



## Proficiency testing of virus diagnostics based on bioinformatics analysis of simulated in silico high-throughput sequencing datasets

**Brinkmann, Annika; Andrusch, Andreas; Belka, Ariane; Wylezich, Claudia; Höper, Dirk; Pohlmann, Anne; Petersen, Thomas Nordahl; Lucas, Pierrick; Blanchard, Yannick; Papa, Anna**

*Published in:*

Journal of Clinical Microbiology

*Link to article, DOI:*

[10.1128/JCM.00466-19](https://doi.org/10.1128/JCM.00466-19)

*Publication date:*

2019

*Document Version*

Peer reviewed version

[Link back to DTU Orbit](#)

*Citation (APA):*

Brinkmann, A., Andrusch, A., Belka, A., Wylezich, C., Höper, D., Pohlmann, A., ... Nitsche, A. (2019). Proficiency testing of virus diagnostics based on bioinformatics analysis of simulated in silico high-throughput sequencing datasets. *Journal of Clinical Microbiology*, 57(8), [e00466-19]. <https://doi.org/10.1128/JCM.00466-19>

---

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

1 Proficiency testing of virus diagnostics based on bioinformatics analysis of simulated *in silico*  
2 high-throughput sequencing datasets

3

4 Annika Brinkmann<sup>a#</sup>, Andreas Andrusch<sup>a</sup>, Ariane Belka<sup>b</sup>, Claudia Wylezich<sup>b</sup>, Dirk Höper<sup>b</sup>, Anne  
5 Pohlmann<sup>b</sup>, Thomas Nordahl Petersen<sup>c</sup>, Pierrick Lucas<sup>d</sup>, Yannick Blanchard<sup>d</sup>, Anna Papa<sup>e</sup>,  
6 Angeliki Melidou<sup>e</sup>, Bas B. Oude Munnink<sup>f</sup>, Jelle Matthijnsens<sup>g</sup>, Ward Deboutte<sup>g</sup>, Richard J.  
7 Ellis<sup>h</sup>, Florian Hansmann<sup>i</sup>, Wolfgang Baumgärtner<sup>i</sup>, Erhard van der Vries<sup>j</sup>, Albert Osterhaus<sup>k</sup>,  
8 Cesare Camma<sup>l</sup>, Iolanda Mangone<sup>l</sup>, Alessio Lorusso<sup>l</sup>, Maurilia Maracci<sup>l</sup>, Alexandra Nunes<sup>m</sup>,  
9 Miguel Pinto<sup>m</sup>, Vítor Borges<sup>m</sup>, Annelies Kroneman<sup>n</sup>, Dennis Schmitz<sup>f,n</sup>, Victor Max Corman<sup>o</sup>,  
10 Christian Drosten<sup>o</sup>, Terry C. Jones<sup>o,p</sup>, Rene S. Hendriksen<sup>c</sup>, Frank M. Aarestrup<sup>c</sup>, Marion  
11 Koopmans<sup>f</sup>, Martin Beer<sup>b</sup>, Andreas Nitsche<sup>a</sup>

12

13 <sup>a</sup>Robert Koch Institute, Centre for Biological Threats and Special Pathogens 1, Berlin, Germany

14 <sup>b</sup>Friedrich-Loeffler-Institut, Institute of Diagnostic Virology, Greifswald–Insel Riems, Germany

15 <sup>c</sup>Technical University of Denmark, National Food Institute, WHO Collaborating Center for  
16 Antimicrobial Resistance in Foodborne Pathogens and Genomics and European Union Reference  
17 Laboratory for Antimicrobial Resistance, Kgs. Lyngby, Denmark

18 <sup>d</sup>French Agency for Food, Environmental and Occupational Health & Safety, Laboratory of  
19 Ploufragan, Unit of Viral Genetics and Biosafety, Ploufragan, France

20 <sup>e</sup>Aristotle University of Thessaloniki, School of Medicine, Microbiology Department,  
21 Thessaloniki, Greece

22 <sup>f</sup>Erasmus Medical Centre, Department of Viroscience, Rotterdam, The Netherlands

23 <sup>g</sup>REGA Institute KU Leuven, Leuven, Belgium

24 <sup>h</sup>Animal and Plant Health Agency, Addlestone, United Kingdom

25 <sup>i</sup>University of Veterinary Medicine Hannover, Department of Pathology, Hannover, Germany

26 <sup>j</sup>University of Utrecht, Department of Infectious Diseases and immunity, Utrecht, The

27 Netherlands

28 <sup>k</sup>Artemis One Health Research Institute, Utrecht, the Netherlands

29 <sup>l</sup>Istituto Zooprofilattico Sperimentale dell’Abruzzo e Molise “G. Caporale”, National Reference

30 Center for Whole Genome Sequencing of microbial pathogens: database and bioinformatic

31 analysis, Teramo, Italy

32 <sup>m</sup>Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health (INSA),

33 Lisbon, Portugal

34 <sup>n</sup>National Institute for Public Health and the Environment, Bilthoven, The Netherlands

35 <sup>o</sup>Institute of Virology, Charité-Universitätsmedizin Berlin, Berlin, Germany

36 <sup>p</sup>Center for Pathogen Evolution, Department of Zoology, University of Cambridge, Cambridge,

37 United Kingdom

38

39 Running Head: COMPARE *in silico* Virus Proficiency Test

40

41 **# Address correspondence to Annika Brinkmann, BrinkmannA@rki.de**

42

43 **Keywords:**

44 High-Throughput Sequencing, Proficiency Test, External Quality Assessment, Virus Diagnostics

45

46

47 **Abstract**

48 Quality management and independent assessment of high-throughput sequencing-based virus  
49 diagnostics have not yet been established as a mandatory approach for ensuring comparable  
50 results. Sensitivity and specificity of viral high-throughput sequence data analysis are highly  
51 affected by bioinformatics processing, using publicly available and custom tools and databases,  
52 and differ widely between individuals and institutions.

53 Here, we present the results of the COMPARE (COllaborative Management Platform for  
54 detection and Analyses of [Re-] emerging and foodborne outbreaks in Europe) *in silico* virus  
55 proficiency test. An artificial, simulated *in silico* dataset of Illumina HiSeq sequences was  
56 provided to 13 different European institutes for bioinformatics analysis towards the identification  
57 of viral pathogens in high-throughput sequence data. Comparison of the participants' analyses  
58 shows that the use of different tools, programs, and databases for bioinformatics analyses can  
59 impact the correct identification of viral sequences from a simple dataset. The identification of  
60 slightly mutated and highly divergent virus genomes has been identified as being most  
61 challenging: Furthermore, the interpretation of the results together with a fictitious case report by  
62 the participants showed that in addition to the bioinformatics analysis, the virological evaluation  
63 of the results can be important in clinical settings.

64 External quality assessment and proficiency testing should become an important part of  
65 validating high-throughput sequencing-based virus diagnostics and could improve harmonization,  
66 comparability, and reproducibility of results. Similar to what is established for conventional  
67 laboratory tests like PCR, there is a need for the establishment of international proficiency testing  
68 for bioinformatics pipelines and interpretation of such results.

69

70 **Introduction**

71 High-throughput sequencing (HTS) has become increasingly important for virus diagnostic in  
72 human and veterinary clinical settings and for disease outbreak investigations (1-3). Since the  
73 introduction of the first HTS platform only about one decade ago, sequencing quality and output  
74 have been increasing exponentially, combined with continuing decreased costs per base. Thus,  
75 HTS has become a standard method for molecular diagnostics in many virological laboratories.  
76 The relatively unbiased approach of HTS not only enables the screening of clinical samples for  
77 common and expected viruses, but also allows an open view without preconceptions about which  
78 virus might be present. This approach has led to the discovery of novel viruses in clinical  
79 samples, such as Bas-Congo virus associated with hemorrhagic fever outbreaks in Central Africa  
80 (2), Lujo arenavirus in southern Africa (3) and Borna-like virus as the causative agent of several  
81 cases of encephalitis with fatal outcome in Germany (4). Considering the potential of HTS to  
82 complement or even replace existing 'gold-standard' diagnostic approaches such as polymerase  
83 chain reaction (PCR) and qPCR, quality assessment (QA) and accreditation processes need to be  
84 established to ensure quality, harmonization, comparability and reproducibility of diagnostic  
85 results. While the computational analysis of the immense amount of data produced requires  
86 dedicated computational infrastructure, bioinformatics knowledge or software developed by (bio-  
87 ) informaticians, the interpretation of the results also requires evaluation by an experienced  
88 virologist or physician. In many cases, true positive results can be difficult to discern among large  
89 numbers of false positives, or may be entirely missing from result sets due to false negative  
90 results. Interpretation of results also requires knowledge of anomalies that may arise through  
91 sequencing artefacts or contamination.

92 Proficiency testing (PT) is an external quality assessment (EQA) for evaluating and verifying  
93 sequencing quality and reliability in HTS analyses. The pioneer in EQA and PT for infectious

94 disease applications of HTS has been the Global Microbial Identifier (GMI) initiative, which has  
95 been organizing annual PT's since 2015, focusing on sequencing quality parameters including  
96 detection of antimicrobial resistance gene, Multilocus sequence typing, and phylogenetic analysis  
97 of defined bacterial strains ([https://www.globalmicrobialidentifier.org/workgroups/about-the-](https://www.globalmicrobialidentifier.org/workgroups/about-the-gmi-proficiency-tests)  
98 [gmi-proficiency-tests](https://www.globalmicrobialidentifier.org/workgroups/about-the-gmi-proficiency-tests)) (5). Subsequently, the concept was similarly established regionally for  
99 United States laboratories offered by the FDA (6, 7).

100 COMPARE (COllaborative Management Platform for detection and Analyses of (Re-) emerging  
101 and foodborne outbreaks in Europe, (<http://www.compare-europe.eu/>) is a European Union-  
102 funded programme with participation of institutions with hands-on experience in viral outbreak  
103 investigation and with the vision to improve the identification of (novel) emerging diseases  
104 through HTS technologies. One of the ambitious goals is to establish and enhance quality  
105 management and quality assurance in HTS, including external assessment and inter-laboratory  
106 comparison.

107 In this study, we present the results of the first global PT to assess bioinformatics analysis of  
108 simulated *in silico* clinical HTS virus data offered by the COMPARE network. The viral  
109 sequence dataset was accompanied with a fictitious case report to facilitate a more real scenario  
110 to support the identification of the simulated virus included the dataset.

111

112

### 113 **Tools and programs for bioinformatics analysis**

114 Over the past years, numerous tools, programs, and ready-to-use workflows have been  
115 established, making metagenomics sequence analyses accessible to scientists from all research  
116 fields. Workflows for the typical analysis of HTS data and for the identification of viral  
117 sequences are based on the same general tasks and tools, including quality trimming,

118 background/host subtraction, *de novo* assembly, and sequence alignment and annotation.  
119 Sequence processing usually starts with obligatory quality assessment and trimming, using  
120 programs like FastQC or Trimmomatic, including removal of technical and low-complexity  
121 sequences or filtering of poor-quality reads (8, 9). Following these initial steps, many workflows  
122 include the subtraction of background reads, e.g., host and bacteria, to reduce the total amount of  
123 data and increase specificity, using tools such as BWA (Burrows-Wheeler Alignment Tool) or  
124 Bowtie2 (10, 11). *De novo* assembly of HTS reads into longer, contiguous sequences (contigs),  
125 followed by reference-based identification, has been shown to improve the sensitivity of  
126 pathogen identification. Such analyses depend heavily on the use of assemblers, such as SPAdes  
127 or VELVET, which make use of specific assembly algorithms, such as overlap-layout-consensus  
128 graph or de Bruijn graph algorithms (12, 13). Alignment tools like BLAST, DIAMOND, Kraken,  
129 and Usearch are among the most important components in the bioinformatics workflows for  
130 pathogen identification and taxonomic assignment of viral sequences (14-17). As command-line  
131 tools for HTS sequence require specific knowledge in bioinformatics, complete workflows and  
132 pipeline approaches were developed, including ready-to-use web-based tools, such as RIEMS  
133 (Reliable Information Extraction from Metagenomic Sequence datasets), PAIPline (PAIPline for  
134 the Automatic Identification of Pathogens), Genome detective, and others (18-20). As the  
135 COMPARE *in silico* PT focuses on comparing different tools and software programs for  
136 bioinformatics analyses, an overview of frequently-used programs is given in Table 1. A more  
137 extensive overview of virus metagenomics classification tools and pipelines published between  
138 2010 and 2017 can be found at (<https://compare.cbs.dtu.dk/inventory#pipeline>).

139

140

141 **Methods**142 *Organization*

143 The virus PT was initiated by the COMPARE network and organized by the Robert Koch  
144 Institute. Invitations to participate were free of charge for research groups experienced in  
145 analyzing HTS datasets, and were announced through email and the COMPARE website.

146 Participants were asked to analyze an *in silico* HTS dataset, with the main goal being to identify  
147 the viral reads with their bioinformatics tools and workflows of choice and to interpret the  
148 obtained results including final diagnostic conclusions.

149 An artificial, simulated *in silico* dataset of >6 million single-end 150bp long Illumina HiSeq  
150 sequences derived from viral genomes, human chromosomes and bacterial DNA was provided to  
151 13 different European institutes for bioinformatics analysis towards the identification of viral  
152 pathogens in high-throughput sequence data. In order to assess how different level of experience  
153 and/or bioinformatics methodologies affects the outputs and interpretation, participants were  
154 allowed to use their bioinformatics tools and workflows of choice. Participants were invited to  
155 report the PT results via an online survey within eight weeks (from September 16, 2016 until  
156 November 16, 2016). Overall results were anonymized by the organizers but each participant was  
157 provided with the identifier for their own results.

158

159 *In silico HTS dataset*

160 The simulated *in silico* dataset consisted of a total of 6,339,908 reads (Table 2), based on a  
161 single-end 150-bp Illumina HiSeq 2500 run with an empirical read quality score distribution of  
162 Illumina-specific base substitutions. The artificial dataset was simulated with the ART program  
163 (21). Sequences were generated from the Human Genome Reference Consortium Build38  
164 (GRCh38, NCBI accession CM000663–CM000686), *Acinetobacter johnsonii* (NCBI accession



165 NZ\_CP010350.1), *Propionibacterium acnes* (NCBI accession NZ\_CP012647.1) and  
166 *Staphylococcus epidermis* (NCBI accession NZ\_CP009046.1). In addition to human and bacterial  
167 reads, simulated viral sequences of four viruses, Torque teno virus (TTV; NCBI accession  
168 NC\_015783.1), human herpesvirus 1 (HSV-1; NCBI accession NC\_001806.2), measles virus  
169 (MeV; NCBI accession NC\_001498.1) and a novel avian bornavirus (nABV; NCBI accession  
170 JN014950.1) were included in different numbers and with different levels of similarity to known  
171 viruses present in databases (Table 2). TTV and HSV-1 were included in the panel as the easiest  
172 sequences to identify (with 1,917 and 2,000 reads respectively, and 100% nucleotide identity  
173 with the reference sequences), followed by a slightly altered MeV (1,000 reads, with 82%  
174 nucleotide identity to the reference genome) and, as the likely most difficult taxon, nABV (only  
175 500 reads and 55% nucleotide identity to reference JN014950.1). The dataset has been uploaded  
176 to the European Nucleotide Archive with the study accession number PRJEB32470.

177

#### 178 *Participants*

179 Thirteen participants applied for the COMPARE virus PT and completed the survey within the  
180 given timeframe. Participants were registered from Belgium (n = 1), Denmark (n = 1), France (n  
181 = 1), Germany (n = 4), Greece (n = 1), Italy (n = 1), The Netherlands (n = 2), Portugal (n = 1) and  
182 United Kingdom (n = 1). The 13 participants represented 13 different institutes or organizations.  
183 Information about the participants' background is given in Table 4.

184

#### 185 *Case report*

186 To simulate clinical relevance and to set the background for evaluation of the bioinformatics  
187 results, the following fictitious case report was provided with the dataset:

188 *Recently, a 14-year-old boy from Berlin, Germany, was hospitalized with sudden blindness,*  
189 *reduced consciousness and movement disorders. The patient's mother reported developmental*  
190 *disorders starting one year ago, with concentration problems, uncontrolled fits of rage, overall*  
191 *decreasing performance in school and occasional compulsive head nods. Unfortunately, the*  
192 *patient had received neither medical examination nor treatment, but had attended psychological*  
193 *treatment, assuming behavioral problems.*

194 *Magnet resonance tomography of the patient's brain showed white and gray matter lesions and*  
195 *gliosis. Soon after hospitalization, the patient showed a persistent vegetative state and died.*

196 *A sample of the boy's brain tissue was sequenced using the Illumina HiSeq 2500 platform,*  
197 *resulting in approximately 6 million single end reads of 150 bp each.*

198 This case of subacute sclerosing panencephalitis (SSPE) can be caused by a persistent infection  
199 with a mutated MeV (22). However, the symptoms described could also be caused by HSV-1 and  
200 borna-like viruses (4, 23).

201

#### 202 *Reported PT results*

203 Results were collected using the Robert Koch Institute's online survey software VOXCO. The  
204 survey contained 23 questions including general participant information and specifications about  
205 the programs used, parameter settings, computer specifications as well as the final results of the  
206 PT, including an evaluation of the case. The responses were collected as single or multiple  
207 options from a multiple-choice questionnaire with additional free text for remarks and comments.

208

#### 209 *Analysis of PT results*

210 The results were evaluated based on sensitivity (true positive rate i.e. fraction of true virus reads  
211 that were identified), specificity and total time of the bioinformatics analysis (Table 3). The time

212 of analysis was evaluated based on the computational time only, without time for preparation and  
213 discussion of the bioinformatics results. Correlation of the time of analysis with computer and  
214 server specifications was only based on use of online analysis, personal computer, server and  
215 high-performance virtual machine. Although the pathogen identification by HTS-related  
216 metagenomics should naturally involve experienced qualified health professionals, participants  
217 were dared to attempt an interpretation regardless of the background of the team performing  
218 bioinformatics. In this context, no qualitative and quantitative scoring was performed in this part.

219

## 220 **Results**

221

### 222 *PT results*

223 The results of the PT were evaluated based on sensitivity, specificity, total turnaround time, and  
224 interpretation of results (Table 3). HSV-1 was identified by all participants (Tables 3-4, Fig. 1).  
225 For most of the participants, the identified read numbers for HSV-1 were complete or near  
226 complete (actual HSV-1 read count = 2,000). One participant identified more reads of HSV-1  
227 than present in the dataset (participant 7; 8,361 reads identified).

228 TTV (actual read count = 1917) and MeV were identified by all participants except for one  
229 (participant 4) (Tables 3-4, Fig. 1). For TTV, the read numbers identified were complete or  
230 almost complete for all participants, with the exception of participant 9 who was only able to  
231 identify 29% of the TTV reads. For the mutated MeV (actual read count = 1000), seven out of 13  
232 participants were able to identify complete or almost complete read numbers (participants 3, 5, 6,  
233 8, 10, 11, 12), whereas five participants (participants 1, 2, 9, 13) identified only 21%, 46%, 49%  
234 and 34% of the total number of 1000 reads, respectively (Table 3). Participant 4 was unable to

235 identify MeV and participant 7 assigned too many reads (1,411) as originating from the mutated  
236 MeV.

237 The divergent nABV (actual read count = 500) proved to be the most challenging target and was  
238 identified by only four of the participants (participants 3, 5, 6, 12) (Tables 3-4, Fig. 1). The  
239 overall specificity for all bioinformatics workflows was high, with only participant 6 identifying  
240 43 reads of a chordopoxvirus as a false positive result.

241

242 The total time of analysis ranged widely, from three hours (participant 1) to 216 hours (online  
243 analysis of 15 hours with additional 201 hours of waiting time for server availability, participant  
244 4) (Table 5). Most workflows were calculated on a server system; two participants used a  
245 personal computer and two participants a virtual machine. One calculation was executed through  
246 an external public server.

247 Most of the workflows used in the COMPARE virus PT were quite similar, with the same basic  
248 tasks applied in different order (Fig. 2). Most workflows started with trimming and quality  
249 filtering, then subtraction of background reads, assembly of remaining reads, and a final  
250 reference-based viral read assignment (Fig. 1). Databases used were custom-made or full  
251 databases from NCBI nt/nr GenBank (participants 1-4, 6-11, 13). Participants 5 and 12 used viral  
252 sequences from NCBI GenBank only, while participant 7 also included a database for human  
253 pathogenic viruses (ViPR) (<https://www.viprbrc.org/brc/home.spg?decorator=vipr>).

254 All groups were also asked to correlate the results based on the bioinformatics analysis with the  
255 clinical symptoms described in the case report (Table 4). HSV-1 was suspected as the disease-  
256 causing agent by three groups and MeV was identified by six groups. An MeV infection with  
257 HSV-1 possibly affecting the course of disease was named by two groups. nAVB was interpreted  
258 as the single causative agent by two groups.

259

260 **Discussion**

261 HTS-based virus diagnostics requires a complex multistep processing, including laboratory  
262 preparation, assessment of quality of sequences produced, computationally challenging analytic  
263 validation of sequence reads, and post-analytic interpretation of results. Therefore, not only  
264 comprehensive technical skills, but also bioinformatic, biological, and medical knowledge are of  
265 paramount importance for proper analyses of HTS data for virus diagnostics.

266 HTS data can comprise several hundred thousand to many millions of reads from a single  
267 sequenced sample. Handling and analyzing such amounts of data pose computational challenges  
268 and currently require know-how and expertise in bioinformatics. Depending on the laboratory  
269 procedure, identification of viral reads from clinical metagenomics data is negatively affected by  
270 low virus-to-host sequence ratios and high viral mutation rates, making reference-based sequence  
271 assignments for highly divergent viruses challenging (24).

272 The *In silico* bioinformatics analysis of HTS data can be separated into an analytic and a post-  
273 analytic step. The analytic step includes the processing of sequence reads with software tools or  
274 scripts assembled into workflows and pipelines. The post-analytic step is the evaluation of the  
275 results obtained from the bioinformatics analysis, regarding pathogen identification often  
276 involving the interpretation by an experienced qualified health professional to correlate  
277 bioinformatics results with the clinical and epidemiological patient information.

278 The bioinformatics analysis and the technical identification of viral reads from the HTS dataset,  
279 was shown to have a decreasing success as sequences became more divergent from reference  
280 strains, exemplified by MeV with 82% identity on nucleotide level to its closest relative and  
281 nABV with just 52 % identity on nucleotide level to other bornaviruses, identified by only four of  
282 the 13 participants. MeV and TTV were missed by participant 4 whose analysis was based on the

283 Kraken tool and an in-house workflow. Kraken is known to align sequence reads to the reference  
284 sequences with a high specificity and low sensitivity, making the alignment of mutated and  
285 divergent virus reads difficult (15). As Kraken uses a user-specific reference database the TTV  
286 may have been absent from the custom database, since Kraken was also used by participant 7,  
287 who was able to identify both MeV and TTV. It is noted that the use of different databases is an  
288 obstacle in bioinformatics analysis of HTS data. So far, there were only unified, curated virus  
289 reference databases for Influenza viruses (EpiFlu) (25), HIV (26) and human pathogenic viruses  
290 (ViPR) (27). Recently, viral reference databases for bioinformatics analysis of HTS data have  
291 been developed (<https://hive.biochemistry.gwu.edu/rvdb>), (<https://rvdb-prot.pasteur.fr/>) (28).  
292 NCBI offers the most extensive collection of viral genomes, but the lack of curation and  
293 verification of submitted sequences often leads to false positive and false negative results. To  
294 overcome such problems, reference-independent tools for virus detection of HTS data have been  
295 developed, also making the discovery of novel viruses feasible without any knowledge of the  
296 reference genome (29). All of the participants who were able to identify the divergent nABV  
297 used workflows based on protein alignment approaches, including BLASTx/p, USEARCH, or  
298 DIAMOND, which are known to be highly sensitive (14, 17). The identification of such highly-  
299 divergent viruses is still challenging and cannot be accomplished by workflows based on  
300 nucleotide-only reference-based alignment approaches. DIAMOND (double index alignment of  
301 next-generation sequencing data), which became available in 2015, was specifically designed for  
302 such sensitive analysis of HTS data at the protein level, and is up to 20,000 times faster than  
303 BLAST programs. Compared to other alignment tools which seem to have a trade-off between  
304 speed and sensitivity, DIAMOND offers superior sensitivity for the detection of mutated and  
305 divergent viral sequences (14). However, the detection of such highly divergent viral sequences  
306 in patient samples is rare, and virus discovery is not a routine part of clinical virus diagnostics.

307 In terms of specificity, all workflows were highly specific, with only workflow 6 showing the  
308 identification of a chordopoxvirus which was not present in the dataset. Such false positives, as  
309 well as the excessive number of HSV-1 and MeV reads found by participant 7 (8,361 of 2,000  
310 reads and 1,411 of 1,000 reads, respectively) can derive, for example, from low-complexity reads  
311 in the dataset which are aligned to low-complexity or repetitive sequences of the viral reference  
312 genomes, from inappropriate matching score limits during filtering, or inappropriate algorithm  
313 parameters. Furthermore, custom databases and viral references from NCBI can include  
314 sequences of human origin which can lead to false positive results, which in some cases can  
315 result in the non-reporting of other matches due to default algorithm reporting limits.

316 The total time of all workflows differed widely from only three hours to 216 hours (15 hours for  
317 the analysis and 201 hours waiting time for available servers). One of the fastest participants was  
318 participant 1 who needed only 3 hours to perform the calculations on a scalable high-performance  
319 national virtual machine, whereas the slowest workflow (participant 4; 216 hours) was calculated  
320 on a personal computer, through an external public server where bioinformatics software jobs are  
321 queued among many other users (Fig. 1, Table 6). However, participant 5 also performed analysis  
322 on a notebook but within a much shorter time (26 hours). Overall, workflows exclusively  
323 specified for virus detection or using only viral or refSeq databases did not clearly correlate with  
324 faster workflow times compared to full metagenomics analyses. However, the specific  
325 composition of each database was not provided. To finally evaluate the performance of each  
326 bioinformatics workflow regarding the time of analysis, all workflows should be run on the same  
327 computer system, but such standardization was not practical for this PT evaluation.

328 The COMPARE virus PT has further shown that analytic and post-analytic evaluation is both of  
329 importance, as similar analytic results can be interpreted very differently, depending on the  
330 analyzing participant. Unlike standard routine virus diagnostic approaches such as polymerase

331 chain reaction, where a medical hypothesis of relevance is tested either positive or negative, HTS  
332 offers an extensive and largely unbiased catalogue of results. The etiological agent of a patient  
333 sample can be masked by false positives, sequencing contaminants, commensal viruses of the  
334 human virome, or viruses of yet unknown importance. Furthermore, the causative viral agent of a  
335 disease may be present in very low read numbers because viral loads may be low, depending on  
336 the timing of sampling and the sample matrix. RNA viruses, some of which are the most  
337 pathogenic human viruses, usually have smaller genomes than DNA viruses (30, 31). Therefore,  
338 low read numbers from an RNA virus might be dismissed, resulting in a false negative. To assess  
339 sequencing results, some workflows and pipelines use cutoffs for read numbers so as to reduce  
340 false positives, but may in the process make the detection of low read-number matches less  
341 likely.

342 As the analysis of HTS data for virus diagnostics requires bioinformatics as well as virological  
343 knowledge, the collaboration of both disciplines has been emphasized (32). Furthermore,  
344 automated pipelines for HTS-based virus diagnostics with unbiased evaluation of pathogenicity  
345 and relevance of the detected pathogen have been implemented, which can render analysis and  
346 interpretation of HTS sequence results more harmonized (33).

347 A robust approach to viral diagnostics using HTS requires further refinement and validation. The  
348 COMPARE *in silico* PT is limited by the low complexity of the simulated dataset. *In Vivo*  
349 sequence datasets can consist of a high diversity of the background and microbiome of the host,  
350 which further increases the difficulty to identify viral reads. Further proficiency schemes with *in*  
351 *vivo* datasets and samples and wider collaboration are required to make progress. A second *in*  
352 *silico* PT organized by the COMPARE network has focused on the interpretation of the  
353 significance of food-borne pathogens in a simulated dataset (data not published). Again, the  
354 interpretation of the results was shown to be one of the most diverse and critical points in HTS



355 data analysis. Furthermore, third-generation sequencing technologies, such as the MinION from  
356 Oxford Nanopore Technologies, are becoming available in many laboratories and field settings  
357 due to low cost and short sequencing times (34-36). However, analysis tools developed for  
358 second-generation sequencing technologies, such as Illumina, may not be applicable for third-  
359 generation sequencing data, due to the low sequencing accuracy of approximately 85 % and of  
360 and the length of the sequences, which can be up to 2Mbp (37-39). Consequently, future PTs  
361 should also include the use of third-generation sequencing technologies, as those are likely to  
362 become part of future routine laboratory diagnostics.

363

#### 364 **Conclusion**

365 The present availability of External Quality Assessment for HTS-based virus identification is  
366 limited. The COMPARE *in silico* virus PT has shown that numerous tools and different  
367 workflows are used for virus analysis of HTS data, and results of such workflows differ in  
368 sensitivity and specificity. At the present time, there are no standard procedures for virome  
369 analyses, and the sharing, comparing, and reliable production of results of such analyses are  
370 difficult.

371 Finally, there is a clear need for creating updated and highly curated, publicly freely available  
372 databases for harmonized identification of virus in virome datasets, as well as mechanisms for  
373 conducting continuous ringtails to ensure the quality of virus diagnostic and characterization in  
374 clinical diagnostic and public and veterinary health laboratories.

375 **Acknowledgments**

376 This study was supported by EU Horizon 2020 funding for COMPARE Europe (grant agreement  
377 no. 643476). We thank Ursula Erikli for copyediting.

378

379 **References**

- 380 1. McMullan LK, Folk SM, Kelly AJ, MacNeil A, Goldsmith CS, Metcalfe MG, Batten BC, Albarino CG,  
381 Zaki SR, Rollin PE, Nicholson WL, Nichol ST. 2012. A new phlebovirus associated with severe  
382 febrile illness in Missouri. *N Engl J Med* 367:834-41.
- 383 2. Grard G, Fair JN, Lee D, Slikas E, Steffen I, Muyembe JJ, Sittler T, Veeraraghavan N, Ruby JG,  
384 Wang C, Makuwa M, Mulembakani P, Tesh RB, Mazet J, Rimo AW, Taylor T, Schneider BS,  
385 Simmons G, Delwart E, Wolfe ND, Chiu CY, Leroy EM. 2012. A novel rhabdovirus associated with  
386 acute hemorrhagic fever in central Africa. *PLoS Pathog* 8:e1002924.
- 387 3. Briese T, Paweska JT, McMullan LK, Hutchison SK, Street C, Palacios G, Khristova ML, Weyer J,  
388 Swanepoel R, Egholm M, Nichol ST, Lipkin WI. 2009. Genetic detection and characterization of  
389 Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. *PLoS Pathog*  
390 5:e1000455.
- 391 4. Hoffmann B, Tappe D, Hoper D, Herden C, Boldt A, Mawrin C, Niederstrasser O, Muller T, Jenckel  
392 M, van der Grinten E, Lutter C, Abendroth B, Teifke JP, Cadar D, Schmidt-Chanasit J, Ulrich RG,  
393 Beer M. 2015. A Variegated Squirrel Bornavirus Associated with Fatal Human Encephalitis. *N Engl*  
394 *J Med* 373:154-62.
- 395 5. Moran-Gilad J, Sintchenko V, Pedersen SK, Wolfgang WJ, Pettengill J, Strain E, Hendriksen RS,  
396 Global Microbial Identifier initiative's Working G. 2015. Proficiency testing for bacterial whole  
397 genome sequencing: an end-user survey of current capabilities, requirements and priorities. *BMC*  
398 *Infect Dis* 15:174.
- 399 6. Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW, Timme R. 2016. Practical Value  
400 of Food Pathogen Traceability through Building a Whole-Genome Sequencing Network and  
401 Database. *J Clin Microbiol* 54:1975-83.
- 402 7. Timme RE, Rand H, Sanchez Leon M, Hoffmann M, Strain E, Allard M, Roberson D, Baugher JD.  
403 2018. GenomeTrakr proficiency testing for foodborne pathogen surveillance: an exercise from  
404 2015. *Microb Genom* 4.
- 405 8. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data.  
406 *Bioinformatics* 30:2114-20.
- 407 9. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data.
- 408 10. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.  
409 *Bioinformatics* 25:1754-60.
- 410 11. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357-9.
- 411 12. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI,  
412 Pham S, Pribelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA.  
413 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J*  
414 *Comput Biol* 19:455-77.
- 415 13. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn  
416 graphs. *Genome Res* 18:821-9.

- 417 14. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat*  
418 *Methods* 12:59-60.
- 419 15. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact  
420 alignments. *Genome Biology* 15:R46.
- 421 16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool.  
422 *Journal of Molecular Biology* 215:403-410.
- 423 17. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*  
424 26:2460-1.
- 425 18. Scheuch M, Hoper D, Beer M. 2015. RIEMS: a software pipeline for sensitive and comprehensive  
426 taxonomic classification of reads from metagenomics datasets. *BMC Bioinformatics* 16:69.
- 427 19. Vilsker M, Moosa Y, Nooij S, Fonseca V, Ghysens Y, Dumon K, Pauwels R, Alcantara LC, Vanden  
428 Eynden E, Vandamme AM, Deforche K, de Oliveira T. 2018. Genome Detective: An Automated  
429 System for Virus Identification from High-throughput sequencing data. *Bioinformatics*  
430 doi:10.1093/bioinformatics/bty695.
- 431 20. Andrusch A, Dabrowski PW, Klenner J, Tausch SH, Kohl C, Osman AA, Renard BY, Nitsche A.  
432 2018. PAIPline: pathogen identification in metagenomic and clinical next generation sequencing  
433 samples. *Bioinformatics* 34:i715-i721.
- 434 21. Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator.  
435 *Bioinformatics* 28:593-4.
- 436 22. Rota PA, Moss WJ, Takeda M, de Swart RL, Thompson KM, Goodson JL. 2016. Measles. *Nature*  
437 *Reviews Disease Primers* 2:16049.
- 438 23. Bradshaw MJ, Venkatesan A. 2016. Herpes Simplex Virus-1 Encephalitis in Adults:  
439 Pathophysiology, Diagnosis, and Management. *Neurotherapeutics* 13:493-508.
- 440 24. Hoper D, Mettenleiter TC, Beer M. 2016. Metagenomic approaches to identifying infectious agents.  
441 *Rev Sci Tech* 35:83-93.
- 442 25. Shu Y, McCauley J. 2017. GISAID: Global initiative on sharing all influenza data - from vision to  
443 reality. *Euro Surveill* 22.
- 444 26. Druce M, Hulo C, Masson P, Sommer P, Xenarios I, Le Mercier P, De Oliveira T. 2016. Improving  
445 HIV proteome annotation: new features of BioAfrica HIV Proteomics Resource. *Database (Oxford)*  
446 2016.
- 447 27. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, Liu M, Kumar S, Zaremba S, Gu  
448 Z, Zhou L, Larson CN, Dietrich J, Klem EB, Scheuermann RH. 2012. ViPR: an open bioinformatics  
449 database and analysis resource for virology research. *Nucleic Acids Res* 40:D593-8.
- 450 28. Goodacre N, Aljanahi A, Nandakumar S, Mikailov M, Khan AS. 2018. A Reference Viral Database  
451 (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus  
452 Detection. *mSphere* 3:e00069-18.
- 453 29. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, Xie X, Sun F. 2018. Identifying viruses from  
454 metagenomic data by deep learning. *arXiv e-prints*.
- 455 30. ME JW, Adair K, Brierley L. 2013. RNA Viruses: A Case Study of the Biology of Emerging Infectious  
456 Diseases. *Microbiol Spectr* 1.
- 457 31. Woolhouse ME, Brierley L, McCaffery C, Lycett S. 2016. Assessing the Epidemic Potential of RNA  
458 and DNA Viruses. *Emerg Infect Dis* 22:2037-2044.
- 459 32. Hufsky F, Ibrahim B, Beer M, Deng L, Mercier PL, McMahon DP, Palmarini M, Thiel V, Marz M.  
460 2018. Virologists-Heroes need weapons. *PLoS Pathog* 14:e1006771.
- 461 33. Tausch SH, Loka TP, Schulze JM, Andrusch A, Klenner J, Dabrowski PW, Lindner MS, Nitsche A,  
462 Renard BY. 2018. doi:10.1101/402370.
- 463 34. Kafetzopoulou LE, Efthymiadis K, Lewandowski K, Crook A, Carter D, Osborne J, Aarons E,  
464 Hewson R, Hiscox JA, Carroll MW, Vipond R, Pullan ST. 2018. Assessment of metagenomic

- 465 Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and  
466 dengue viruses directly from clinical samples. *Eurosurveillance* 23:1800228.
- 467 35. Cheng J, Hu H, Kang Y, Chen W, Fang W, Wang K, Zhang Q, Fu A, Zhou S, Cheng C, Cao Q,  
468 Wang F, Lee S, Zhou Z. 2018. Identification of pathogens in culture-negative infective endocarditis  
469 cases by metagenomic analysis. *Ann Clin Microbiol Antimicrob* 17:43.
- 470 36. Kafetzopoulou LE, Pullan ST, Lemey P, Suchard MA, Ehichioya DU, Pahlmann M, Thielebein A,  
471 Hinzmann J, Oestereich L, Wozniak DM, Efthymiadis K, Schachten D, Koenig F, Matjeschk J,  
472 Lorenzen S, Lumley S, Ighodalo Y, Adomeh DI, Olorok T, Omomoh E, Omiunu R, Agbukor J, Ebo  
473 B, Aiyepada J, Ebhodaghe P, Osiemi B, Ehikhametalor S, Akhilomen P, Airende M, Esumeh R,  
474 Muoebonam E, Giwa R, Ekanem A, Igenegbale G, Odigie G, Okonofua G, Enigbe R, Oyakhilome J,  
475 Yerumoh EO, Odia I, Aire C, Okonofua M, Atafa R, Tobin E, Asogun D, Akpede N, Okokhere PO,  
476 Rafiu MO, Iraoyah KO, Iruolagbe CO, et al. 2019. Metagenomic sequencing at the epicenter of the  
477 Nigeria 2018 Lassa fever outbreak. *Science* 363:74-77.
- 478 37. Payne A, Holmes N, Rakyan V, Loose M. 2018. BulkVis: a graphical viewer for Oxford nanopore  
479 bulk FAST5 files. *Bioinformatics* doi:10.1093/bioinformatics/bty841.
- 480 38. Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, Pregno G,  
481 Giachino D, Mandrile G, Espejo Valle-Inclan J, Korzelius J, de Bruijn E, Cuppen E, Talkowski ME,  
482 Marschall T, de Ridder J, Kloosterman WP. 2017. Mapping and phasing of structural variation in  
483 patient genomes using nanopore sequencing. *Nat Commun* 8:1326.
- 484 39. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dillthey AT, Fiddes  
485 IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan  
486 AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2018. Nanopore  
487 sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* 36:338-345.
- 488 40. Bellod Cisneros JL, Lund O. 2017. KmerFinderJS: A Client-Server Method For Fast Species Typing  
489 Of Bacteria Over Slow Internet Connections. *bioRxiv* doi:10.1101/145284.
- 490 41. Petersen TN, Lukjancenko O, Thomsen MCF, Maddalena Sperotto M, Lund O, Moller Aarestrup F,  
491 Sicheritz-Ponten T. 2017. MGmapper: Reference based mapping and taxonomy annotation of  
492 metagenomics sequence reads. *PLoS One* 12:e0176469.
- 493 42. Minot SS, Krumm N, Greenfield NB. 2015. One Codex: A Sensitive and Accurate Data Platform for  
494 Genomic Microbial Identification. *bioRxiv* doi:10.1101/027607.
- 495 43. Gaidatzis D, Lerch A, Hahne F, Stadler MB. 2015. QuasR: quantification and annotation of short  
496 reads in R. *Bioinformatics* 31:1130-2.
- 497 44. Jiang H, Lei R, Ding S-W, Zhu S. 2014. Skewer: a fast and accurate adapter trimmer for next-  
498 generation sequencing paired-end reads. *BMC Bioinformatics* 15:182.
- 499 45. Zaharia M, J. Bolosky W, Curtis K, Fox A, Patterson D, Shenker S, Stoica I, Karp R, Sittler T. 2011.  
500 Faster and More Accurate Sequence Alignment with SNAP, vol 1111.
- 501 46. Flygare S, Simmon K, Miller C, Qiao Y, Kennedy B, Di Sera T, Graf EH, Tardif KD, Kapusta A,  
502 Ryneerson S, Stockmann C, Queen K, Tong S, Voelkerding KV, Blaschke A, Byington CL, Jain S,  
503 Pavia A, Ampofo K, Eilbeck K, Marth G, Yandell M, Schlaberg R. 2016. Taxonomer: an interactive  
504 metagenomics analysis portal for universal pathogen detection and host mRNA expression  
505 profiling. *Genome Biol* 17:111.
- 506
- 507
- 508

509 Figure legends

510

511 **FIG 1 Identified viral read numbers for Torque Teno virus (TTV), human herpesvirus**

512 **(HSV-1), measles virus (MeV), and new avian bornavirus (nABV) by participant**

513 **(numbered 1-13).**

514

515 **FIG 2 Simplified comparison of different bioinformatics workflows for virus identification**

516 **used in the COMPARE virus proficiency test**

517 **+ Human herpesvirus, + Torque teno virus; + Measles virus; + Avian bornavirus**

518

519 **TABLE 1 Tools and programs for analysis of HTS data used in the COMPARE Virus Proficiency Test, in alphabetical order**

Program	Application	Description/relevance for viral HTS	URL
<b>BWA (10)</b>	Alignment (nucleotide)	BWA (Burrows-Wheeler Alignment Tool) to align efficiently short sequencing reads against a large reference genome. Based on string matching with Burrows-Wheeler transform (BWT).	<a href="http://bio-bwa.sourceforge.net/">http://bio-bwa.sourceforge.net/</a>
<b>DIAMOND (14)</b>	Alignment (protein)	Double-index alignment of NGS data. Shown to be up to 20,000 times faster than comparable programs, with high sensitivity.	<a href="http://ab.inf.uni-tuebingen.de/software/diamond/">http://ab.inf.uni-tuebingen.de/software/diamond/</a>
<b>FastQC (9)</b>	Quality control, trimming	Generates base quality scores and sequence contents, sequence length distributions, identification of duplicate or overrepresented sequences, adapter, and k-mer contents.	<a href="https://www.bioinformatics.babraham.ac.uk/projects/fastqc/">https://www.bioinformatics.babraham.ac.uk/projects/fastqc/</a>
<b>Kmerfinder (40)</b>	Taxonomic assignment	Online user interface also allows the prediction of human and vertebrate viruses.	<a href="https://cge.cbs.dtu.dk/services/KmerFinder/">https://cge.cbs.dtu.dk/services/KmerFinder/</a>
<b>Kraken (15)</b>	Alignment (nucleotide)	Only uses exact alignments for its taxonomic classification with high speed and less computational requirements.	<a href="https://ccb.jhu.edu/software/kraken/">https://ccb.jhu.edu/software/kraken/</a>
<b>MetaPhlan</b>	Taxonomic assignment	Metagenomic Phylogenetic Analysis is a tool for the taxonomic assignment of microbial communities. High accuracy and speed are supported by only high-confidence matches. Such approaches allow the assignment of 25,000 microbial reads per second but might fail with	<a href="https://bitbucket.org/biobakery/metaphlan2">https://bitbucket.org/biobakery/metaphlan2</a>

		viral genomes which often lack common markers and genes.	
<b>MGMapper</b> (41)	Pipeline	Online tool for processing, assigning, and analyzing HTS sequences.	<a href="https://cge.cbs.dtu.dk/services/MGmapper/">https://cge.cbs.dtu.dk/services/MGmapper/</a> , <a href="https://bitbucket.org/genomicepidemiology/mgmapper">https://bitbucket.org/genomicepidemiology/mgmapper</a>
<b>MIRA</b>	<i>De novo</i> assembly	Mimicking Intelligent Read Assembly, overlap-layout-consensus graph (OLC) assembler for metagenomics data from several sequencing platforms. Assembles the most as well as the largest contigs compared to other <i>de novo</i> assembly programs, as well as produces the highest number of contigs which could be assigned to a viral taxon.	<a href="https://sourceforge.net/projects/mira-assembler/">https://sourceforge.net/projects/mira-assembler/</a>
<b>NCBI BLAST (16)</b>	Alignment (nucleotide and protein)	Basic local alignment search tool. Offers very sensitive online and stand-alone alignments of nucleotides, translated nucleotides, and protein sequences.	<a href="https://blast.ncbi.nlm.nih.gov/Blast.cgi">https://blast.ncbi.nlm.nih.gov/Blast.cgi</a>
<b>One Codex</b> (42)	Taxonomic assignment	Web-based data platform for k-mer based taxonomic classification. Very high degree of sensitivity and specificity, even when analyzing highly divergent and mutated sequences.	<a href="https://www.onecodex.com/">https://www.onecodex.com/</a>
<b>PAIPline</b> (20)	Pipeline	Pipeline for metagenomic analysis of HTS data.	<a href="https://gitlab.com/rki_bioinformatics/pipeline">https://gitlab.com/rki_bioinformatics/pipeline</a>

<b>QUASR (43)</b>	Pipeline	Combination of several R packages and external software for HTS read analysis. Part of the Bioconductor project.	<a href="http://www.bioconductor.org">http://www.bioconductor.org</a>
<b>RIEMS (18)</b>	Pipeline	Pipeline for metagenomics sequence analysis, combining several established programs and tools for pathogen detection in one automated workflow. Separated into a workflow of accurate and fast “basic analysis” and a more sensitive “further analysis”.	<a href="https://www.fli.de/en/institutes/institute-of-diagnostic-virology-ivd/laboratories-working-groups/laboratory-for-ngs-and-microarray-diagnostics/">https://www.fli.de/en/institutes/institute-of-diagnostic-virology-ivd/laboratories-working-groups/laboratory-for-ngs-and-microarray-diagnostics/</a>
<b>Skewer (44)</b>	Quality control, trimming	Trimming of primer and adapter sequences focusing on the characteristics of paired-end and mate-pair reads. A statistical scheme based on quality values allows the accurate trimming of adapters with mismatches.	<a href="https://sourceforge.net/projects/skewer/">https://sourceforge.net/projects/skewer/</a>
<b>SNAP (45)</b>	Alignment (nucleotide)	Up to 10 to 100 times faster than similar alignment programs but offers greater sensitivity due to richer error acceptance.	<a href="http://snap.cs.berkeley.edu/">http://snap.cs.berkeley.edu/</a>
<b>SPAdes, MetaSPAdes (12)</b>	<i>De novo</i> assembly	De Bruijn graph assembler. MetaSPAdes specifically addresses the challenges that arise with complex metagenomics data.	<a href="http://cab.spbu.ru/software/spades/">http://cab.spbu.ru/software/spades/</a>
<b>Taxonomer (46)</b>	Taxonomic assignment	Web-based tool for nucleotide- and protein-based read assignment. User-friendly interactive result visualization. Based on exact k-mer matching	<a href="http://taxonomer.iobio.io/">http://taxonomer.iobio.io/</a>



		with low error tolerance. Speed up to ~32 million reads/minute. Furthermore, protein-based read identification offers the detection of divergent viral sequences but is based on exact k-mer matching without error allowance.	
<b>Trimmomatic (8)</b>	Quality control, trimming	Paired-end sequence reads can be cut from technical sequences as adapters, primers, or low-quality bases. Has been shown to improve considerably downstream analyses, for example <i>de novo</i> assembly (increasing contig size up to 77 %) and alignment (increasing alignment rates from 7 % to 78 %).	<a href="http://www.usadellab.org/cms/index.php?page=trimmomatic">www.usadellab.org/cms/index.php?page=trimmomatic</a>
<b>USEARCH (17)</b>	Alignment (protein)	Exceptionally high speed for protein or translated nucleotide read alignment. The sensitivity of USEARCH is comparable to the NCBI protein BLAST, but USEARCH is ~350 times faster.	<a href="https://www.drive5.com/usearch/">https://www.drive5.com/usearch/</a>
<b>Velvet (13)</b>	<i>De novo</i> assembly	Can be used for <i>de novo</i> assemblies of short HTS reads using the de Bruijn algorithm. <i>de novo</i> assembly using Velvet can be achieved in as little as 14 minutes.	<a href="https://www.ebi.ac.uk/~zerbino/velvet/">https://www.ebi.ac.uk/~zerbino/velvet/</a>

520

521

522 **TABLE 2 Composition of the simulated sequence dataset. Total number of reads are**523 **6,339,908.**

<b>Organism</b>	<b>#Reads</b>	<b>Nucleotide sequence identity with reference (%)</b>
Human	4,834,491	100
<i>Acinetobacter johnsonii</i>	500,000	100
<i>Propionibacterium acnes</i>	500,000	100
<i>Staphylococcus epidermidis</i>	500,000	100
<b>Torque teno virus</b>	1,917	100
<b>Human herpesvirus 1</b>	2,000	100
<b>Measles virus</b>	1,000	82
<b>(Novel) Avian bornavirus</b>	500	55

524

525 **TABLE 3 Sensitivity and specificity for identified reads of the COMPARE virus proficiency**526 **test. Participants were numbered randomly.**

	<b>Torque teno virus</b>	<b>Human herpesvirus</b>	<b>Measles virus</b>	<b>Avian bornavirus</b>	<b>No false positive result</b>	<b>Time of analysis (h)</b>
#1	1	0.99	0.21	0	√	3
#2	1	1.01	0.46	0	√	15.5
#3	0.96	0.96	1	1	√	60
#4	0	0.10	0	0	√	216
#5	1	0.98	1	1	√	26

#6	1	0.84	1	1	-	12
#7	0.94	4.00	1.41	0	√	6
#8	1	1.04	0.99	0	√	7
#9	0.29	0.84	0.49	0	√	5
#10	1	1	1	0	√	48
#11	1	1	1	0	√	14
#12	1	1	1.02	0.23	√	18
#13	1.02	0.90	0.34	0	√	48

527

528 **TABLE 4 Interpretation of bioinformatics results.** Abbreviations: TTV = Torque teno virus,

529 HSV-1 = human herpesvirus 1, MeV = measles virus, nABV = new avian bornavirus

	Results bioinformatics	Results diagnostics	Participants' background
#1	TTV, HSV-1, MeV	HSV-1	Bioinformatics
#2	TTV, HSV-1, MeV	HSV-1	Food & environmental health
#3	TTV, HSV-1, MeV, nABV	SSPE/HSV-1	Veterinarian, virology
#4	HSV-1	HSV-1	University, virology
#5	TTV, HSV-1, MeV, nABV	nABV	Virology
#6	TTV, HSV-1, MeV, nABV	nABV	Medical research
#7	TTV, HSV-1, MeV	SSPE	Animal and plant health
#8	TTV, HSV-1, MeV	SSPE	Veterinarian, virology
#9	TTV, HSV-1, MeV	SSPE	Public health
#10	TTV, HSV-1, MeV	SSPE	Public health
#11	TTV, HSV-1, MeV	SSPE	Public health and environment

#12 TTV, HSV-1, MeV, nABV	SSPE/HSV-1	Diagnostics, virology
#13 TTV, HSV-1, MeV	SSPE	Virology

530

531

532 **TABLE 5 Total time of computational analysis, maximum computer/server specifications,**  
 533 **and reference databases used.**

	Time of analysis (h)	Database	Operating system	CPU	CPU Mhz	RAM (GB)
#1	3	NCBI nt	UNIX Ubuntu 16.04	VM	VM	VM
#2	15.5	NCBI nt	LTS	56	1270	378
#3	60	NCBI nt/nr	CentOS 6	24	2400	64
#4	216	NCBI nt	Windows XP	intel core i5	2300	8
#5	26	NCBI viral db	OS X	2	na	na
#6	12	NCBI nr VIPR and	Ubuntu 14.04 BioLinux	32	2000	503
#7	6	NCBI nt	Ubuntu 14.04	8	3.6	16
#8	7	NCBI nt	CentOS 6.5	64	2300	250
#9	5	NCBI nr	Ubuntu 12.04.5	na	3800	50
#10	48	NCBI nt	CentOS 6.5	2 × AMD Opteron	2200	32
					VM,	VM,
#11	14	NCBI nt/nr	RHEL	VM, variable	variable	variable
#12	18	NCBI viral db	Linux Mint	Intel Xenon	6 × 2.67	25

28

X5650

Ghz

Ubuntu 14.04.4 2 × AMD Opteron 24 × 2.2

#13 48

NCBI nt

LTS

6174

GHz

128

---

534 db=database; na= not available; nr=non-redundant; nt=nucleotide; VM=virtual machine

535



