7562.

884   Mühlemann B, Vinner L, Margaryan A, Wilhelmson H, Castro C de la F, Allentoft ME,

885   Damgaard P de B, Hansen AJ, Nielsen SH, Strand LM, et al. 2020. Diverse variola virus

886   (smallpox) strains were widespread in northern Europe in the Viking Age. *Science*

887   [Internet] 369. Available from: https://science.sciencemag.org/content/369/6502/eaaw8977

888   O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B,

889   Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq)

890   database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic*

891   *Acids Res.* 44:D733–D745.

892   Pérez-Brocal V, Moya A. 2018. The analysis of the oral DNA virome reveals which viruses

893   are widespread and rare among healthy young adults in Valencia (Spain). *PLOS ONE*

894   13:e0191867.

895   Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, Mayer G. 2018.

896   Systematic evaluation of error rates and causes in short samples in next-generation

897   sequencing. *Sci. Rep.* 8:10950.

898   Rascovan N, Sjögren K-G, Kristiansen K, Nielsen R, Willerslev E, Desnues C, Rasmussen

899   S. 2019. Emergence and Spread of Basal Lineages of Yersinia pestis during the Neolithic

900   Decline. *Cell* 176:295-305.e10.

901   Renaud G, Hanghøj K, Willerslev E, Orlando L. 2017. gargammel: a sequence simulator

902   for ancient DNA. *Bioinformatics* 33:577–579.

903   Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. 2012. Temporal Patterns of

904   Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. *PLOS ONE*

905   7:e34131.

906   Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. 2012.

907   Metagenomic microbial community profiling using unique clade-specific marker genes.

908   *Nat. Methods* 9:811–814.

909   Siqueira JD, Dominguez-Bello MG, Contreras M, Lander O, Caballero-Arias H, Xutao D,

910   Noya-Alarcon O, Delwart E. 2018. Complex virome in feces from Amerindian children in

911   isolated Amazonian villages. *Nat. Commun.* 9:4270.

912   Stern J, Miller G, Li X, Saxena D. 2019. Virome and bacteriome: two sides of the same

913   coin. *Curr. Opin. Virol.* 37:37–43.

914   Taubenberger JK, Reid AH, Lourens RM, Wang R, Jin G, Fanning TG. 2005.

915   Characterization of the 1918 influenza virus polymerase genes. *Nature* 437:889.

916   Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A,

917   Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic

918   profiling. *Nat. Methods* 12:902–903.

919   Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J. 2001. Insertional

920   polymorphisms of full-length endogenous retroviruses in humans. *Curr. Biol.* 11:1531–

921   1535.

922   Vågene ÅJ, Herbig A, Campana MG, García NMR, Warinner C, Sabin S, Spyrou MA,

923   Valtueña AA, Huson D, Tuross N, et al. 2018. Salmonella enterica genomes from victims

924   of a major sixteenth-century epidemic in Mexico. *Nat. Ecol. Evol.* 2:520.

925   Velsko IM, Frantz LAF, Herbig A, Larson G, Warinner C. 2018. Selection of Appropriate

926   Metagenome Taxonomic Classifiers for Ancient Microbiome Research. *mSystems*

927   [Internet] 3. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6050634/

928   Warinner C, Herbig A, Mann A, Fellows Yates JA, Weiß CL, Burbano HA, Orlando L,

929   Krause J. 2017. A Robust Framework for Microbial Archaeology. *Annu. Rev. Genomics*

930   *Hum. Genet.* 18:321–356.

931   Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2.

932   *Genome Biol.* 20:257.

933   Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using

934   exact alignments. *Genome Biol.* 15:R46.

935   Worobey M, Watts TD, McKay RA, Suchard MA, Granade T, Teuwen DE, Koblin BA,

936   Heneine W, Lemey P, Jaffe HW. 2016. 1970s and 'Patient 0' HIV-1 genomes illuminate

937   early HIV/AIDS history in North America. *Nature* 539:98–101.

938

Characteristics of the five 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 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.
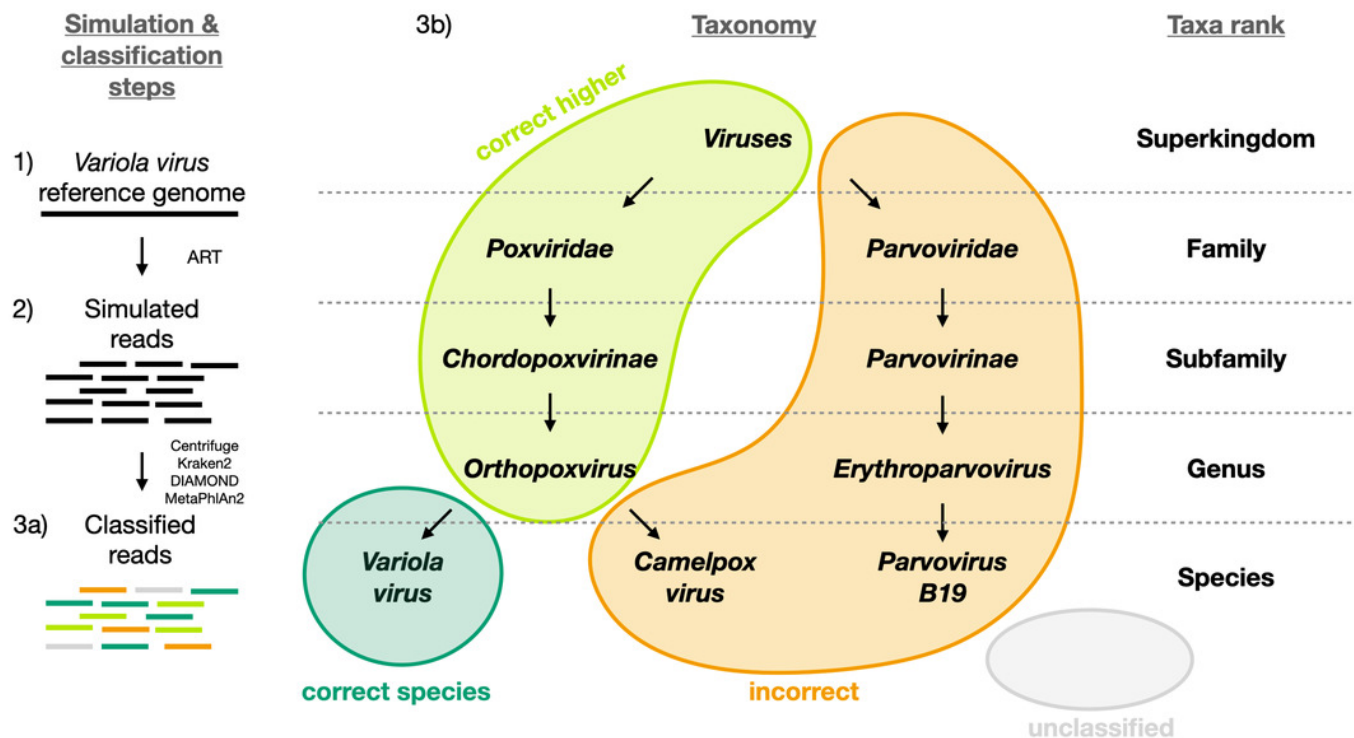
Figure 2

Observed deamination damage across the reads.

Observed frequency of substitutions across the reads simulated from the viral sequence NC_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).
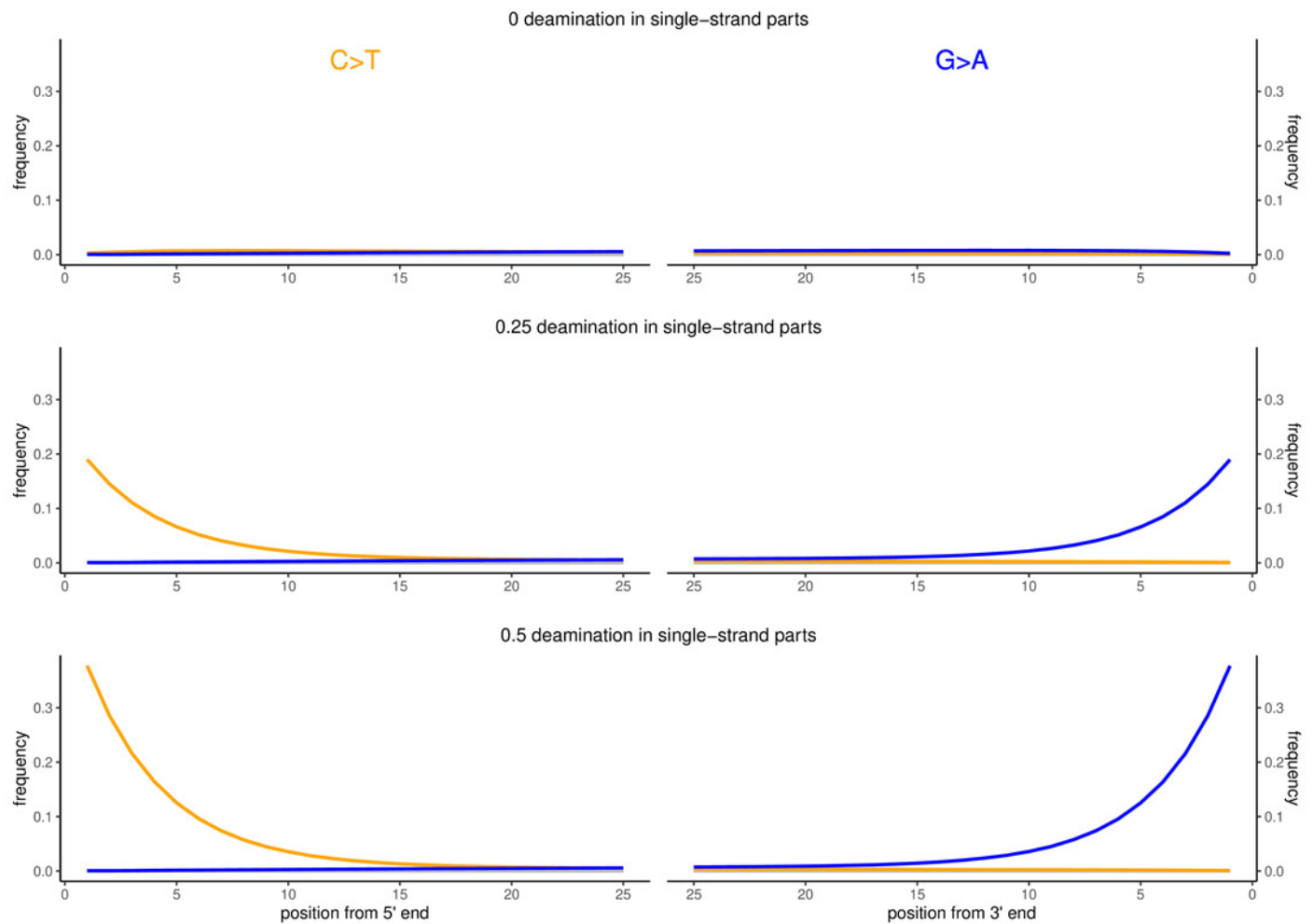
Figure 3

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).

A. Deamination simulations

B. Sequencing error simulations
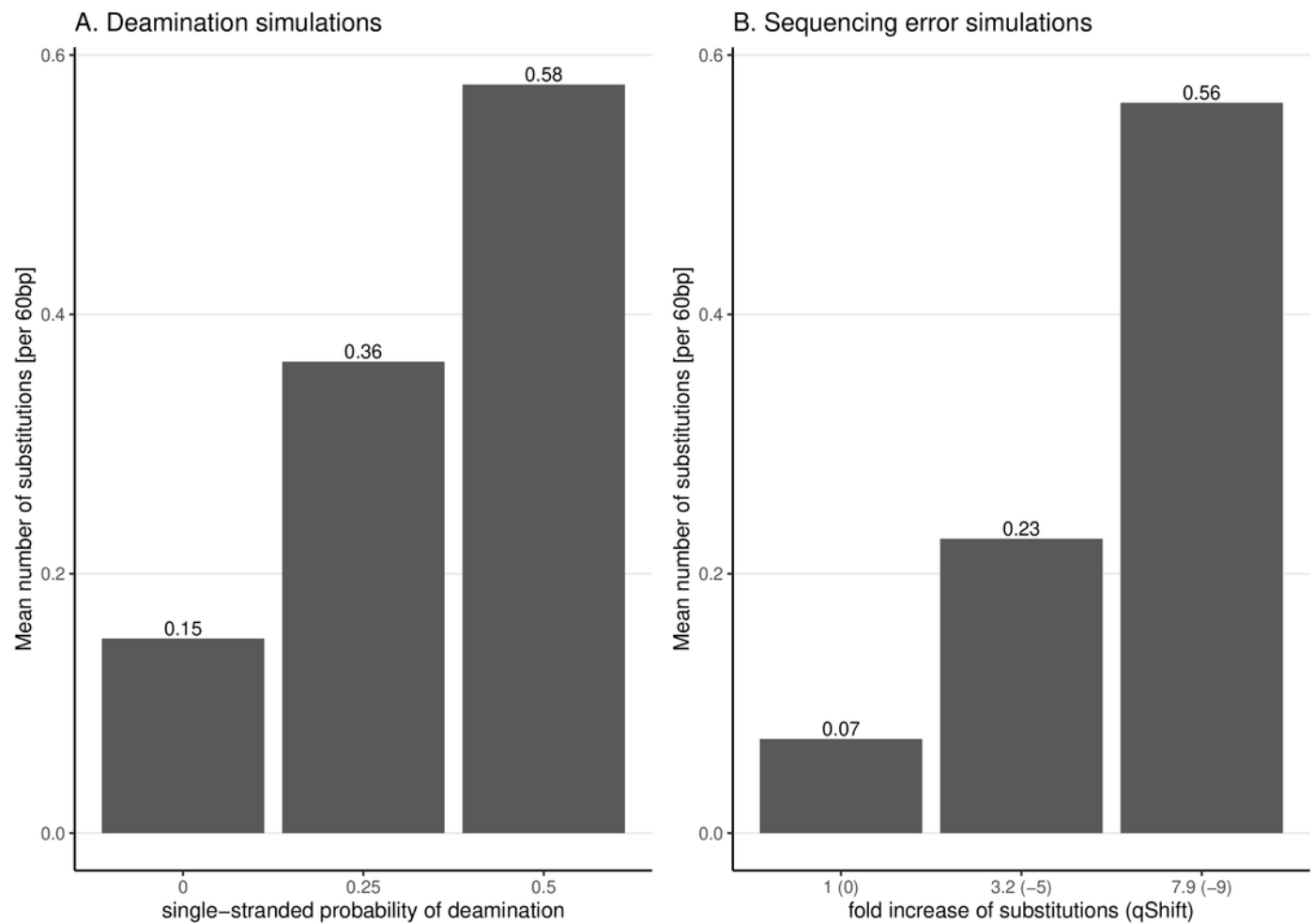
Figure 4

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

B. Total number of correctly detected viruses

C. Mean number of spurious extra taxa
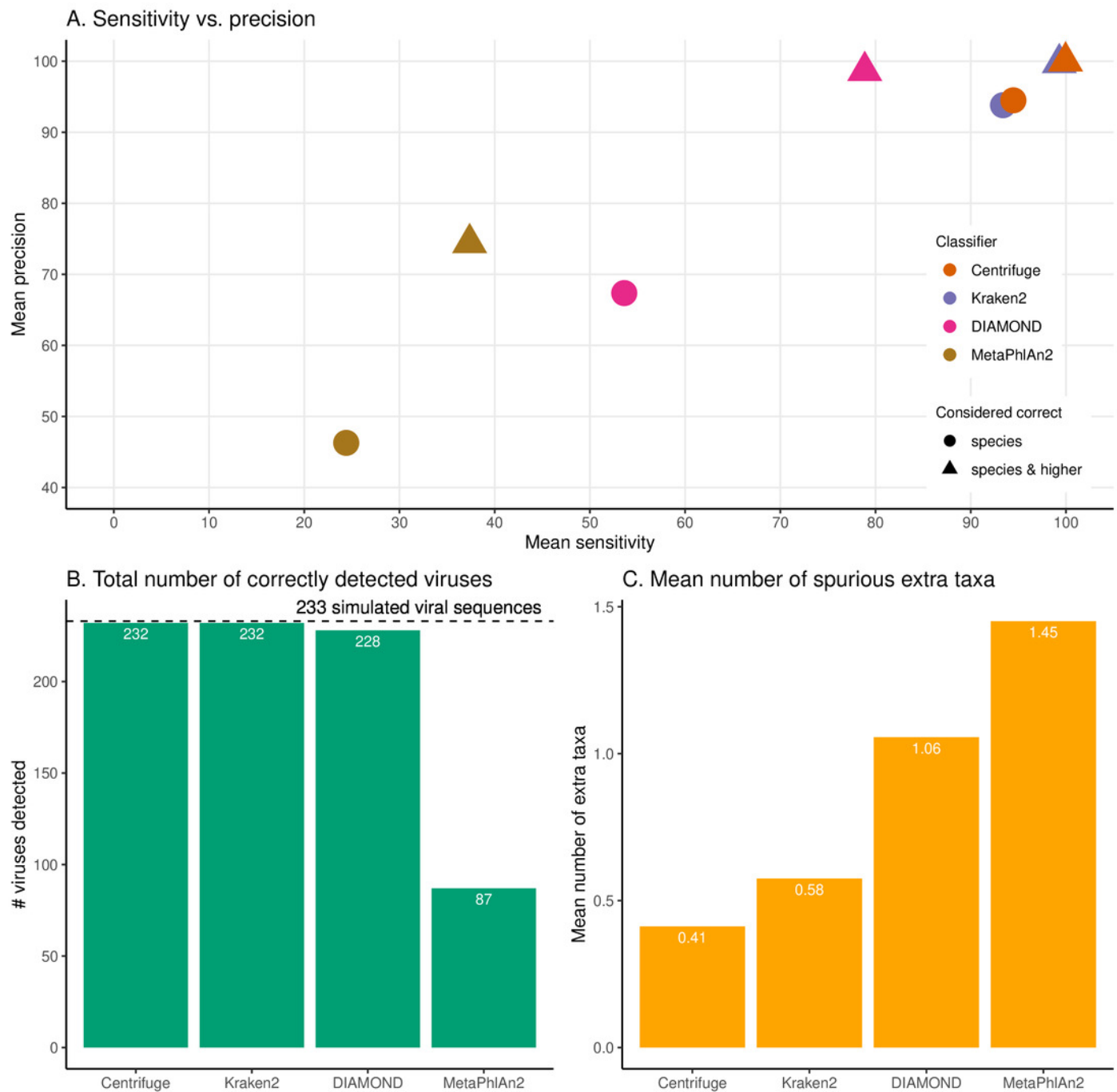
Figure 5

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.
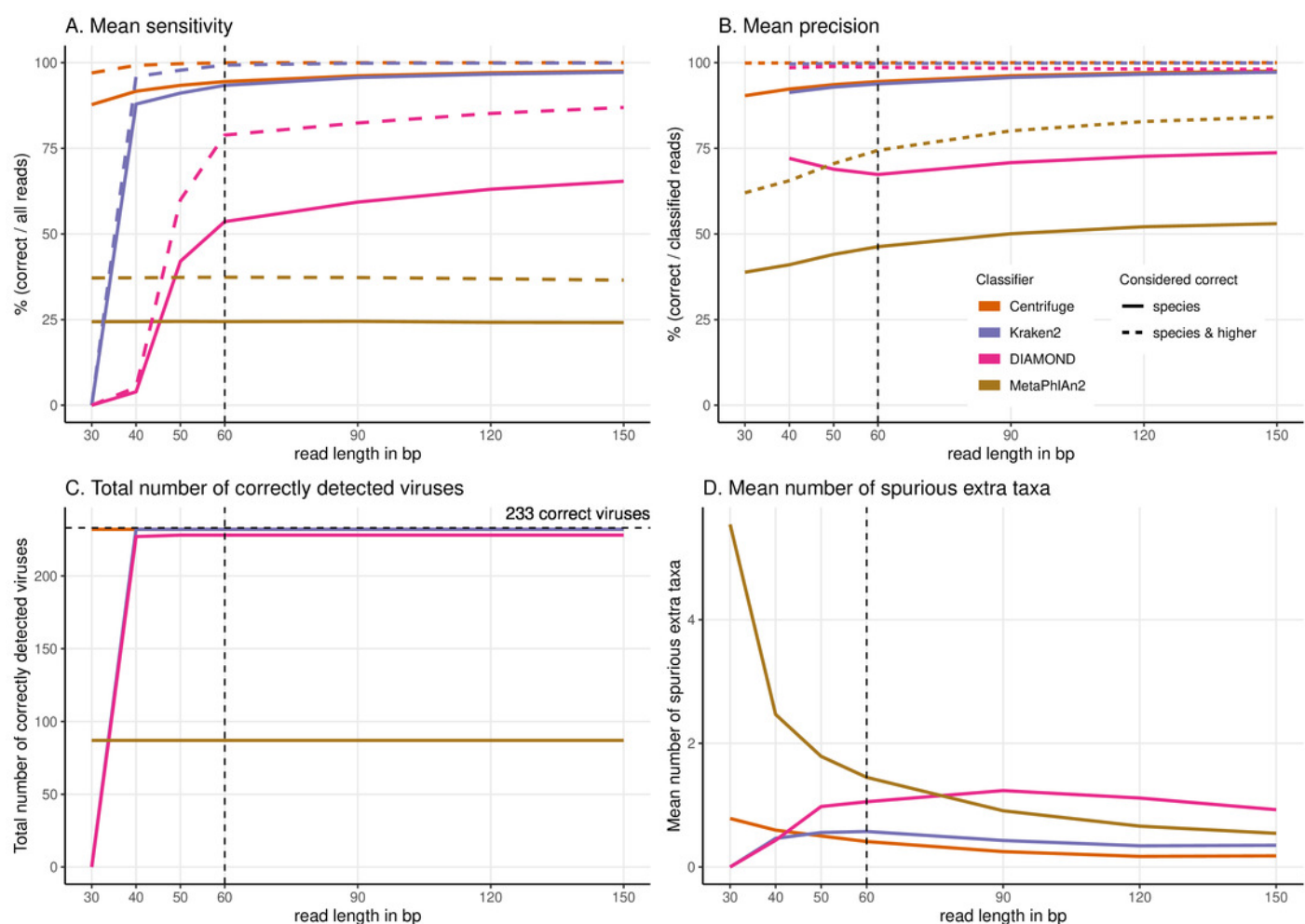
Figure 6

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.

A. Mean sensitivity

B. Mean precision

C. Total number of correctly detected viruses

D. Mean number of spurious extra taxa
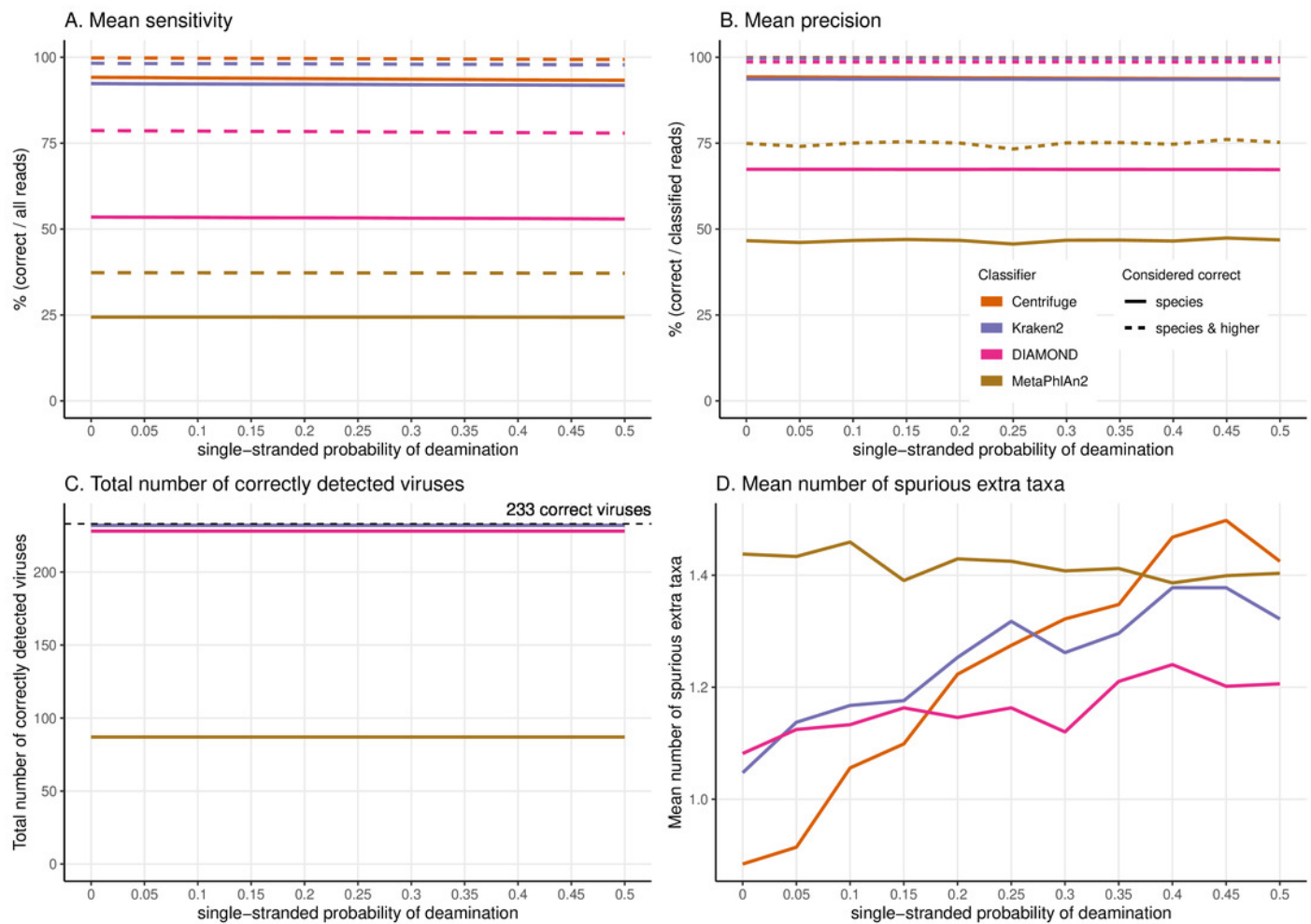
Figure 7

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.

A. Mean sensitivity

B. Mean precision

C. Total number of correctly detected viruses

D. Mean number of spurious extra taxa
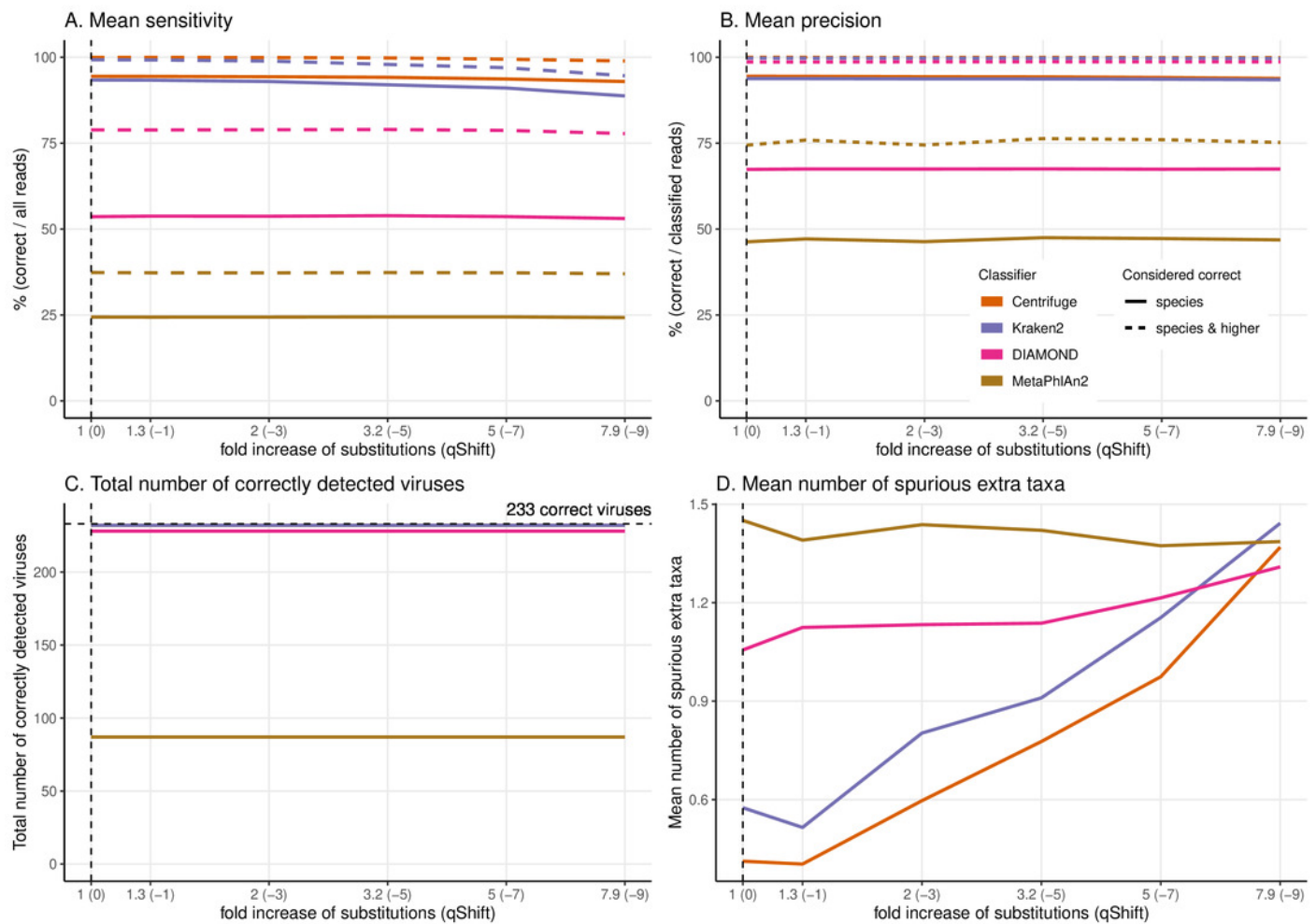
Figure 8

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.
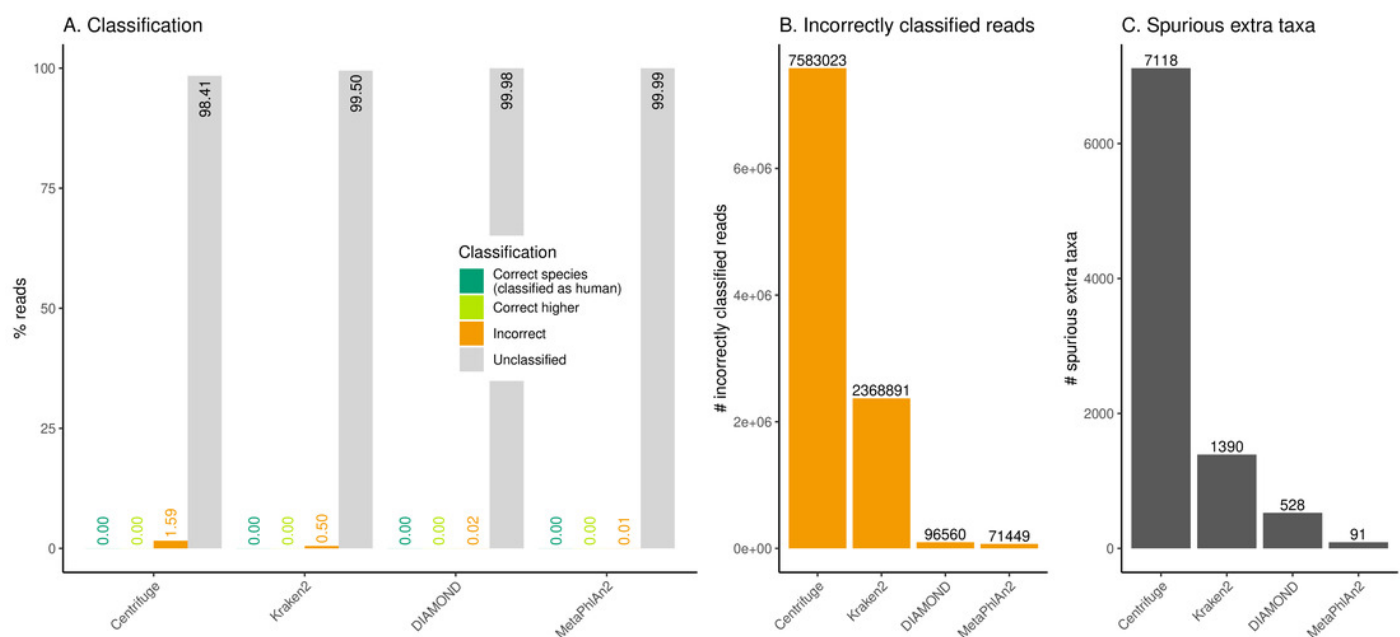
Figure 9

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).