Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples

Jutte J.C. de Vries[1], Julianne R. Brown[2], Nicole Fischer[3], Igor A. Sidorov[1], Sofia Morfopoulou[2a], Jiabin Huang[3], Bas B. Oude Munnink[4], Arzu Sayiner[5], Alihan Bulgurcu[5], Christophe Rodriguez[6], Guillaume Gricourt[6], Els Keyaerts[7], Leen Beller[7], Claudia Bachofen[8], Jakub Kubacki[8], Samuel Cordey[9], Florian Laubscher[9], Dennis Schmitz[10], Martin Beer[11], Dirk Hoeper[11], Michael Huber[12], Verena Kufner[12], Maryam Zaheri[12], Aitana Lebrand[13], Anna Papa[14], Sander van Boheemen[4], Aloys C.M. Kroes[1], Judith Breuer[2, 2a], F. Xavier Lopez-Labrador[15, 16], Eric C.J. Claas[1], on behalf of the ESCV Network on Next-Generation Sequencing

Word count: 2431

Author contributions: Conceptualization: JV, EC. Methodology: JV, EC, FXL, NF, IS, BO, SB. Data analysis: JV, JuB, NF, IS, SM, JH, BO, AS, AB, CR, GG, EK, LB, CB, JK, SC, FL, DS, MB, DH, MH, VK, MZ, AL, AP. Visualization: JV, AL. First draft: JV. Reviewing and editing: all authors

[1] Clinical Microbiological Laboratory, department of Medical Microbiology, Leiden University Medical Center, Leiden, the Netherlands; jjcdevries@lumc.nl, I.A.Sidorov@lumc.nl, A.C.M.Kroes@lumc.nl, E.C.J.Claas@lumc.nl

[2] Microbiology, Virology and Infection Prevention & Control, Great Ormond Street Hospital for Children NHS Foundation Trust, London, United Kingdom; julianne.brown@gosh.nhs.uk, breuej@gosh.nhs.uk

[2a] Division of Infection and Immunity, University College London, London, United Kingdom; sofia.morfopoulou.10@ucl.ac.uk

[3] University Medical Center Hamburg-Eppendorf, UKE Institute for Medical Microbiology, Virology and Hygiene, Germany; nfischer@uke.de, j.huang@uke.de

[4] Viroscience, Erasmus Medical Center, Rotterdam, the Netherlands; b.oudemunnink@erasmusmc.nl, s.vanboheemen@erasmusmc.nl

[5] Dokuz Eylul University, Medical Faculty, Izmir, Turkey; arzu.sayiner@deu.edu.tr, alihanbulgurcu@gmail.com

[6] Hospital Henri Mondor, Paris, France; christophe.rodriguez@aphp.fr, guillaume.gricourt@aphp.fr

[7] Laboratory of Clinical and Epidemiological Virology (Rega Institute), KU Leuven, Belgium; els.keyaerts@kuleuven.be, leen.beller@kuleuven.be

[8] Institute of Virology, University of Zurich, Switzerland; claudia.bachofen@uzh.ch, jakub.kubacki@uzh.ch

[9] Laboratory of Virology, University Hospitals of Geneva, Geneva, Switzerland; Samuel.Cordey@hcuge.ch, Florian.Laubscher@hcuge.ch

[10] RIVM National Institute for Public Health and Environment, Bilthoven, the Netherlands; Dennis.Schmitz@RIVM.nl

[11] Friedrich-Loeffler-Institute, Institute of Diagnostic Virology, Greifswald, Germany; martin.beer@fli.de, dirk.hoeper@fli.de

[12] Institute of Medical Virology, University of Zurich, Switzerland; huber.michael@virology.uzh.ch, kufner.verena@virology.uzh.ch, zaheri.maryam@virology.uzh.ch

[13] Swiss Institute of Bioinformatics, Geneva, Switzerland; aitana.lebrand@sib.swiss

[14] Department of Microbiology, Medical School, Aristotle University of Thessaloniki, Greece; annap@auth.gr

[15] Virology Laboratory, Genomics and Health Area, Center for Public Health Research (FISABIO-Public Health), Generalitat Valenciana and Microbiology & Ecology Department, University of Valencia, Spain; F.Xavier.Lopez@uv.es

[16] CIBERESP, Instituto de Salud Carlos III, Spain; F.Xavier.Lopez@uv.es

Abstract

Metagenomic sequencing is increasingly being used in clinical settings for difficult to diagnose cases. The performance of viral metagenomic protocols relies to a large extent on the bioinformatic analysis. In this study, the European Society for Clinical Virology (ESCV) Network on NGS (ENNGS) initiated a benchmark of metagenomic pipelines currently used in clinical virological laboratories.

Methods

Metagenomic datasets from 13 clinical samples from patients with encephalitis or viral respiratory infections characterized by PCR were selected. The datasets were analysed with 13 different pipelines currently used in virological diagnostic laboratories of participating ENNGS members. The pipelines and classification tools were: Centrifuge, DAMIAN, DIAMOND, DNASTAR, FEVIR, Genome Detective, Jovian, MetaMIC, MetaMix, One Codex, RIEMS, VirMet, and Taxonomer. Performance, characteristics, clinical use, and user-friendliness of these pipelines were analysed.

Results

Overall, viral pathogens with high loads were detected by all the evaluated metagenomic pipelines. In contrast, lower abundance pathogens and mixed infections were only detected by 3/13 pipelines, namely DNASTAR, FEVIR, and MetaMix. Overall sensitivity ranged from 80% (10/13) to 100% (13/13 datasets). Overall positive predictive value ranged from 71-100%. The majority of the pipelines classified sequences based on nucleotide similarity (8/13), only a minority used amino acid similarity, and 6 of the 13 pipelines assembled sequences *de novo*. No clear differences in performance were detected that correlated with these classification approaches. Read counts of target viruses varied between the pipelines over a range of 2-3 log, indicating differences in limit of detection.

Conclusion

A wide variety of viral metagenomic pipelines is currently used in the participating clinical diagnostic laboratories. Detection of low abundant viral pathogens and mixed infections remains a challenge, implicating the need for standardization and validation of metagenomic analysis for clinical diagnostic use. Future studies should address the selective effects due to the choice of different reference viral databases.

Introduction

Viral metagenomic next-generation sequencing (mNGS) is increasingly being used in virology laboratories for the diagnosis of patients with suspected but unexplained infectious diseases. The current main clinical application of viral metagenomics is for diagnosing encephalitis of unknown cause [1, 2], but metagenomic sequencing is considered useful in a growing number of other clinical syndromes [3-6]. Although many wet-lab challenges need to be faced as well [14], the performance of metagenomic methods is largely dependent on accurate bioinformatic analysis, and both classification algorithms and databases are crucial factors determining the overall performance of the pipelines [7] [55]. A wide range of metagenomic pipelines and taxonomic classifiers have been developed, commonly for the purpose of biodiversity studies analysing the composition of the microbiome in different cohorts. In contrast, when applying metagenomics to patient diagnostics, potential false-negative and false-positive bioinformatic classification results can have significant consequences for patient care. Most reports on bioinformatic tools for metagenomic analysis for virus diagnostics typically describe algorithms and validations of single in-house developed pipelines developed by the authors themselves [8-12]. Most reports on bioinformatic tools for metagenomic analysis for virus diagnostics typically describe algorithms and validations of single in-house developed pipelines developed by the authors themselves [13], and recently a metagenomic benchmarking trial among Swiss virology laboratories has been conducted [7]. Recently, ESCV Network on NGS (ENNGS) recommendations for the introduction of next-generation sequencing in clinical virology, part II: bioinformatic analysis and reporting were published [55], aiming to address the challenges involved. While a professional External Quality Assessment (EQA) program is currently in preparation by Quality Control for Molecular Diagnostics (QCMD), the ENNGS [14] [55] conducted the presented benchmark of bioinformatic pipelines of the participating diagnostic laboratories using viral metagenomic datasets derived from clinical samples, in order to assist laboratories with selection and optimization of tools to be implemented for clinical use.

Methods

Datasets

To exclude differences in wet-lab procedures, the same raw, untrimmed metagenomic datasets were provided, so that the participants had standardized datasets for bioinformatic analysis.

In total, 13 clinical metagenomic datasets from samples well-characterized by RT-PCR [15-18] were selected from patients with encephalitis or respiratory complaints, including: cerebrospinal fluid (CSF, n=4), brain biopsies (n=3), nasopharyngeal swabs (n=3), nasal washings (n=1), bronchoalveolar lavage (n=1), and a plasma sample (n=1). RT-PCR panel results and Cq-values are included in the result section. The pathogens in the 13 datasets are depicted in Table 2.

For samples processed at the Great Ormond Street Hospital, London (GOSH), mRNA from the three brain biopsy samples was sequenced on an Illumina NextSeq500 instrument using an 81 bp paired-run after library preparation using Illumina's TruSeq Stranded mRNA LT sample preparation kit (p/n RS-122-2101) according to the manufacturer's instructions [19]. The other samples were spiked with Equine Arteritis Virus (EAV) and Phocid Herpes Virus (PhHV) internal controls preceding total nucleic acid extraction using the MagNAPure 96 DNA and Viral NA Small Volume Kit (Roche Diagnostics, Almere, the Netherlands) and sequenced on Illumina NextSeq500 (respiratory samples) or NovaSeq6000 (CSF samples, plasma) instruments using 150 bp paired-end runs after library preparation using New England BioLabs' NEBNext Ultra Directional RNA Library preparation kit for Illumina with in-house adaptations in order to enable simultaneous detection of both DNA and RNA viruses, at the Leiden University Medical Center (LUMC) [4, 20]. Three of the CSF samples were sequenced after enrichment using capture probes targeting vertebrate viruses [21]. Human reads from the output FASTQ files were removed after mapping them to human reference genome GRCh38 [22] with Bowtie2 version 2.3.4 [23] before the datasets were uploaded to various data sharing platforms (see below).

131

132 Data sharing

133 The FASTQ datasets were and remain publicly available for user-friendly downloading at
134 https://veb.lumc.nl/CliniMG (hosted by the dept. MM, LUMC, Leiden), and part of the datasets were
135 additionally accessible via a COMPARE Data Hub at http://www.ebi.ac.uk/ena/pathogens (hosted by
136 the European Bioinformatics Institute, EMBL-EBI) [24].

137

138 Bioinformatic pipelines

139 The datasets were analysed in a blinded fashion by the participants, with the (viral) metagenomic
140 pipelines and classification tools (**Figure 1 and Table 1**) used at their diagnostic laboratories:
141 Centrifuge [25], DAMIAN [26, 27], DIAMOND [28], DNASTAR [29], FEVIR [30], Genome Detective
142 [31], Jovian [32], MetaMIC [33], MetaMix [34, 35], One Codex [36], RIEMS [37, 38], Taxonomer [39],
143 and VirMet [40]. DAMIAN was run by two participants in combination with a different database
144 (pipeline A and B), and one participant run both Centrifuge and GenomeDetective. Details of the
145 algorithms are described in **Table 1**.

146

147 Performance characteristics

148 Both qualitative and quantitative performance of the pipelines were analysed with real-time PCR
149 results as gold standard. The following parameters available for all pipelines were considered:
150 pathogen detection, taxonomic classification level and target read count. Additionally, horizontal
151 genome coverage (if available), computational time, user-friendliness and output formats were
152 considered. Since EAV and PhHV were added as internal controls and not reported by the

153 participants (due to default reporting criteria, or absence in the database)they were not included in

154 the comparative analysis.

Results

Metagenomic pipeline characteristics

In total 13 different metagenomic pipelines and classification tools were in use in the 13 participating diagnostic laboratories. Clinical use, classification and output characteristics of the pipelines and tools utilized are shown in **Figure 1** and **Table 1**. The majority of the pipelines were developed or adapted at a local site, while four pipelines were commercially available and web-based: DNASTAR (Madison, WI, USA), Genome Detective (Emweb bv, Herent, Belgium), One Codex (San Francisco, USA), and Taxonomer (Utah, USA). DAMIAN and Centrifuge are publicly available as an open source software. Both classification tools and reference databases differed among participants (and were fixed for end-users of the commercially available pipelines); (adapted versions of) NCBI's nucleotide and RefSeq databases were most commonly used to generate reference databases. Six of the 13 pipelines assembled sequence reads *de novo*, whereas the others classified unassembled reads. The majority of the pipelines classified reads based on nucleotide similarity (8/13), and a minority used amino acid similarity (2/13), or a combination of both (3/13 pipelines). Parameters used by the participants for defining a positive result were the number of virus reads, horizontal genome coverage (some of the participants), and a cut-off based on posterior-probability scores of the species presence (MetaMix) and ROC-curves. Output formats varied, the majority had a user-friendly output format: excel, PDF or interactive webpage. Examples of these user-friendly output formats are shown in **Supplementary Figure S1**.

Detection of PCR targeted viral pathogens; sensitivity

The qualitative and quantitative results of the pipeline benchmarking for viruses detected by RT-PCR are shown in **Table 2 and Figure 2**. Overall, higher abundance viral pathogens (Cq-value < 28) were detected by all metagenomic pipelines evaluated. In contrast, viral pathogens with RT-PCR Cq-value

of 28 and higher including mixed virus infections were only detected by 3/13 pipelines, namely DNASTAR, FEVIR, and MetaMix. Although participants analysed the same FASTQ files, read counts of the target viruses varied from one to several orders of magnitude across pipelines. Also, read counts (all datasets combined) achieved by participants did not correlate well with the viral load as measured by RT-PCR (R=-0.07, P-value 0.5), however it must be noted that wet lab procedures varied per set of samples, including protocols with and without viral enrichment, which had potential impact on the viral read counts and thus on correlation with Cq-values. Overall sensitivity of the pipelines at sample level was 77% (10/13) - 100% (13/13 samples, mixed infections counted as one) (**Table 2 and Supplementary Table 2**). At viral mNGS hit level, overall sensitivity was 80% (12/15) - 100% (15/15 viral hits) (**Supplementary Table 4**). One of the participants reported normalized reads including the genome length, using the following formula: RPKM = (number of reads mapped to virus genome Y * 10^6) / (total number of reads * length of genome in kp). This formula was also used to normalize the reads of all study pipelines shown in Figure 2.

Taxonomic level of classification

The taxonomic levels of classification and typing of pathogenic viruses by the metagenomic pipelines with the settings used and reported by the participants are shown in **Figure 3** and **Supplementary Table 3**. The classification level is dependent on the database used, algorithm settings (classification of reads to the lowest common ancestor, LCA, in case of multiple hits), and the participant's default reporting levels based on either in-house validation data or clinical relevancy. Species level classification was the most common level reported. Serotype and strain level were identified by tools that were combined with NCBI's nt database without the LCA setting. DAMIAN was the only tool to report classification at the isolate level.

For the Adenovirus sample (#13), virus types reported were not consistent between different pipelines: human Adenovirus type 31 (DIAMOND, Jovian, DNASTAR, VirMet), type 12 (DAMIAN), type 31 or 61 (metaMIC), indicating that type classification was not always correct. Type 12 and 31 are both from subgroup A Adenoviruses, whereas type 61 is a type 31 recombinant virus.

Additional virus hits and positive predictive value

Additional viruses, either not tested for by RT-PCR or RT-PCR negative were reported by 11 out of 13 pipelines, and in one or more samples (**Supplementary Table 4**). The following additional viruses were reported by multiple pipelines and absent in the negative run control (dataset not available for the participants): human retrovirus RD114 (2-2102 reads, up to 28% genome coverage), feline leukemia virus (2-1406 reads), torque-teno virus (TTV) (18-66 reads, up to 7% genome coverage), polyomaviruses (5-41 reads, up to 37% genome coverage), Bovine viral diarrhea virus (BVDV) (6-220 reads, likely FBS contaminants), human metapneumovirus (HMPV) (15-21 reads, 9% genome coverage), human rhinovirus (HRV) (2-4 reads, up to 5% genome coverage), human parainfluenzavirus-4 (PIV-4) (2-6 reads) and Dengue virus (18-370 reads). RT-PCR data were available for some of the additional viruses detected (**Supplementary Table 4**). When considering viral mNGS hits with negative RT-PCR results: CoV-NL63 (1 read), PIV-4 (2-6 reads), HRV-C (2-4 reads), CoV-OC43 (5 reads), INF-B (2 reads), the positive predictive value ranged from 71-100% (**Figure 4**). It must be noted that for these mNGS hits, no distinction could be made between assignments of sequences genuinely present e.g. by index hopping (which was suspected given the low number of reads), false negative by PCR due to primers/probes mismatches, and false positive assignments. When considering the mNGS findings without available RT-PCR results, retrovirus RD114, leukemia viruses, TTV, and polyomaviruses sequences may actually be present given their association with the host (integrated or commensal).

Reporting criteria

Reporting criteria used by the participants are shown in **Table 1**: a threshold for number of reads, for genome coverage (number of nucleotides and proportion of the genome, or a certain number of genome regions covered), based on reference or in-house validation studies. A BLAST analysis of matching sequences was commonly used by the study participants to exclude false positive (or to confirm true positive) hits. Some participants indicated that for clinical samples outside of the current benchmark, they required a confirmatory PCR before reporting while others indicated that this was not needed based on experiences from their validation studies.

Discussion

This study aimed to benchmark the combination of bioinformatic tools and databases currently in use in diagnostic virology laboratories from the ESCV ENNGS network. The data presented here support bioinformatic selection and optimization of software for the implementation of viral metagenomic sequencing for pathogen detection in clinical samples. To our knowledge, this is the first large-scale international benchmarking study using datasets from clinical samples and pipelines currently applied in a large series of clinical diagnostic laboratories.

The study showed that the pipelines of all the participating laboratories succeeded in detecting viral pathogens with relative high viral loads (Cq-values <28), whereas lower abundant pathogens and mixed infections were only detected by some of the pipelines, namely DNASTAR, FEVIR, and MetaMix. These results are in line with other reports [7]. With regard to mixed infections, the less abundant viruses were generally missed, possibly due to the low number of reads, or reporting considerations. For the missed CoV-HKU1 virus, potential primer cross-reactivity with CoV-NL63 viruses was excluded by *in silico* analysis. The databases used in the pipelines were mostly custom-made, based on either NCBI's RefSeq [41] or nt database [42]. All of the participants used different classification tools, though no selection of laboratories using different tools was made in advance. Given the inclusion of different types of pipelines including commercially available ones with fixed databases, it was not feasible to compare the different tools with one standardised database at the local sites. Two of the three pipelines that reached 100% sensitivity included NCBI's nt database but this was also seen using a pipeline with NCBI's RefSeq database. Pipelines with NCBI's nt database scored both low and maximum precision. The design did allow for comparison of the complete pipeline in use for clinical diagnostics, from QC to reporting algorithms including posterior probability scores. No clear differences were observed in terms of performance based on nucleotide-based classification versus amino acid-based classification and *de novo* assembly-based algorithms versus read based classification: whereas amino-acid based classification may be more sensitive for

detecting variants, two of the three pipelines with 100% sensitivity used nucleotide-based classification (DNASTAR, FEVIR). High precision was reached by pipelines that used *de novo* assembly but this was not essential: 3/8 pipelines with 100% precision did not use *de novo* assembly (Centrifuge, Taxonomer, One Codex).

Reported read counts and genome coverage varied between pipelines up to several orders of magnitude (for read counts), explaining in part the differences observed in limits of detection for samples with very low viral load. Possibly, differences in reporting of unique versus non-uniquely mapped sequence reads may be related to this difference. Sensitivity and positive predictive value were measured, conveniently avoiding the proportion of true negative findings given the immense but unknown number of negative mNGS hits without RT-PCR data needed for specificity calculations. This aspect remains a limitation intrinsically linked to mNGS validations with clinical datasets, though datasets from negative matrix samples and/or negative controls would have been contributable for specificity calculations and correction for contaminants by the participants respectively. Positive predictive value calculations were hampered by the intrinsic inability to distinguish between sequences actually present in the dataset that might be undetected by RT-PCR because, for instance, primer mismatches, index hopping or contaminant sequences introduced during library preparation.This may partially be overcome by defining mNGS consensus results as alternative golden standard, however in diagnostic settings e.g. index hopping reads should not be labelled positive despite being actually present in the dataset. A study design using synthetic datasets this may enable a more accurate estimation of the specificity and PPV *in silico* however these estimates would deviate significantly from the ones in real-life conditions, where has to be dealt with interfering factors such as the 'kitome', present in every single dataset. The current comparison aimed at the entire bioinformatic workflow including thresholds for reporting and corrections for interfering real-life factors.

It is important to note that participants likely have optimized their interpretation algorithm including cut-offs for their specific workflow from library preparation to sequencing. A different wet lab procedure (sequencer with or without index hopping, preparation with or without probe enrichment) will require new validation and indexing of the determined cut-off values and probability values. Because this was a dry lab comparison exercise, the participants could not follow their routine wet lab workflow and confirmatory PCR steps, which may have affected the reporting of results. Therefore no conclusions can be drawn on the limit of detection of the full metagenomic workflows used in each specific laboratory.

Genome coverage and depth was not always taken into account by the participating laboratories, however can be an effective parameter to distinguish between (PCR-)contaminants, often indicated by high depth at a small (PCR amplicon) region of the genome, and true positives [21, 55]. In five of participating laboratories a cut-off of one single read was chosen for defining a positive mNGS result. While potentially at higher risk of reporting false positive results, the PPV of these pipelines ranged from 72 up to 100%, indicating that this cut-off was dependent on the overall steps of the analysis and reporting. ROC analysis was used to find the optimal balance between sensitivity and specificity [20].

Finally, our taxonomic results are in line with data available from other groups [43]: the pipelines performed well at species level but deeper level classification was subject to less reliable classification in some cases.

In conclusion, a wide variety of viral metagenomic pipelines with overall high sensitivity are currently used in the ESCV ENNGS participating clinical diagnostic laboratories. Detection of low abundance viral pathogens or mixed infections remains a challenge, implicating the need for standardization and validation of metagenomic analysis for clinical diagnostic use [44]. The algorithm for defining positive results and rejecting false positive results is critical and should be evaluated individually for every workflow, which includes genome extraction, library preparation, sequencer and bioinformatic

pipeline. Identification of deeper taxonomic levels is challenging, dependent on the individual types present in the reference database, and should be validated separately to prevent misidentification.

314 Acknowledgements

315 We thank the COMPARE study group (https://www.compare-europe.eu/) and the EMBL-EBI

316 (https://www.ebi.ac.uk/) for the availability of the Data Hubs.

317

318 Funding

319 MH was supported by the Clinical Research Priority Program 'Comprehensive Genomic Pathogen

320 Detection' of the University of Zurich

321 FXL receives funding from Instituto de Salud Carlos III, Spain (Grant numbers PI18/01824 and

322 PI18/01759 and CIBEResp).

323

324 Conflict of interest: none

Pipeline no (alphabetical order)	1	2	3 and 3A	4	5	6	7	8	9	10	11	12	13
Classification tool or pipeline	Centrifuge [25] v1.0.1-beta	DAMIAN [26, 27] v190628	DIAMOND [28] v0.9.13.114	DNASTAR [29] Lasergene v16	FEVIR [30] V1	Genome Detective [31] v1.110	Jovian [32] V0.9.6	MetaMIC [33] v2.1.1	MetaMix [34] [35] v1.2	One Codex [36] v1	RIEMS [37, 38] v4.0	Taxonomer [39] 2020-U	VirMet [40] v1.1.1
Clinical usage by participant	Patient care	Experimental	Experimental	Patient care	Patient care	Patient care[a]	Patient care	Patient care[a]	Patient care[a]	Experimental	Patient care	Experimental	Patient care
In-house/ commercially available	Open-source software	Open-source software	In-house	Commercial	In-house	Commercial	In-house	In-house	In-house	Commercial	In-house	Commercial	In-house
Local/web-based	Local	Local	Local	Local (cloud optional)	Local	Web-based	Local	Local	Web-based (hosted on Bluebee)	Web-based	Local	Web-based	Local
De novo assembly	N	Y	Y [45]	N (optional)	N	Y	Y	Y	Y	N	Y	N	N
Alignment of NT/ AA	NT	NT, AA	AA	NT	NT	NT, AA	NT	NT	AA	NT	NT	NT, AA	NT
Database used by participant viral/bacterial (version)	Viruses, bacteria, fungi, archae; RefSeq (compressed index) V2019-04-04	NCBI's nt and nr v2019-05-17, PFAM 30.0	Viruses; NCBI's non-redundant protein database/ pipeline 3B: NCBI's nt V0.9.22	Based on NCBI's nt V2020-01-08	Viruses; based on Virosaurus [46] v90v_2018_11	Viruses; based on RefSeq (filtering: Swissprot Uniref 90) v2018-11-14	NCBI's nt, v2019-11-30 a.o. (compressed index)	Bacteria, viruses, fungi; v2.1.1 based on NCBI's nt (complete)	Human, Environmental, bacteria, Viruses RefSeq protein v2017	Bacteria, viruses, fungi, archaea, protozoa, One Codex DB v2019-5-1	NCBI's nt (complete) v2019-3-16	Based on NCBI's nt (v May 2019)	Based on NCBI's nt (selection of viral full genomes without compression) v224
Paired reads as input option	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N
Trimming and QC tools	Trimmomatic	Trimmomatic	Trimmomatic [47]	Included in DNASTAR	Included in virusscan [48] 1.0		Trimmomatic, fastqc, multiqc	HMN Trimmer	TrimGalore!, prinseq	Cutadapt	Roche 454 newbler	N	Seqtk, prinseq
Exclusion of human reads	N	Y	N/pipeline 3B:Y[b]	Y	N	Y	Y	Y	Y	Y	N	N	Y
Output	Script,	Excel	Script,	User-	Script	Web	Web	PDF	Web	Web	PDF	Web	PDF

type	Krona		krona	friendly		interface, interactive	interface, interactive		interface, interactive, PDF and excel	interface, interactive, PDF and excel		interface, interactive	
Visualization of genome coverage	N	N	N	Y	N	Y	N	N	N (CLC Genomics Workbench)	N	Y	N (free version)/ Y (paid)	Y
Computational time for analysis per sample (CPU/RAM)	~10 min (24 CPUs, 0.3 GB RAM)	90 min (56 CPUs, 125 GB RAM)	90min (36 CPUs, RAM=23 gb)	3 min (4 CPUs, 64 GB RAM	1 min (176 CPUs, 384 GB RAM)	~10 min (web-based)	12 min (20 CPUs, U GB RAM)	18 min (U CPUs, U GB RAM)	60-180 min (web-based)	35-40 min (web-based)	U (48 CPUs, 768 GB RAM)	10 min (web-based)	20 min (16 CPUs, 64 GB RAM)
Cut-off for defining positive result used	≥15 reads [20]	Contig length => 400 bp	1 read	1 read	≥300 nt coverage	≥3 regions, distributed [49]	U	Above background : environmental sample	≥3 regions >10 reads [50] [51] Probability score	1 read	1 read	1 read	≥3 reads, distributed, >100x than NC/other samples
Confirmatory analysis required for clinical reporting	BLAST, PCR	PCR for clinical cases	BLAST, PCR for clinical cases	BLAST, PCR for clinical cases	BLAST	U	U	PCR not required (based on validation) [52-54]	BLAST, coverage (PCR not required based on validation)	Not required	BLAST, PCR	BLAST	U

Table 1. Clinical use, classification and output characteristics of metagenomic pipelines analysed.

[a] Within the scope of accreditation

[b] Mapping of the trimmed reads to HG38 by Bowtie with "very-sensitive" option

AA; amino acid, NT; nucleotide, U; Undisclosed

References

1. Brown, J.R., T. Bharucha, and J. Breuer, *Encephalitis diagnosis using metagenomics: application of next generation sequencing for undiagnosed cases.* J Infect, 2018. **76**(3): p. 225-240.
2. Wilson, M.R., et al., *Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis.* N Engl J Med, 2019. **380**(24): p. 2327-2340.
3. Jerome, H., et al., *Metagenomic next-generation sequencing aids the diagnosis of viral infections in febrile returning travellers.* J Infect, 2019. **79**(4): p. 383-388.
4. van Boheemen, S., et al., *Retrospective Validation of a Metagenomic Sequencing Protocol for Combined Detection of RNA and DNA Viruses Using Respiratory Samples from Pediatric Patients.* J Mol Diagn, 2020. **22**(2): p. 196-207.
5. Lewandowska, D.W., et al., *Metagenomic sequencing complements routine diagnostics in identifying viral pathogens in lung transplant recipients with unknown etiology of respiratory infection.* PLoS One, 2017. **12**(5): p. e0177340.
6. Kufner, V., et al., *Two Years of Viral Metagenomics in a Tertiary Diagnostics Unit: Evaluation of the First 105 Cases.* Genes (Basel), 2019. **10**(9).
7. Junier, T., et al., *Viral Metagenomics in the Clinical Realm: Lessons Learned from a Swiss-Wide Ring Trial.* Genes (Basel), 2019. **10**(9).
8. Chen, J., J. Huang, and Y. Sun, *TAR-VIR: a pipeline for TARgeted VIRal strain reconstruction from metagenomic data.* BMC Bioinformatics, 2019. **20**(1): p. 305.
9. Miller, S., et al., *Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid.* Genome Res, 2019. **29**(5): p. 831-842.
10. Paez-Espino, D., et al., *Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data.* Nat Protoc, 2017. **12**(8): p. 1673-1682.
11. Li, Y., et al., *VIP: an integrated pipeline for metagenomics of virus identification and discovery.* Sci Rep, 2016. **6**: p. 23774.
12. Nooij, S., et al., *Overview of Virus Metagenomic Classification Methods and Their Biological Applications.* Front Microbiol, 2018. **9**: p. 749.
13. Brinkmann, A., et al., *Proficiency Testing of Virus Diagnostics Based on Bioinformatics Analysis of Simulated In Silico High-Throughput Sequencing Data Sets.* J Clin Microbiol, 2019. **57**(8).
14. Lopez-Labrador F.X., B.J.R., Fischer N., Harvala H., Van Boheemen S., Cinek O, Sayiner A, Vasehus Madsen T, Auvinen E. et al., *Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: wet lab procedure.* J Clin Virol, 2020; Dec.
15. Kalpoe, J.S., et al., *Validation of clinical application of cytomegalovirus plasma DNA load measurement and definition of treatment criteria by analysis of correlation to antigen detection.* J Clin Microbiol, 2004. **42**(4): p. 1498-504.
16. Read, S.J. and J.B. Kurtz, *Laboratory diagnosis of common viral infections of the central nervous system by using a single multiplex PCR screening assay.* J Clin Microbiol, 1999. **37**(5): p. 1352-5.
17. Lankester, A.C., et al., *Epstein-Barr virus (EBV)-DNA quantification in pediatric allogenic stem cell recipients: prediction of EBV-associated lymphoproliferative disease.* Blood, 2002. **99**(7): p. 2630-1.

18. Loens, K., et al., *Performance of different mono- and multiplex nucleic acid amplification tests on a multipathogen external quality assessment panel.* J Clin Microbiol, 2012. **50**(3): p. 977-87.
19. Morfopoulou, S., et al., *Deep sequencing reveals persistence of cell-associated mumps vaccine virus in chronic encephalitis.* Acta Neuropathol, 2017. **133**(1): p. 139-147.
20. van Rijn, A.L., et al., *The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease.* PLoS One, 2019. **14**(10): p. e0223952.
21. Carbo, E.C., et al., *Improved diagnosis of viral encephalitis in adult and pediatric hematological patients using viral metagenomics.* J Clin Virol, 2020. **130**: p. 104566.
22. https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/ (Accessed July.
23. B. Langmead, S.L.S., Fast gapped-read alignment with Bowtie 2, Nat. Methods and h.d.o.n. 9 (4) (2012) 357–359, Apr.
24. Amid, C., et al., *The COMPARE Data Hubs.* Database-the Journal of Biological Databases and Curation, 2019: p.1-14.
25. Kim, D., et al., *Centrifuge: rapid and sensitive classification of metagenomic sequences.* Genome Res, 2016. **26**(12): p. 1721-1729.
26. Alawi, M., et al., *DAMIAN: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples.* Sci Rep, 2019. **9**(1): p. 16841.
27. https://sourceforge.net/projects/damian-pd.
28. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using DIAMOND.* Nat Methods, 2015. **12**(1): p. 59-60.
29. https://www.dnastar.com/software/lasergene/.
30. Fernandes, J.F., et al., *Unbiased metagenomic next-generation sequencing of blood from hospitalized febrile children in Gabon.* Emerg Microbes Infect, 2020. **9**(1): p. 1242-1244.
31. Vilsker, M., et al., *Genome Detective: an automated system for virus identification from high-throughput sequencing data.* Bioinformatics, 2019. **35**(5): p. 871-873 www.genomedetective.com.
32. https://github.com/DennisSchmitz/Jovian.
33. Rodriguez, C., et al., *Pathogen identification by shotgun metagenomics of patients with necrotizing soft-tissue infections.* Br J Dermatol, 2019.
34. Morfopoulou, S. and V. Plagnol, *Bayesian mixture analysis for metagenomic community profiling.* Bioinformatics, 2015. **31**(18): p. 2930-8.
35. https://cran.r-project.org/web/packages/metaMix/index.html.
36. Minot, S.S., *One Codex: a sensitive and accurate data platform for genomic microbial identification.* bioRxiv, 2015.
37. Scheuch, M., D. Hoper, and M. Beer, *RIEMS: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets.* BMC Bioinformatics, 2015. **16**: p. 69.
38. https://github.com/EBI-COMMUNITY/fli-RIEMS.
39. Flygare, S., et al., *Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling.* Genome Biol, 2016. **17**(1): p. 111.
40. https://github.com/medvir/VirMet and https://github.com/medvir/shiny-server/tree/master/NGS/VirMetRunAnalysis.

41. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D733-45.
42. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2011. **39**(Database issue): p. D32-7.
43. Sczyrba, A., et al., *Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software.* Nat Methods, 2017. **14**(11): p. 1063-1071.
44. Bharucha, T., et al., *STROBE-metagenomics: a STROBE extension statement to guide the reporting of metagenomics studies.* Lancet Infect Dis, 2020. **20**(10): p. e251-e260.
45. Nurk, S., et al., *metaSPAdes: a new versatile metagenomic assembler.* Genome Res, 2017. **27**(5): p. 824-834.
46. https://viralzone.expasy.org/8676.
47. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.
48. https://github.com/sib-swiss/virusscan.
49. Carbo EC, B.E., Karelioti E, Sidorov I, Feltkamp MCW, Von dem Borne PA, Verschuuren Jan JGM, Kroes ACM, Claas ECJ, De Vries JJC, *Improved diagnosis of viral encephalitis in adults and pediatric hematological patients using viral metagenomics.* bioRxiv, 2020.
50. Mongkolrattanothai, K. and J. Dien Bard, *The utility of direct specimen detection by Sanger sequencing in hospitalized pediatric patients.* Diagn Microbiol Infect Dis, 2017. **87**(2): p. 100-102.
51. Kawada, J., et al., *Identification of Viruses in Cases of Pediatric Acute Encephalitis and Encephalopathy Using Next-Generation Sequencing.* Sci Rep, 2016. **6**: p. 33452.
52. Rodriguez, C., et al., *Pathogen identification by shotgun metagenomics of patients with necrotizing soft-tissue infections.* Br J Dermatol, 2020. **183**(1): p. 105-113.
53. Rodriguez, C., et al., *Fatal Measles Inclusion-Body Encephalitis in Adult with Untreated AIDS, France.* Emerg Infect Dis, 2020. **26**(9): p. 2231-2234.
54. Rodriguez, C., et al., *Fatal Encephalitis Caused by Cristoli Virus, an Emerging Orthobunyavirus, France.* Emerg Infect Dis, 2020. **26**(6): p. 1287-1290.
55. De Vries JJC, et al. Recommendations for the introduction of nex**t-generation sequencing in clinical virology, part II: bioinformatic analysis and reporting. J Clin Virol 2021, 104812** https://doi.org/10.1016/j.jcv.2021.104812

Figure 1. Workflow of bioinformatic analysis of (viral) metagenomics data with the pipelines and classification tools used by participants in the current study.

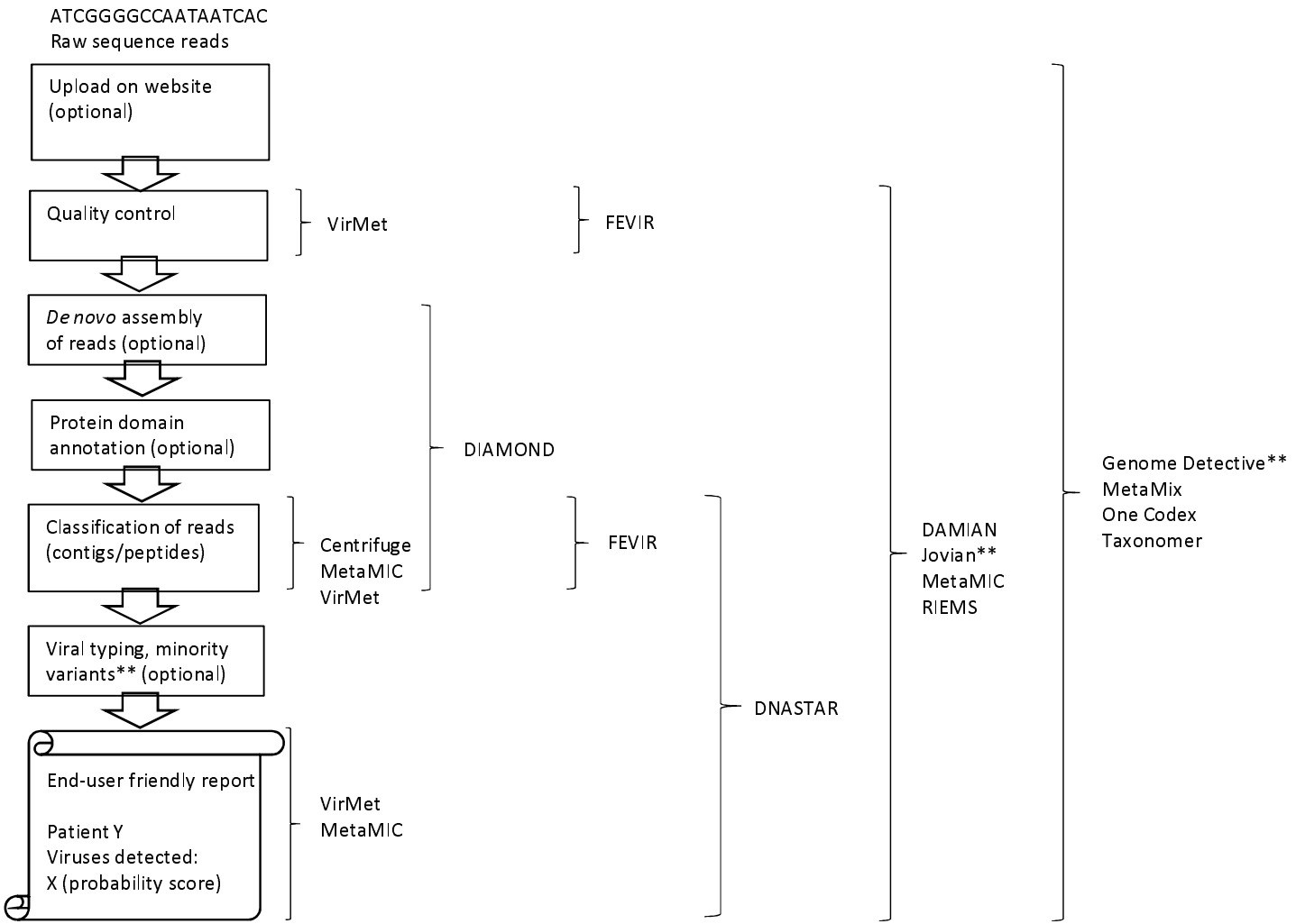


Table 2. Qualitative and quantitative results: raw sequence read count categories of the PCR positive viruses reported by the metagenomic pipelines using datasets from 13 clinical samples, per classification tool (complete pipeline details can be found in table 1).
CSF; cerebrospinal fluid, NP; nasopharyngeal, BAL; bronchoalveolar lavage, and in legend: ND; not detected

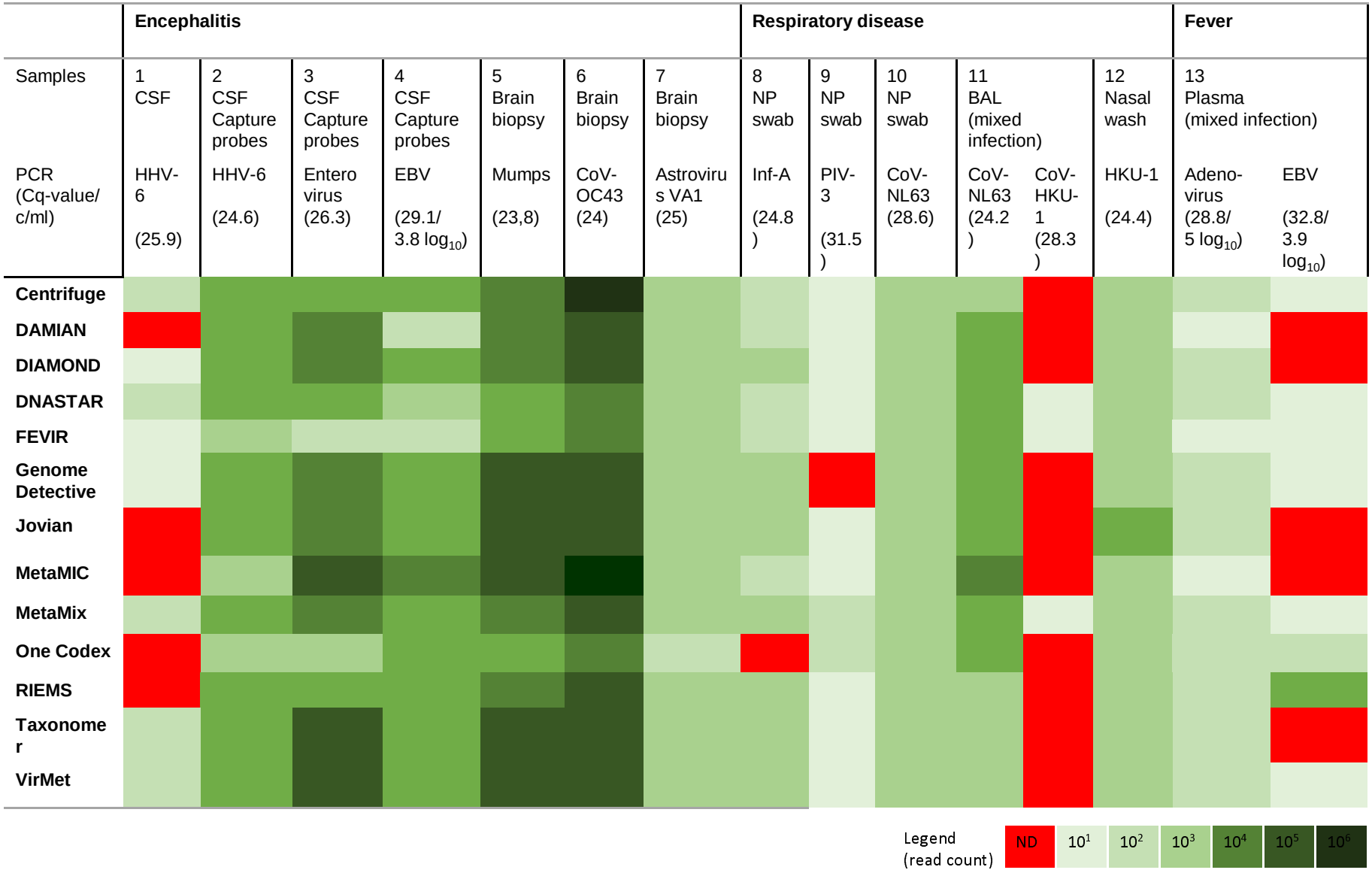


Figure 2. Sequence read counts (Y) versus RT-PCR Cq values (X) of the PCR positive viruses using datasets from 13 clinical samples, per classification tool with read counts reported by the participants (complete pipeline details can be found in table 1).
Each vertical series of dots represents one clinical sample.
The different wet lab methods used are marked (^ mRNA sequencing, $ RNA/DNA sequencing, and #: a captured approach using probes targeting vertebrate viruses).
RPKM; reads per kilobase of genome per million mapped reads

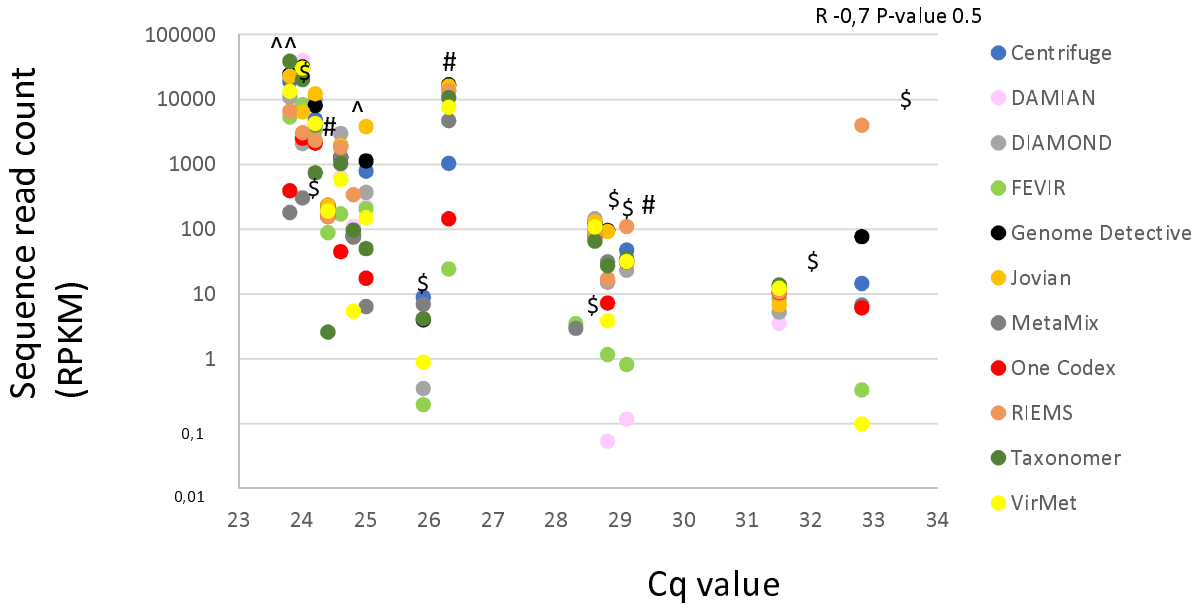
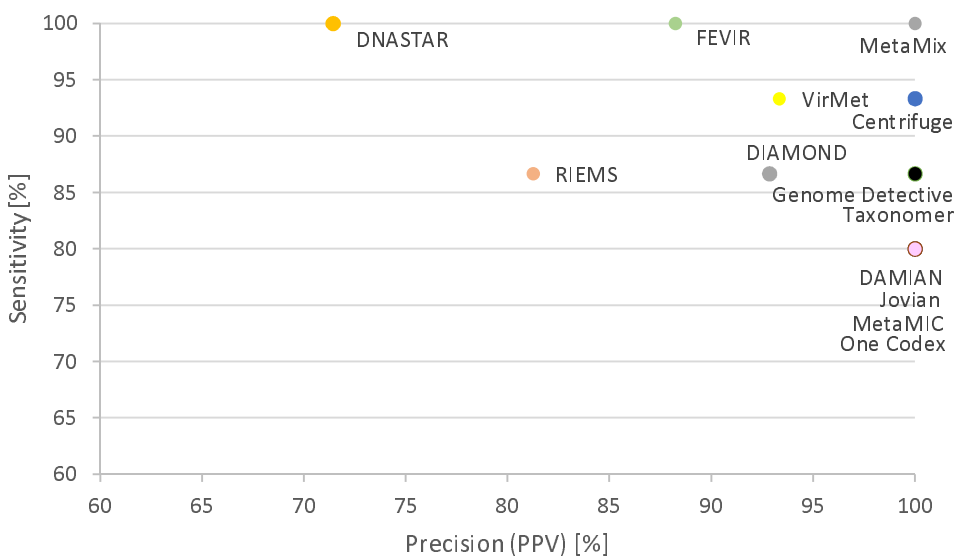


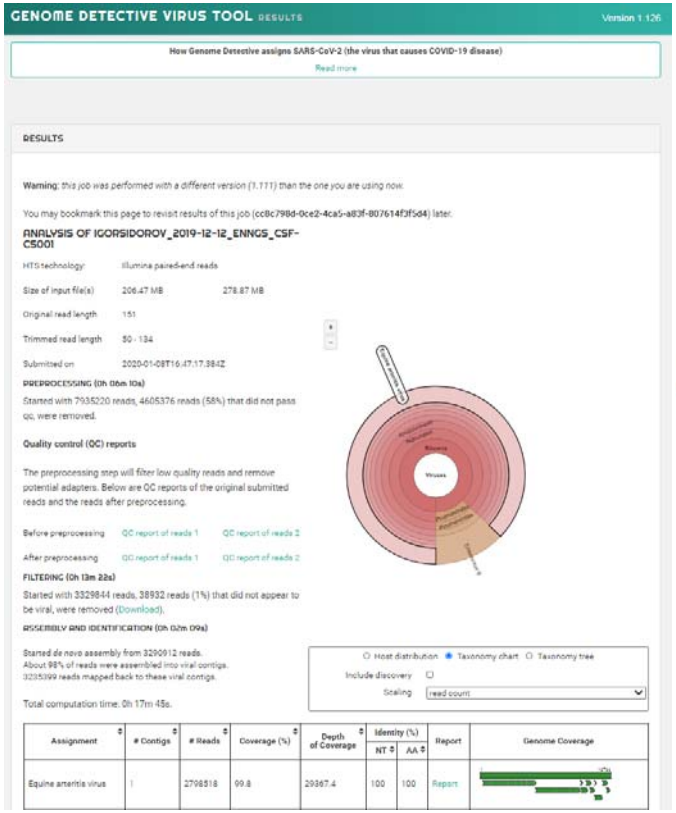
Figure 3. (Taxonomic) level of classification and typing of the pathogenic viruses identified using the combination of tools and databases, as reported by participating diagnostic laboratories. Depicted are the number of target viruses per classification level.
RefSeq; NCBI's RefSeq data base (or an adapted version), NT; NCBI's nucleotide database (or an adapted version); LCA; lowest common ancestor.
*Taxonomic assignment method described in [25,26]

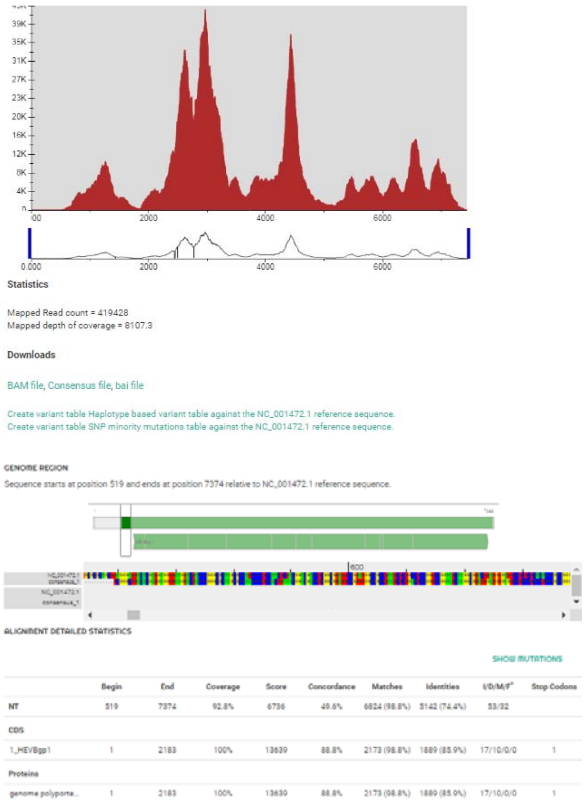
Figure 4. Sensitivity and positive predictive value based on the hits reported by the participants, for the different pipelines/classification tools.



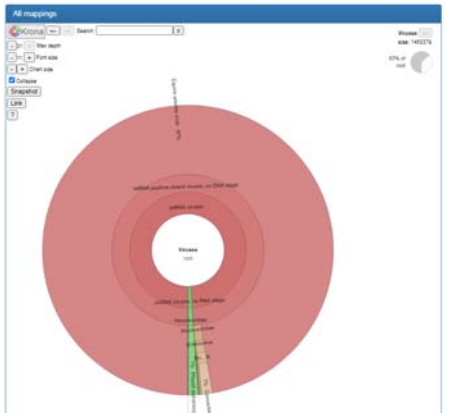
Supplementary Figure S1. User-friendly output formats of metagenomic pipelines and tools tested, command line formats excluded.

Genome Detective





Cenrifuge/Krona plots



A

Sample Name	cs068_contiglen400
Read Pairs	11723458
Read Pairs (PF)	11702298
Single Reads	0
Single Reads (PF)	21160
Average Fragment Size	262
Fragment Size Standard Deviation	83
Reads aligning to host (est. perc.)	0.5

Species	Absolute Abundance	Relative Abundance	Assembly Length	Contigs	Orfs	Maximum Identity (nucl)	Maximum Identity (prot)
Influenza C virus	5845	1.612%	12910	7	27	99.483	-1
Alphaarterivirus equid	640	0.177%	12545	5	42	100	-1
Moraxella catarrhalis	76358	21.065%	1838485	1716	5070	100	100
Rothia mucilaginosa	43973	12.131%	14728	23	72	100	98.352
Streptococcus pneumoniae	38291	10.563%	783428	894	2370	100	100
uncultured bacterium	36029	9.939%	60681	88	175	100	88.462
Moraxella bovis	28466	7.853%	571467	621	1681	99.281	99.167
Streptococcus parasanguinis	18833	5.195%	401551	509	1313	100	100
Atopobium parvulum	12025	3.317%	18548	28	84	99.695	100
Cardiobacterium hominis	4529	1.249%	3563	3	16	98.667	78.392
Streptococcus salivarius	2064	0.569%	14449	14	39	100	42.857
Acinetobacter baumannii	1512	0.417%	6643	7	46	-1	88.06
Streptococcus mitis	558	0.154%	22252	36	62	99.632	97.059
Neisseria meningitidis	118	0.033%	4998	6	25	98.887	79.882
Klebsiella pneumoniae	56	0.015%	2910	3	25	-1	52.83
Veillonella parvula	4	0.001%	462	1	2	97.186	-1
Moraxella phage Mcat17	49	0.014%	1356	1	4	98.304	-1
Streptococcus anginosus	9672	2.668%	26025	41	96	100	52.727
Lactobacillus gasseri	9596	2.647%	5091	6	11	100	-1
Moraxella nonliquefaciens	7413	2.045%	135928	195	390	100	100
Prevotella denticola	4615	1.273%	2572	5	11	99.368	-1
Streptococcus mutans	4589	1.266%	2559	4	6	100	-1
Fusobacterium nucleatum	4207	1.161%	1941	4	6	100	-1

B

Contig_ID	Length	Abundance	Abundance/Length	Assignment
315326	2340	999	0.426923077	Influenza C virus
315411	2039	981	0.481118195	Influenza C virus (C/Catalonia/1266/2009)
315312	2412	914	0.37893864	Influenza C virus (C/Miyagi/25/2004)
316527	938	872	0.929637527	Influenza C virus (C/India/P119564/2011)
315525	1763	838	0.475326149	Influenza C virus
315342	2259	778	0.344400177	Influenza C virus (C/Miyagi/25/2004)

Domains

Contig_ID	Orf_ID	Accession	Name	Description	Viral Root
315326	1329636	PF00604	Flu_PB2	Influenza RNA-dependent RNA polymerase subunit PB2	Orthomyxoviridae
315411	1330160	PF00509	Hemagglutinin	Haemagglutinin	Orthomyxoviridae
315411	1330160	PF02710	Hema_HEFG	Haemagglutinin domain of haemagglutinin-esterase-fusion glycoprotein	ssRNA viruses
315411	1330160	PF08720	Hema_stalk	Influenza C hemagglutinin stalk	Influenza C virus (C/Ann Arbor/1/50)
315411	1330160	PF03996	Hema_esterase	Haemagglutinin esterase	ssRNA viruses
315312	1329545	PF00602	Flu_PB1	Influenza RNA-dependent RNA polymerase subunit PB1	Orthomyxoviridae
315312	1329545	PF02404	SCF	Stem cell factor	
315312	1329542	PF10523	BEN	BEN domain	dsDNA viruses, no RNA stage
316527	1334337	PF03506	Flu_C_NS1	Influenza C non-structural protein (NS1)	Influenza C virus (C/Ann Arbor/1/50)
316527	1334337	PF03555	Flu_C_NS2	Influenza C non-structural protein (NS2)	Influenza C virus (C/Ann Arbor/1/50)
316527	1334337	PF09014	Sushi_2	Beta-2-glycoprotein-1 fifth domain	
316527	1334335	PF09172	DUF1943	Domain of unknown function (DUF1943)	
315525	1330771	PF00506	Flu_NP	Influenza virus nucleoprotein	Orthomyxoviridae
315525	1330771	PF10211	Ax_dynein_light	Axonemal dynein light chain	
315342	1329737	PF05010	TACC	Transforming acidic coiled-coil-containing protein (TACC)	
315342	1329737	PF00603	Flu_PA	Influenza RNA-dependent RNA polymerase subunit PA	Orthomyxoviridae
315342	1329739	PF09416	UPF1_Zn_bind	RNA helicase (UPF2 interacting domain)	
316101	1333102	PF03021	CM2	Influenza C virus M2 protein	Influenza C virus (C/Ann Arbor/1/50)
316101	1333102	PF03026	CM1	Influenza C virus M1 protein	Influenza C virus (C/Ann Arbor/1/50)

Metamix/Bluebee

RNA-Seq Encephalitis Diagnostics

Pipeline Run Details

User Reference:	GOSHmeta3	Pipeline:	GOSH RNA-Seq Encephalitis Diagnostics 1.2.0
Request Date:	Sep, 10 2019 10:58:33	Start Date:	Sep, 10 2019 11:00:23
Duration:	14h 58m 33s	Requestor:	Dr. Julianne Brown
User Tags:			

Input Data

UCLGNS1212-13M1974-B_S7_R1_001.fastq.gz

File Name:	UCLGNS1212-13M1974-B_S7_R1_001.fastq.gz	File Path:	UCLGNS1212-13M1974-B_S7_R1_001.fastq.gz
Size:	5.57 GB	Format:	FASTQ
Creation Date:	Sep, 10 2019 08:55:02	User Tags:	
Run In Tags:	GOSHmeta3	Connector Tags:	Upload

UCLGNS1212-13M1974-B_S7_R2_001.fastq.gz

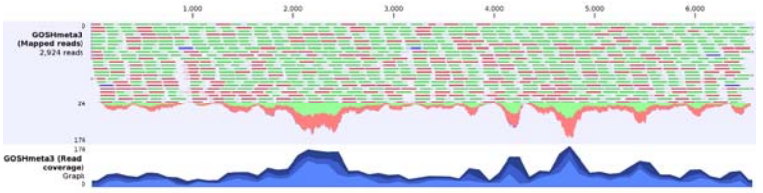
File Name:	UCLGNS1212-13M1974-B_S7_R2_001.fastq.gz	File Path:	UCLGNS1212-13M1974-B_S7_R2_001.fastq.gz
Size:	5.7 GB	Format:	FASTQ
Creation Date:	Sep, 10 2019 08:47:18	User Tags:	
Run In Tags:	GOSHmeta3	Connector Tags:	Upload

Results

	"taxonID"	"scientName"	"finalAssignments"	"poster.prob"	"log10BF"
"8"	"unknown"	"unknown"	30490	1	NA
"7"	"9606"	"Homo sapiens"	28586	1	28977.6477200774
"6"	"645687"	"Astrovirus VA1"	2423	1	9562.99329606601
"1"	"10090"	"Mus musculus"	536	1	684.019570605247
"2"	"28090"	"Acinetobacter lwoffii"	25	1	135.6328430578
"4"	"469"	"Acinetobacter"	19	0.99	57.62766626128
"3"	"43675"	"Rothia mucilaginosa"	14	1	109.876588922052
"5"	"488"	"Neisseria mucosa"	11	0.94	14.9840642137569

List of detected species (presentSpecies_assignedReads.tsv)

Command line (CLC Genomics Workbench) PDF coverage plot

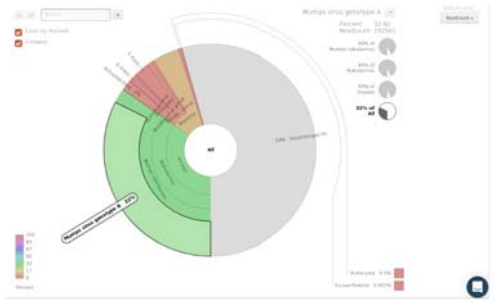


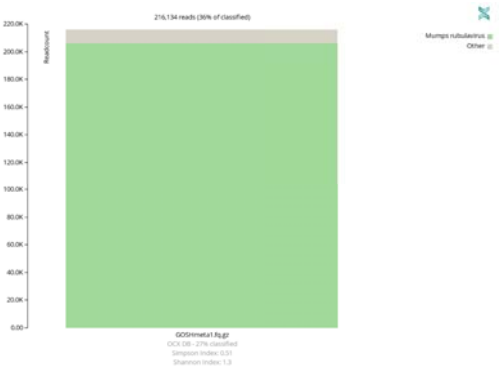
RIEMS

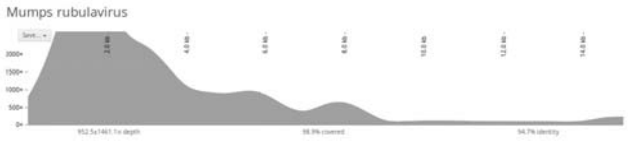
Family	*Species*	SkTax	FunTax	Tax	counts	Pre Screening	Mapping	Mapping2	Assembly	Ident Assembly	Megablast vs ntdb	Ident Megablast vs ntdb	Blastn vs Organism	Ident Blastn vs Organism	Blastn vs ntdb	Ident Blastn vs ntdb
Pithoviridae	*Pithovirus LCPAC102*	10239	2023203	2506587	4	0	0	0	0	0	0	0	0	0	4	76.92-88
	Pithovirus LCPAC104	10239	2023203	2506589	2	0	0	0	0	0	0	0	0	0	2	76.54-78.2
Peribunyaviridae	*Impatiens necrotic spot tospovirus*	10239	1980416	1933294	1	0	0	0	0	0	0	0	0	0	1	75.26
	Tomato spotted wilt tospovirus	10239	1980416	1933298	1	0	0	0	0	0	0	0	0	0	1	72.36
Nudiviridae	*Heliothis zea nudivirus*	10239	1511852	29250	1	0	0	0	0	0	0	0	0	0	1	81.82
Mimiviridae	*Hokovirus HKV1*	10239	549779	1977638	5	0	0	0	0	0	0	0	0	0	5	71.74-85.19
	Indivirus ILV1	10239	549779	1977633	1	0	0	0	0	0	0	0	0	0	1	86.96
	Bodo saltans virus	10239	549779	2024608	1	0	0	0	0	0	0	0	0	0	1	88.89
	Tupanvirus soda lake	10239	549779	2126985	1	0	0	0	0	0	0	0	0	0	1	84.48
	Mimivirus sp SH	10239	549779	2496520	1	0	0	0	0	0	0	0	0	0	1	80
	Mimivirus LCMiAC01	10239	549779	2506608	1	0	0	0	0	0	0	0	0	0	1	77.53
	Mimivirus LCMiAC02	10239	549779	2506609	1	0	0	0	0	0	0	0	0	0	1	87.23
Papillomaviridae	*Alphapapillomavirus 9*	10239	151340	337041	1	0	0	0	0	0	0	0	0	0	1	90.7
Astroviridae	*Astrovirus VA1HMO C*	10239	39733	1602211	1421	0	1125	111	129	95.19-100	31	88.14-100	10	82.35-94.84	15	86.28-100
	Human astrovirus	10239	39733	1868658	8	0	0	0	2	100	4	96.06-98.65	2	78.33-89.47	0	0
	Mamastrovirus 9	10239	39733	1239573	6	0	0	0	0	0	6	94.02-100	0	0	0	0

OneCodex

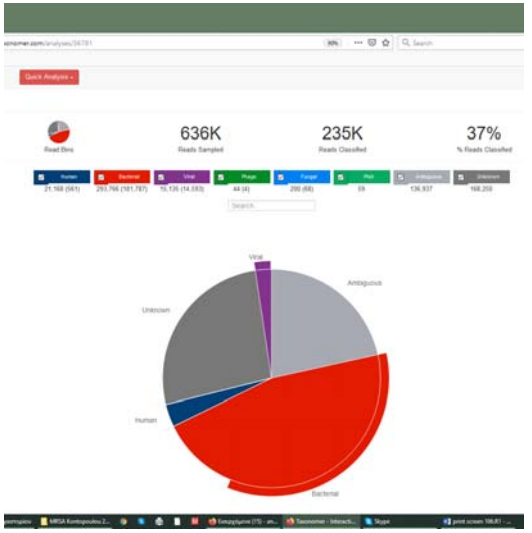
Organism Name	Rank	Tax ID	% of All Reads	# of Reads	# of Reads (w/ Children)	Est. Depth	Est. Abundance
Mumps rubulavirus	species	1979165	0,606542997	13511	206072	1665,2489	0,999342612
Cutibacterium acnes	species	1747	0,002873122	64	73	0,2731787	0,000163939
Pseudomonas fluorescens	species	294			24	0,0082842	4,97146E-06
Pseudomonas putida	species	303			7	0,000401	2,4064E-07
Ralstonia pickettii	species	329	0,00017957	4	19	0,0158012	9,48256E-06
Xanthomonas campestris	species	339	0,001616131	36	36	0,0009556	5,73464E-07
Methyloru brum extorquens	species	408			6	0,0007724	4,63546E-07
Moraxella nonliquefaciens	species	478	0,002424197	54	54	0,0029298	1,75824E-06
Moraxella sp.	species	479	0,000808066	18	18		





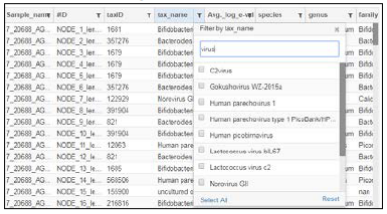


Taxonomer



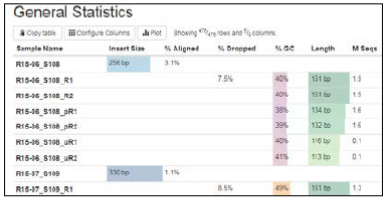
Jovian

A. Filterable spreadsheets

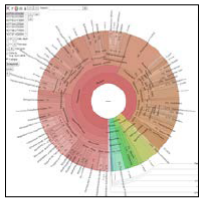


B. Heatmaps

C. MultiQC report



D. Krona plot



E. IGVjs scaffold alignment viewer for in-depth analyses

VirMet Report

NA

Run Information

Run name	191010_M01274_0212_000000000-ENNR1
Sample name	ENNGS_7059_CSF_RNA_S01CAP
Quality-filtered reads (DNA/RNA)	NA/3.556 M
Report date	2020-02-27
Analyzed by	Maryam Zaheri

Domain Level Taxonomy Profile

domain	reads	percent	abundance
human	0	0.000	+
bacterial	4298	0.121	+
fungal	0	0.000	+
bovine	4744	0.133	+
viral	2017278	56.722	++++++++++++++++++++
undetermined	1530103	43.024	+++++++++++++++

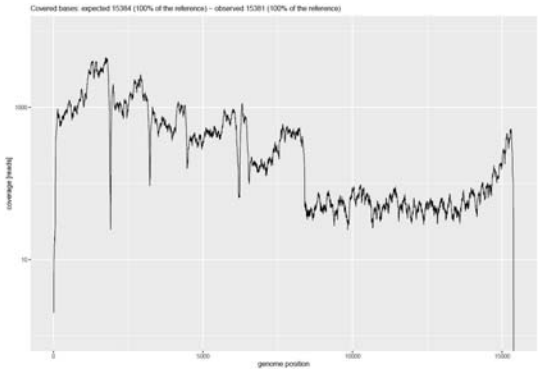
Detected[1] Virus Species

species	RNA workflow	total
Equine arteritis virus	1396102	1396102
Enterovirus B	201611	201611
Phocid alphaherpesvirus 1	15879	15879
Human metapneumovirus	21	21

[1]excluding phages and endogenous retroviruses

Rejected Virus Species

species	RNA workflow	total
Cynomolgus cytomegalovirus	340886	340886
Caulobacter phage Cer29	61874	61874
Enterovirus C	507	507
Feline leukemia virus	147	147
RD114 retrovirus	108	108
Staphylococcus phage Andhra	67	67
Human mastadenovirus C	59	59
Enterovirus J	6	6
Enterovirus A	4	4
Escherichia virus phiX174	4	4
Enterovirus D	2	2
Human betaherpesvirus 6A	1	1



PDF coverage plot (command line)