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

Total number of authors:
36

Published in:
Journal of Clinical Microbiology

Link to article, DOI:
10.1128/JCM.00466-19

Publication date:
2019

Document Version
Peer reviewed version

JCM Accepted Manuscript Posted Online 5 June 2019
J. Clin. Microbiol. doi:10.1128/JCM.00466-19

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

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

[a]Robert Koch Institute, Centre for Biological Threats and Special Pathogens 1, Berlin, Germany
[b]Friedrich-Loeffler-Institut, Institute of Diagnostic Virology, Greifswald–Insel Riems, Germany
[c]Technical University of Denmark, National Food Institute, WHO Collaborating Center for Antimicrobial Resistance in Foodborne Pathogens and Genomics and European Union Reference Laboratory for Antimicrobial Resistance, Kgs. Lyngby, Denmark
[d]French Agency for Food, Environmental and Occupational Health & Safety, Laboratory of Ploufragan, Unit of Viral Genetics and Biosafety, Ploufragan, France
[e]Aristotle University of Thessaloniki, School of Medicine, Microbiology Department, Thessaloniki, Greece
[f]Erasmus Medical Centre, Department of Viroscience, Rotterdam, The Netherlands
[g]REGA Institute KU Leuven, Leuven, Belgium
[h]Animal and Plant Health Agency, Addlestone, United Kingdom

[i]University of Veterinary Medicine Hannover, Department of Pathology, Hannover, Germany

[j]University of Utrecht, Department of Infectious Diseases and immunity, Utrecht, The Netherlands

[k]Artemis One Health Research Institute, Utrecht, the Netherlands

[l]Istituto Zooprofilattico Sperimentale dell'Abruzzo e Molise "G. Caporale", National Reference Center for Whole Genome Sequencing of microbial pathogens: database and bioinformatic analysis, Teramo, Italy

[m]Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health (INSA), Lisbon, Portugal

[n]National Institute for Public Health and the Environment, Bilthoven, The Netherlands

[o]Institute of Virology, Charité-Universitätsmedizin Berlin, Berlin, Germany

[p]Center for Pathogen Evolution, Department of Zoology, University of Cambridge, Cambridge, United Kingdom

Running Head: COMPARE *in silico* Virus Proficiency Test

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

Keywords:

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

Abstract

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

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

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

Introduction

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

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

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

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

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

Tools and programs for bioinformatics analysis

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

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

Journal of Clinical Microbiology

Methods

Organization

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

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

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

In silico HTS dataset

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

NZ_CP010350.1), *Proprionibacterium acnes* (NCBI accession NZ_CP012647.1) and *Staphylococcus epidermis* (NCBI accession NZ_CP009046.1). In addition to human and bacterial reads, simulated viral sequences of four viruses, Torque teno virus (TTV; NCBI accession NC_015783.1), human herpesvirus 1 (HSV-1; NCBI accession NC_001806.2), measles virus (MeV; NCBI accession NC_001498.1) and a novel avian bornavirus (nABV; NCBI accession JN014950.1) were included in different numbers and with different levels of similarity to known viruses present in databases (Table 2). TTV and HSV-1 were included in the panel as the easiest sequences to identify (with 1,917 and 2,000 reads respectively, and 100% nucleotide identity with the reference sequences), followed by a slightly altered MeV (1,000 reads, with 82% nucleotide identity to the reference genome) and, as the likely most difficult taxon, nABV (only 500 reads and 55% nucleotide identity to reference JN014950.1). The dataset has been uploaded to the European Nucleotide Archive with the study accession number PRJEB32470.

Participants

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

Case report

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

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

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

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

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

Reported PT results

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

Analysis of PT results

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

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

Results

PT results

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

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

Journal of Clinical Microbiology

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

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

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

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

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

Discussion

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

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

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

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

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

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

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

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

Journal of Clinical Microbiology

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

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

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

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

Conclusion

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

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

Journal of Clinical Microbiology

Acknowledgments

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

References

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

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

Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. Eurosurveillance 23:1800228.

35. Cheng J, Hu H, Kang Y, Chen W, Fang W, Wang K, Zhang Q, Fu A, Zhou S, Cheng C, Cao Q, Wang F, Lee S, Zhou Z. 2018. Identification of pathogens in culture-negative infective endocarditis cases by metagenomic analysis. Ann Clin Microbiol Antimicrob 17:43.

36. Kafetzopoulou LE, Pullan ST, Lemey P, Suchard MA, Ehichioya DU, Pahlmann M, Thielebein A, Hinzmann J, Oestereich L, Wozniak DM, Efthymiadis K, Schachten D, Koenig F, Matjeschk J, Lorenzen S, Lumley S, Ighodalo Y, Adomeh DI, Olokor T, Omomoh E, Omiunu R, Agbukor J, Ebo B, Aiyepada J, Ebhodaghe P, Osiemi B, Ehikhametalor S, Akhilomen P, Airende M, Esumeh R, Muoebonam E, Giwa R, Ekanem A, Igenegbale G, Odigie G, Okonofua G, Enigbe R, Oyakhilome J, Yerumoh EO, Odia I, Aire C, Okonofua M, Atafo R, Tobin E, Asogun D, Akpede N, Okokhere PO, Rafiu MO, Iraoyah KO, Iruolagbe CO, et al. 2019. Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. Science 363:74-77.

37. Payne A, Holmes N, Rakyan V, Loose M. 2018. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics doi:10.1093/bioinformatics/bty841.

38. Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, Pregno G, Giachino D, Mandrile G, Espejo Valle-Inclan J, Korzelius J, de Bruijn E, Cuppen E, Talkowski ME, Marschall T, de Ridder J, Kloosterman WP. 2017. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun 8:1326.

39. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36:338-345.

40. Bellod Cisneros JL, Lund O. 2017. KmerFinderJS: A Client-Server Method For Fast Species Typing Of Bacteria Over Slow Internet Connections. bioRxiv doi:10.1101/145284.

41. Petersen TN, Lukjancenko O, Thomsen MCF, Maddalena Sperotto M, Lund O, Moller Aarestrup F, Sicheritz-Ponten T. 2017. MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence reads. PLoS One 12:e0176469.

42. Minot SS, Krumm N, Greenfield NB. 2015. One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification. bioRxiv doi:10.1101/027607.

43. Gaidatzis D, Lerch A, Hahne F, Stadler MB. 2015. QuasR: quantification and annotation of short reads in R. Bioinformatics 31:1130-2.

44. Jiang H, Lei R, Ding S-W, Zhu S. 2014. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics 15:182.

45. Zaharia M, J. Bolosky W, Curtis K, Fox A, Patterson D, Shenker S, Stoica I, Karp R, Sittler T. 2011. Faster and More Accurate Sequence Alignment with SNAP, vol 1111.

46. Flygare S, Simmon K, Miller C, Qiao Y, Kennedy B, Di Sera T, Graf EH, Tardif KD, Kapusta A, Rynearson S, Stockmann C, Queen K, Tong S, Voelkerding KV, Blaschke A, Byington CL, Jain S, Pavia A, Ampofo K, Eilbeck K, Marth G, Yandell M, Schlaberg R. 2016. Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. Genome Biol 17:111.

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
BWA (10)	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).	http://bio-bwa.sourceforge.net/
DIAMOND (14)	Alignment (protein)	Double-index alignment of NGS data. Shown to be up to 20,000 times faster than comparable programs, with high sensitivity.	http://ab.inf.uni-tuebingen.de/software/diamond/
FastQC (9)	Quality control, trimming	Generates base quality scores and sequence contents, sequence length distributions, identification of duplicate or overrepresented sequences, adapter, and k-mer contents.	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Kmerfinder (40)	Taxonomic assignment	Online user interface also allows the prediction of human and vertebrate viruses.	https://cge.cbs.dtu.dk//services/KmerFinder/
Kraken (15)	Alignment (nucleotide)	Only uses exact alignments for its taxonomic classification with high speed and less computational requirements.	https://ccb.jhu.edu/software/kraken/
MetaPhlAn	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	https://bitbucket.org/biobakery/metaphlan2

		viral genomes which often lack common markers and genes.	
MGMapper (41)	Pipeline	Online tool for processing, assigning, and analyzing HTS sequences.	https://cge.cbs.dtu.dk/services/MGmapper/, https://bitbucket.org/genomicepidemiology/mgmapper
MIRA	*De novo* 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 *de novo* assembly programs, as well as produces the highest number of contigs which could be assigned to a viral taxon.	https://sourceforge.net/projects/mira-assembler/
NCBI BLAST (16)	Alignment (nucleotide and protein)	Basic local alignment search tool. Offers very sensitive online and stand-alone alignments of nucleotides, translated nucleotides, and protein sequences.	https://blast.ncbi.nlm.nih.gov/Blast.cgi
One Codex (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.	https://www.onecodex.com/
PAIPline (20)	Pipeline	Pipeline for metagenomic analysis of HTS data.	https://gitlab.com/rki_bioinformatics/paipline

QUASR (43)	Pipeline	Combination of several R packages and external software for HTS read analysis. Part of the Bioconductor project.	http://www.bioconductor.org
RIEMS (18)	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".	https://www.fli.de/en/institutes/institute-of-diagnostic-virology-ivd/laboratories-working-groups/laboratory-for-ngs-and-microarray-diagnostics/
Skewer (44)	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.	https://sourceforge.net/projects/skewer/
SNAP (45)	Alignment (nucleotide)	Up to 10 to 100 times faster than similar alignment programs but offers greater sensitivity due to richer error acceptance.	http://snap.cs.berkeley.edu/
SPAdes, MetaSPAdes (12)	*De novo* assembly	De Bruijn graph assembler. MetaSPAdes specifically addresses the challenges that arise with complex metagenomics data.	http://cab.spbu.ru/software/spades/
Taxonomer (46)	Taxonomic assignment	Web-based tool for nucleotide- and protein-based read assignment. User-friendly interactive result visualization. Based on exact k-mer matching	http://taxonomer.iobio.io/

		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.	
Trimmomatic (8)	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 *de novo* assembly (increasing contig size up to 77 %) and alignment (increasing alignment rates from 7 % to 78 %).	www.usadellab.org/cms/index.php?page=trimmomatic
USEARCH (17)	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.	https://www.drive5.com/usearch/
Velvet (13)	*De novo* assembly	Can be used for *de novo* assemblies of short HTS reads using the de Bruijn algorithm. *de novo* assembly using Velvet can be achieved in as little as 14 minutes.	https://www.ebi.ac.uk/~zerbino/velvet/

520

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

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

TABLE 3 Sensitivity and specificity for identified reads of the COMPARE virus proficiency test. Particiants were numbered randomly.

	Torque teno virus	Human herpesvirus	Measles virus	Avian bornavirus	No false positive result	Time of analysis (h)
#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

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

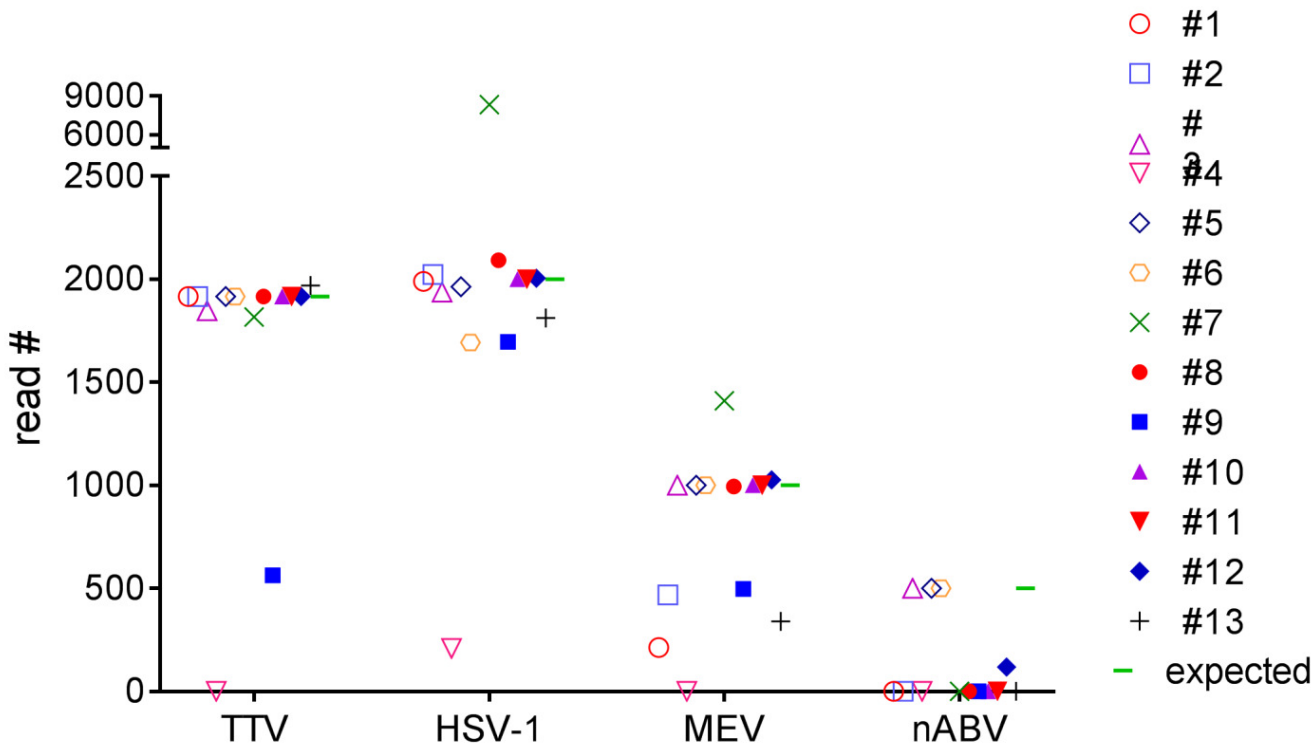
#12	TTV, HSV-1, MeV, nABV	SSPE/HSV-1	Diagnostics, virology
#13	TTV, HSV-1, MeV	SSPE	Virology

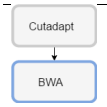
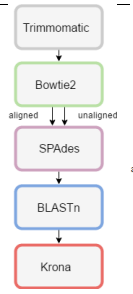
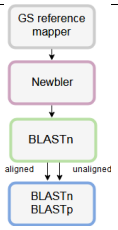
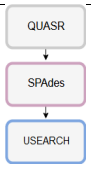
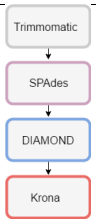
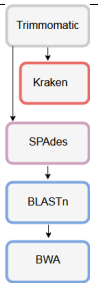
TABLE 5 Total time of computational analysis, maximum computer/server specifications, and reference databases used.

	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
#11	14	NCBI nt/nr	RHEL	VM, variable	VM, variable	VM, variable
#12	18	NCBI viral db	Linux Mint	Intel Xenon	6 × 2.67	25

				X5650	Ghz	
				Ubuntu 14.04.4 2 × AMD Opteron	24 × 2.2	
	#13	48	NCBI nt	LTS	6174	GHz	128

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



Workflow 1	Workflow 2	Workflow 3	Workflow 4	Workflow 5	Workflow 6	Workflow 7
+++ 3 h	+++ 15.5 h	++++ 60 h	+ 216 h	++++ 26 h	++++ 12 h	+++ 6 h
			 * QuasiBAM used for quality metrics and viral consensus sequences			

Workflow 8	Workflow 9	Workflow 10	Workflow 11	Workflow 12	Workflow 13	
+++ 7 h	+++ 5 h	+++ 48 h	+++ 14 h	++++ 18 h	+++ 48 h