Recommendations for the introduction of metagenomic next-generation sequencing in clinical virology, part II: bioinformatic analysis and reporting

Vries, J.J.C. de; Brown, J.R.; Couto, N.; Beer, M.; Mercier, P. le; Sidorov, I.; ... ; ESCV Network Next-Generation Seque

Journal of Clinical Virology 138 (2021) 104812

Contents lists available at ScienceDirect

Journal of Clinical Virology

journal homepage: www.elsevier.com/locate/jcv

Recommendations for the introduction of metagenomic next-generation sequencing in clinical virology, part II: bioinformatic analysis and reporting

Jutte J.C. de Vries [a,*], Julianne R. Brown [b], Natacha Couto [c], Martin Beer [d], Philippe Le Mercier [e], Igor Sidorov [a], Anna Papa [f], Nicole Fischer [g], Bas B. Oude Munnink [h], Christophe Rodriquez [i], Maryam Zaheri [j], Arzu Sayiner [k], Mario Hönemann [l], Alba Pérez-Cataluña [m], Ellen C. Carbo [a], Claudia Bachofen [n], Jakub Kubacki [n], Dennis Schmitz [o], Katerina Tsioka [f], Sébastien Matamoros [p], Dirk Höper [d], Marta Hernandez [q], Elisabeth Puchhammer-Stöckl [r], Aitana Lebrand [e], Michael Huber [j], Peter Simmonds [s], Eric C.J. Claas [a], F. Xavier López-Labrador [t,u,v,**], on behalf of the ESCV Network on Next-Generation Sequencing

[a] *Clinical Microbiological Laboratory, department of Medical Microbiology, Leiden University Medical Center, Leiden, the Netherlands*
[b] *Microbiology, Virology and Infection Prevention & Control, Great Ormond Street Hospital for Children NHS Foundation Trust, London, United Kingdom*
[c] *Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom*
[d] *Friedrich-Loeffler-Institute, Institute of Diagnostic Virology, Greifswald, Germany*
[e] *Swiss Institute of Bioinformatics, Geneva, Switzerland*
[f] *Department of Microbiology, Medical School, Aristotle University of Thessaloniki, Greece*
[g] *University Medical Center Hamburg-Eppendorf, UKE Institute for Medical Microbiology, Virology and Hygiene, Germany*
[h] *Viroscience, Erasmus Medical Center, Rotterdam, the Netherlands*
[i] *Department of Virology, University hospital Henri Mondor, Assistance Public des Hopitaux de Paris, Créteil, France*
[j] *Institute of Medical Virology, University of Zurich, Switzerland*
[k] *Dokuz Eylul University, Medical Faculty, Department of Medical Microbiology, Izmir, Turkey*
[l] *Institute of Virology, Leipzig University, Leipzig, Germany*
[m] *Department of Preservation and Food Safety Technologies, IATA-CSIC, Paterna, Valencia, Spain*
[n] *Institute of Virology, University of Zurich, Switzerland*
[o] *RIVM National Institute for Public Health and Environment, Bilthoven, the Netherlands*
[p] *Medical Microbiology and Infection Control, Amsterdam UMC, Amsterdam, the Netherlands*
[q] *Laboratory of Molecular Biology and Microbiology, Instituto Tecnologico Agrario de Castilla y Leon, Valladolid, Spain*
[r] *Center of Virology, Medical University Vienna, Vienna, Austria*
[s] *Nuffield Department of Medicine, University of Oxford, Oxford, UK*
[t] *Virology Laboratory, Genomics and Health Area, Centre for Public Health Research (FISABIO-Public Health), Valencia, Spain*
[u] *Department of Microbiology, Medical School, University of Valencia, Spain*
[v] *CIBERESP, Instituto de Salud Carlos III, Madrid, Spain*

ARTICLE INFO

Keywords:
Viral metagenomics
NGS/HTS
Bioinformatics

ABSTRACT

Metagenomic next-generation sequencing (mNGS) is an untargeted technique for determination of microbial DNA/RNA sequences in a variety of sample types from patients with infectious syndromes. mNGS is still in its early stages of broader translation into clinical applications. To further support the development,

* Corresponding author at: Clinical Microbiological Laboratory, department of Medical Microbiology, Leiden University Medical Center, Leiden, the Netherlands.
** Corresponding author at: Virology Laboratory, Genomics and Health Area, Centre for Public Health Research (FISABIO-Public Health), Valencia, Spain.
E-mail addresses: jjcdevries@lumc.nl (J.J.C. de Vries), julianne.brown@gosh.nhs.uk (J.R. Brown), nmgdc20@bath.ac.uk (N. Couto), martin.beer@fli.de (M. Beer), Philippe.Lemercier@sib.swiss (P. Le Mercier), I.A.Sidorov@lumc.nl (I. Sidorov), annap@auth.gr (A. Papa), nfischer@uke.de (N. Fischer), b.oudemunnink@erasmusmc.nl (B.B. Oude Munnink), christophe.rodriguez@aphp.fr (C. Rodriquez), zaheri.maryam@virology.uzh.ch (M. Zaheri), arzu.sayiner@deu.edu.tr (A. Sayiner), Mario.Hoenemann@medizin.uni-leipzig.de (M. Hönemann), alba.perez@iata.csic.es (A. Pérez-Cataluña), E.C.Carbo@lumc.nl (E.C. Carbo), claudia.bachofen@uzh.ch (C. Bachofen), jakub.kubacki@uzh.ch (J. Kubacki), Dennis.Schmitz@RIVM.nl (D. Schmitz), aik.tsioka@gmail.com (K. Tsioka), s.p.matamoros@amsterdamumc.nl (S. Matamoros), dirk.hoeper@fli.de (D. Höper), hernandez.marta@gmail.com (M. Hernandez), elisabeth.puchhammer@meduniwien.ac.at (E. Puchhammer-Stöckl), aitana.lebrand@sib.swiss (A. Lebrand), huber.michael@virology.uzh.ch (M. Huber), peter.simmonds@ndm.ox.ac.uk (P. Simmonds), E.C.J.Claas@lumc.nl (E.C.J. Claas), F.Xavier.Lopez@uv.es (F.X. López-Labrador).

https://doi.org/10.1016/j.jcv.2021.104812
Received 9 February 2021; Accepted 20 March 2021
Available online 26 March 2021

Pipeline
Diagnostics
Pathogens

implementation, optimization and standardization of mNGS procedures for virus diagnostics, the European Society for Clinical Virology (ESCV) Network on Next-Generation Sequencing (ENNGS) has been established. The aim of ENNGS is to bring together professionals involved in mNGS for viral diagnostics to share methodologies and experiences, and to develop application guidelines. Following the ENNGS publication *Recommendations for the introduction of mNGS in clinical virology, part I: wet lab procedure* in this journal, the current manuscript aims to provide practical recommendations for the bioinformatic analysis of mNGS data and reporting of results to clinicians.

1. Introduction

Metagenomic next-generation sequencing (mNGS) is an untargeted technique for the determination of DNA/RNA sequences in a variety of clinical sample types from patients with infectious syndromes [1–3]. mNGS is suited for identification of any pathogen, including variants that have diverged at typical PCR amplification targets, pathogens not known to be associated with a specific clinical syndrome, and novel pathogens which may remain undetected by target-based methods [4,5]. Despite these clear advantages, mNGS is still in its early stages of translation into clinical application. One of the challenges in the clinical use of mNGS is the current lack of standardization of methods and workflows, including the bioinformatic analysis to ensure a fit-for-purpose, sensitive and specific pathogen detection. The performance of metagenomic methods is heavily dependent on accurate bioinformatic analysis, and both classification algorithms and databases are crucial factors determining the overall performance of available pipelines [6,7]. A wide range of metagenomic pipelines and taxonomic classifiers have been developed, often for the purpose of biodiversity studies analysing the composition of the microbiome in different samples and cohorts. In contrast, when applying mNGS for 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 pipelines developed by the authors themselves [8–12], stressing the need for high quality validation studies. The development of guidelines and recommendations on mNGS bioinformatic analysis methods and reporting will assist the implementation of mNGS in diagnostic laboratories, ensuring the validity of results and thus optimizing patient management.

To support the development and implementation of mNGS procedures for virus diagnostics, a network has been established under the auspices of the European Society for Clinical Virology (ESCV): the ESCV Network on Next-Generation Sequencing (ENNGS). The aim of this network is to bring together professionals involved in mNGS for viral diagnostics, to share materials, methodologies and experiences, and to develop recommendations for the implementation and use of mNGS in clinical diagnostics and Public Health laboratories.

2. Aim and scope

This review aims to provide recommendations for the implementation and validation of bioinformatic analysis methods for viral mNGS, excluding the wet lab part of the process, which has been discussed previously (Part I) [13] and is outside the scope of the current review. We aim to provide practical recommendations for analysis and reporting steps to aid in the successful implementation of fit-for-purpose mNGS procedures in viral diagnostic laboratories.

3. Recommendations

3.1. (Bio)informatic equipment and security

3.1.1. Bioinformatic software, expertise and information technology (IT) equipment

Processing of mNGS data is either provided by specialized bioinformaticians or non-bioinformaticians through user-friendly interfaces to tools and pipelines. Most metagenomic software pipelines are in the public domain and require expertise in bioinformatics. For the hardware part, options are i) the use of local computers, ii) the use of remote, more potent computers, including the use of cloud computing. Although some bioinformatic pipelines can be run in relatively modest desktop servers even directly in the laboratory, the recommended situation for routine clinical metagenomic analysis, which requires considerable computational capacity, is to have access to a cluster server which is usually situated within a dedicated physically separated "core" IT facility with infrastructure for central data processing, either accessible directly or via external providers of the analysis pipelines (Table 1, **Recommendation 1**). User-friendly software considerations are cloud-based platforms with web front-end interfaces which can facilitate direct uploading of the raw files from sequencing instruments and direct downloading of the final output analyses from the server. Examples of these interfaces and platforms are the Galaxy [14] and INSaFLU platforms (https://insaflu.insa.pt/) [15], server hosting (i.e. Amazon web services, Microsoft Azure), or cloud-based software solutions which can be scalable on-demand and frequently at lower operational costs (see Table 2). Finally, "third-generation" small sequencers based on nanopores that have relatively low capacity for metagenomic runs and are currently used for research applications, may simplify and streamline both the laboratory and bioinformatics processes, allowing for real-time analysis on a laptop computer, and futuristically, potentially near the bedside [16–18].

3.1.2. Data security

Data should be protected from unauthorized access and actions, loss, and destruction. Patient privacy should be guaranteed and justified use and governance of personal data, should be considered when implementing metagenomic procedures. The complexity and data management issues associated with NGS have led to an increasing number of diagnostic laboratories to turn to cloud services [19]. Cloud computing facilitates on-demand self-service, broad network access, resource pooling, and metering capabilities, but also means that the end user generally has no control or limited knowledge over the exact location of the provided computational services [19]. Therefore, it is recommended to have written agreements with cloud service providers on the management of protection of information for unauthorized access, use, disclosure, disruption, modification, or destruction, confidentiality, and timely/reliable access to and use of information (**Recommendation 2**). Furthermore, since accreditation of laboratory activities requires that every component of the assay must be verified prior to reporting patient test results, the agreement should include the management of new releases of software versions to enable validation prior to using a new version for patient care.

Table 1
Recommendations for the use of metagenomic sequencing for universal virus diagnostics.

Process step (paragraph)	Recommendations
Bioinformatic software, expertise and information technology (IT) equipment (3.1.1)	1. Given the amount of data and pipelines for metagenomic analysis, the use of a cluster server, usually situated within a dedicated physically separated "core" IT infrastructure facility for central data processing facilities is recommended, either accessible directly or via external providers of the analysis pipelines.
Data security (3.1.2)	1. It is recommended to have written agreements with cloud service providers on the management of protection information for unauthorized access, use, disclosure, disruption, modification, or destruction, confidentiality and timely/reliable access to and use of information. The agreement should also include the management of new releases of software versions to enable validation prior to using a new version for patient care.
Storage of raw data (3.2.1)	1. NGS FASTQ data and metadata files should be stored with file names and folders having unique and identifying names helpful in classifying and sorting (https://www.ukdataservice.ac.uk/manage-data/format/organising.aspx)
Data analysis: version control (3.2.4)	1. It is recommended to use version controlled pipeline tools and external databases used for NGS data analysis of clinical samples. For each tool used in the pipeline, at least the following parameters/options have to be described: date of analysis, name and version of the tools and external databases, as well as user-defined and default values of parameters used for each tool. Additionally, it is recommended to version the overall ensemble of tools, e.g. using a workflow tools/docker containers. Preferably, software should be made available via GitHub or GitLab to automatically handle version control to a large extent.
Reference database (3.2.6)	1. The reference database should consist of genomes that cover the entire genetic diversity of relevant organisms and should be curated in order not to contain any artificial, low-quality or incorrectly named genome sequences. 1. It is recommended to periodically update the reference databases used for taxonomic profiling, and to validate this update. The frequency of the update is dependent on the need to classify at subtype or isolate level, and on the appearance of novel viruses in the updated public databases.
Removal of contaminating sequences (3.2.7)	1. Taxa/sequences detected in the negative run control should be corrected for, either manually or automated. Automated removal of contaminating sequences should be validated.
Normalization of read counts (3.2.8)	1. Normalization of number of reads assigned to certain taxa by the total number of reads and by the genome size of the pathogen is recommended if quantitative or semi-quantitative results are issued.
Datasets for validation (4.1)	1. The bioinformatics pipeline should be evaluated using data from real samples, well-characterized by molecular diagnostic methods, which can be supplemented with analysis using *in silico* datasets.
Pipeline performance (4.2)	1. Pipeline performance: recall (sensitivity), precision (positive predictive value) and/or F1-score should be determined with real data sets from samples with a known status based on golden standard molecular diagnostic methods.
Threshold for defining a positive result (4.3)	1. For pathogen detection, the cut-off for defining a positive result has to be established during the validation phase by comparison with golden standard molecular techniques. Since the threshold is dependent on factors throughout the entire wet lab and analysis workflow, this will have to be determined for every protocol. The distribution of the reads across the genome has to be taken into account.
Result review and reporting (5)	1. Before reporting, the mNGS data need to be technically evaluated and reviewed, for quality, possible laboratory contaminations and plausibility. 1. Hits of known reagent contaminants, misassignments, bacteriophages, and common (retro)viral endogenous sequences should not be reported to the clinician.

3.2. Bioinformatic analysis

3.2.1. Storage of raw data

NGS FASTQ data and metadata files should be stored with file names and folders having unique and identifying names helpful in classifying and sorting (https://www.ukdataservice.ac.uk/manage-data/format/organising.aspx) [20]. (**Recommendation 3**). Recommended is to include e.g. the date of data delivery, the project team or (sub)department, project name, sequencing library number, unique sample identifiers such as sample number and date, with consistency over time and different people. A standardized submission protocol providing metadata and data handling is supported by a Laboratory Information and Management System (LIMS). Original data files saved in the folder as well as the folder itself should have read-only access and files in the folder should keep their original names supporting standards required for method accreditation (name of the FASTQ files containing Illumina reads typically includes flow cell number, sample name, sample number, machine lane number, type of the reads (R1/R2), for instance, "HK2LLDSXX_7074−09-002−001_CTGATCGT-ATATGCGC_L004_R1.fastq" and the name of the FAST5 files containing nanopore (ONT) raw electrical signals typically includes flowcelll number, run id and a consecutive number of the files generated per barcode, for instance, "FAK96194_5138107d5a8425587f0828dd31f396e3ebd774c4_1.fast5" and need to be converted into FASTQ format using for instance GUPPY). Most of the tools for NGS data processing accept files in the compressed formats 'tar', 'zip', or 'gzip'.

3.2.2. Data preprocessing

Sequence data quality can be visualized with e.g. FASTQC [21] (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ and MultiQC [22] and is followed by data pre-processing, which includes the removal of low-quality, low-complexity reads, bases (using PRINSEQ [23]) and sequence adapters using tools like Trimmomatic [24], and Cutadapt [25], with fairly comparable algorithms with minor differences in read counts after trimming. Some tools, e.g. Trimmomatic do

not auto-detect adapters and need an adapter file.

3.2.3. Removal of human sequences

Certain types of data analysis may require removal of ribosomal RNA reads or human reads prior to classification due to the ethical reasons/data protection rights and for speeding up downstream data analysis. Validation of efficacy of removal of human reads [26–29] is recommended in the light of the general data protection regulation.

3.2.4. Data analysis: version control

Downstream mNGS data analysis may be restricted to taxonomic classification of sequence reads or may alternatively include *de novo* assembly of reads into contigs or scaffolds, followed by aligning to a set of genomes, which requires selection of the tools for particular tasks and targets [30–33]. Currently there are no optimal or golden standard tools and different approaches can produce different results for the same FASTQ file. In a recent ENNGS comparison of viral metagenomic pipelines, performance was impacted by the overall components of specific pipelines including algorithm, settings, and database [34].

It is recommended to use version controlled pipeline tools and external databases used for NGS data analysis of clinical samples. For each tool used in the pipeline, at least the following parameters/options have to be described: date of analysis, name and version of the tools and external databases, as well as user-defined and default values of parameters used for each tool, e.g. using the version management tool (Bio)Conda [35]). Additionally, it is recommended to version the overall ensemble of tools, e.g. using a workflow tools/docker containers (e.g. Snakemake [36], Nextflow [37]) (**Recommendation 4**). Subsequently, storage of the workflow and its default settings can be hosted by GitHub/GitLab [38], a platform with built-in version control.

3.2.5. Taxonomic classification algorithms

Taxonomic profiling gives an insight into taxonomic composition of the samples analyzed and results obtained in defining relative abundances of organisms belonging to taxa at different taxonomic levels, for viruses primarily species, genus and family. Dependent on the specific clinical questions addressed, sequences may be further classified below the level of species, such as genotypes (of hepatitis B and C viruses), subtypes (HIV-1), or isolates, although this is beyond the remit of the taxonomy provided by the ICTV and beyond the ability to accurately sub-type varies between pipelines [34].

Reads can be classified using different algorithmic approaches that can handle large number of sequencing reads in a reasonable amount of time [39]. In order to do so, most algorithms use stretches of perfect sequence matches with reference sequences named k-mers. These tools can be divided into three groups: i) DNA-to-DNA classification (BLASTn-like; i.e. megaBLAST [40], Kraken [41]; Centrifuge [42], CLARK [43]), ii) DNA-to-protein (BLASTx-like; i.e DIAMOND [44], Kaiju [45], GenomeDetective [46], SURPI [47], RIEMS [48] and iii) marker-based classification (i.e. MetaPhlAn2 [49]). DNA-to-protein tools can be more sensitive to novel and highly variable sequences due to the lower mutation rates of amino acids compared with nucleotide sequences [45,50,51]. An aspect that should be taken into account when selecting a taxonomic classification algorithm is the precision versus recall trade-off. High recall usually comes at the cost of a decline in precision, meaning that false positive taxa are classified at low abundance levels [34,39,52]. Each read is usually assigned a particular score or confidence level by the taxonomic algorithm and this can be taken into account by any downstream application as a reliability estimator of the classification [53].

3.2.6. Reference database

Selection of the reference database can significantly influence the results of taxonomic classification [7]. The reference database should consist of genomes that cover the entire genetic diversity of relevant organisms and should be curated in order not to contain any artificial, low-quality or incorrectly named genome sequences (**Recommendation 5**). Poorly curated databases containing misannotated reference sequences will lead to false positive results due to incorrect assignment of ambiguously mapped/aligned reads or k-mers. Incomplete databases missing newly discovered or uncommon viral strains can lead to false negative results [54]. Database compression by removal of duplicate sequences [46] is an effective way to save storage space, but compression can lead to a decreased performance in pathogen detection [39]. In general, larger databases enable more accurate sub-typing/classification to isolate level.

Several viral databases are available to the scientific community (examples are shown in Table 3). Use of complete NCBI's GenBank nucleotide database [55] containing sequences assigned to viruses (NCBI: txid10239) contains redundant sequences, requires a lot of computer resources and leads to a number of false-positive virus assignments [7] as the GenBank database entries are not curated. In contrast, the non-redundant NCBI's RefSeq database [56] is relatively small by providing one sequence per species accurately assigned based on ICTV taxonomy, and importantly, well-curated, significantly reducing the number of false-positive assignments to provisional sequences that can be inaccurate. Viruses recently discovered and virus variants highly divergent from NCBI's RefSeq reference sequence may be unidentified, the latter also depending on the stringency of the mapping criteria of the classification algorithm used as described above [4]. In clinical diagnostic practice, NCBI's RefSeq database is commonly used for identification and classification of viruses and resulted in good overall performance in an international benchmark study [34]. Curated vertebrate virus genome databases have been proposed, conveniently for clinical diagnostics lacking non-vertebrate viruses, for example Virosaurus (https://viralzone.expasy.org/8676 [57]) with sequences that are clustered to remove redundancy.

With the exponential growth of the number genome sequences in public databases, it is important to periodically update the reference databases used for taxonomic profiling, and to validate this update (**Recommendation 6**). The frequency of the update is dependent on the need to classify at subtype or isolate level, and on the appearance of

Table 2
Examples of external providers of web-based user-friendly viral metagenomic analysis tools and interfaces.

Service offered	Provider	Scale	Metagenomic pipeline tool	Website	Citation
Software as a Service: Web-based viral metagenomic analysis tools with user-friendly interface	DNASTAR	Cloud/local	Viral metagenomic tools are included: complete service package from assembly to analysis and user-friendly web-based report	www.DNASTAR.com, www.dnastar.com/software/lasergene/	[34]
	Genome Detective	Cloud, local computer		www.genomedetective.com	[34,46]
	One Codex	Cloud, local computer		www.onecodex.com	[34,75]
	Taxonomer	Cloud		www.taxonomer.com	[34,76]
Platform/ Infrastructure as a Service: Web-based platforms with user-friendly interface for hosting in-house tools and pipelines	Galaxy	Cloud, cluster, local computer	Custom (in-house) pipeline provided by user to be translated onto web-based interface – bioinformatic expertise (user) required	https://usegalaxy.org	[14,34]
	BlueBee	Cloud		www.bluebee.com	[34,77]

Table 3
Some of the viral databases available for pathogen/viral mNGS.

Name of database (alphabetically)	Website	Description	Reference
FDA ARGOS	https://argos.igs.umaryland.edu/	Curated database of reference genomes of microbial sequences, for diagnostic use	[78]
ICTV Virus Metadata Resource (VMR)	https://talk.ictvonline.org/taxonomy/vmr/	Curated database of sequences of exemplars for each classified virus species	[79]
NCBI RefSeq	https://www.ncbi.nlm.nih.gov/refseq/	Curated database of annotated genomic, transcript, and protein sequence records: viruses (ca. 8500 complete viral genomes), prokaryotes, eukaryotes	[56]
NCBI GenBank	https://www.ncbi.nlm.nih.gov/genbank/	Uncurated database of all publicly available nucleotide sequences, annotated	[55]
Reference Viral DataBase (RVDB), nucleic version	https://rvdb.dbi.udel.edu/	Curated database of virus nucleotide sequences, available as Unclustered (U-) and Clustered (C-) nucleotide sequence files. Sequences determined to be irrelevant for virus detection are removed.	[80]
Reference Viral DataBase (RVDB, protein version	https://rvdb-prot.pasteur.fr/	Protein version (RVDB-prot and RVDB-prot-HMM) of the curated U-RVDB described above.	[81]
SIB Viral reference sequences	https://viralzone.expasy.org/6096	Curated database of annotated viral genomes generated by the Swiss Institute of Bioinformatics (SIB), including all sequences annotated as complete viral genomes (ca. 70,000) downloaded from GenBank (query "VRL[Division] AND 'complete genome' [ALL]") and subsequently screened for several criteria.	[71]
UniProt Virus proteomes	https://www.uniprot.org/proteomes/	Curated and annotated database of proteomic virus references (ca. 10,000 virus reference proteomes)	[82]
VirMet	https://github.com/medvir/VirMet	In-house database download from GenBank (query "txid10239[orgn] AND ("complete genome"[Title] OR srcdb_refseq[prop]) NOT wgs[PROP] NOT "cellular organisms"[Organism] NOT AC_000001[PACC]: AC_999999[PACC]"	[3]
Virosaurus	https://viralzone.expasy.org/8676	Curated database of virus sequences for clinical metagenomics, clustered (non-redundant), vertebrate viruses (ca. 24,000 sequences) can be downloaded separately or combined with non-vertebrate viruses.	[57,71]
Virus Pathogen Resource (ViPR)	https://www.viprbrc.org	Curated database of virus pathogens (ca. 1.000.000 genomes from ca. 7000 species) in the NIAD Category A–C Priority Pathogen lists and those causing (re) emerging infectious diseases. External sources: GenBank, UniProt, Immune Epitope Database, Protein Data Bank, etc.)	[83]

novel viruses in the updated public databases. Finally, some virus reference sequences contain stretches of human origin which can be initially noticed by consistent appearance of these hits. This type of misannotation can be detected by aligning the assigned sequencing reads with BLAST, whereby the top hits turn out to be of human origin in these cases. Tagging/blacklisting such entries may structurally prevent misannotation of sequences and false positive results.

3.2.7. *Removal of contaminating sequences*

Contamination can be introduced in several steps of the workflow, including nucleic acid extraction kits, reagents and diluents, post-sampling environment (i.e. airborne particles, index switching, cross-overs from past sequencing runs) and misclassification related to the classification algorithms used and/or the reference databases available [58,59]. As mentioned in Part I of these guidelines, positive and negative controls should be included in the sequencing run so post-sequencing contamination removal can be performed either manually or using computational algorithms (**Recommendation 7**). Two examples of such tools include Recentrifuge [59] and the R package Decontam [60]. These algorithms are based on different assumptions: while Recentrifuge classifies candidate contaminating taxa based on the relative frequency in the samples compared to controls and checks for crossover contamination, Decontam assumes that sequences from contaminating taxa are likely to have frequencies that inversely correlate with sample DNA concentration and are also likely to have a higher prevalence in control samples than in true samples (the contaminating species do not have to compete with true species in the negative control). Furthermore, Recentrifuge takes into account the score level of the classifications in every single step provided by the taxonomic classifier, therefore, removing potential false positive taxa introduced by the taxonomic algorithm. It must be taken into account that (low level) sequences detected in the negative run control not uncommonly originate from highly abundance species present in patient samples (e.g. due to index hopping). Automated removal of contaminating sequences should be validated (**Recommendation 7**). Alignment of sequence reads against a contaminant database (using bwa) can also be useful.

3.2.8. *Normalization of read counts*

For quantitative or semi-quantitative results, normalization of number of reads assigned to certain taxa by the total number of reads generated for each sample is useful since the number of generated sequencing reads might be considerably different between samples [61,62]. Additionally, differences in average genome sizes between taxa can also lead to misinterpretation of the results and, therefore, additional normalization by average genome length for each taxonomic group belonging to a certain taxonomic level is required, for example by reporting read counts per Kb of genome length per million reads [47,58] (**Recommendation 8**).

4. Validation of bioinformatic pipelines

4.1. *Datasets for validation*

The bioinformatics pipeline should be evaluated using data from real samples, well-characterized by molecular diagnostic methods, which can be supplemented with analysis using *in silico* datasets [63,64] (**Recommendation 9**). Artificial mNGS reads can be generated using the tools such as ART; CAMISIM [65] or other simulators, reviewed in [66]. By using simulated datasets, the impact of variable amounts of background sequences (e.g. reads of human or bacterial origin), different mutation rates, detection rate of less-related viral genomes, as well as multiple combinations of settings (single vs paired-end reads and different read lengths) can be tested [6].

4.2. Pipeline performance

Pipeline performance: recall (sensitivity), precision (positive predictive value) and/or F1-score should be determined with real data sets from samples with a known status based on golden standard molecular diagnostic methods (**Recommendation 10**). The F1 score is defined as the harmonic mean of sensitivity (recall/true positive rate) and precision [6]. Specificity analysis for mNGS methods is hampered by the immense high number of negative mNGS findings without available PCR result. By calculating the precision, the proportion of unknown true negative findings is conveniently avoided.

The limit of detection of the entire workflow should be determined in line with the intended use of the assay. Assessment of pipeline performance should include base calling, alignment, and target identification.

4.3. Threshold for defining a positive result

For pathogen detection, the threshold for defining a positive result has to be established during the validation phase by comparison with gold standard molecular techniques. Since virus read counts/distribution and thus the threshold is dependent on factors throughout the entire wet lab and analysis workflow, this will have to be determined for every protocol.

Recent validation work suggests that for robust identification of a positive result, non-overlapping reads mapping to three or more different genomic regions of the organism identified should be present [1,9,67]. A threshold based on read distribution seems more accurate than a threshold based (only) on the number of reads: high read numbers from amplicon contaminants will be mistakenly reported, and a few reads distributed over several genome locations of a pathogen may be missed when setting a strict threshold based on read counts [1,68,69]. Therefore, confirmation of positive results should include mapping reads to a relevant reference sequence of the identified organism/s resulting in genome coverage information, either as an automated part of the pipeline or as a secondary analysis (**Recommendation 11**). It must be noted that identification of bacteria would require different criteria [70].

4.4. Ring trials

Benchmarking of a variety of pipelines [63,64] has recently been performed by the ENNGS using datasets from clinical samples using RT-PCR as a gold standard [34]. A wide variety of viral metagenomic pipelines was used in the participating clinical diagnostic laboratories. In the benchmark, detection of low abundant viral pathogens and mixed infections remained a challenge. Benchmarks are required for accreditation purposes, can reveal less effective components of a workflow, and moreover, can point out best practices with regard to the common aim of the participants, the use of mNGS for clinical diagnostics. A ring trial organized by the Swiss Institute of Bioinformatics encountered the performance of both the wet and the dry lab procedures [71]. The QCMD has initiated a EQA scheme of metagenomic workflows in 2020 (Q4) using spiked samples. Clinical labs providing mNGS service should participate in ring trials or a formal EQA scheme where available; schemes that test both wet lab and bioinformatics are preferable.

5. Results Review and reporting

Before reporting, the mNGS data need to be technically evaluated and reviewed, for quality, possible laboratory contaminations and plausibility (**Recommendation 12**), which may be done in an interdisciplinary team consisting of molecular microbiology, bioinformatics, and clinical virology expertise [72]. This technical team should consider the quality of the run and the expected number of spike-in control reads. (Kit) contaminants, or sequences also detected in the no-template controls should be corrected for. For the evaluation and confirmation of a viral infection, the depth of coverage and number of different genome regions covered have to be taken into account (Fig. 1). Potential false positive hits based on classification misassignments can be manually detected using BLAST. Confirmatory PCRs targeting mNGS sequence hits are useful (in the early phase of implementation).

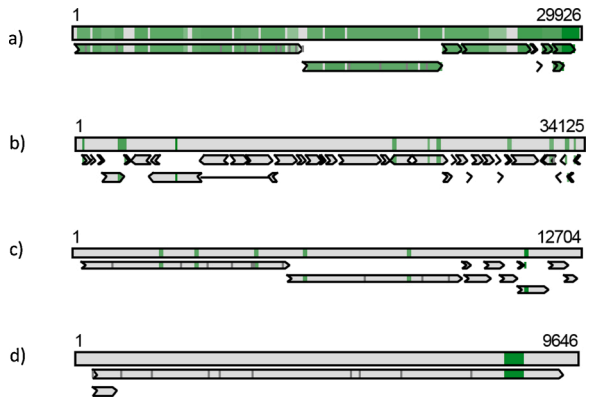

Fig. 1. Examples of coverage plots [46] with rue positive mNGS findings (a–c) confirmed by PCR in real clinical samples: a) human coronavirus HKU-1, 3951 reads, 89 % genome coverage, b) human mastadenovirus A, 19 reads, 8% genome coverage, >3 genome locations, and c) spiked-in equine arteritis virus, 14 reads, 5% genome coverage >3 genome locations, and d) an example of a false positive mNGS finding plotting a mapped hepatitis C virus amplicon contaminant, 133,213 reads, 4% coverage but only 1 genome location.
Top bar represents nucleotide alignment, bottom bar(s) represents amino acid alignment, green zone: matching sequences.
Distribution of reads over the genome is an important parameter for defining a positive result.

After technical review, the result of mNGS should be reported to the clinician in a compact format and facilitate decision making with regard to the treatment strategy and further diagnostic steps. Thus, the reports should be comprehensible, but yet easy to read and contain only clinically relevant or potentially relevant information. The essence of diagnostics is to identify potentially clinically relevant findings and interpret their significance. Therefore, hits of known reagent contaminants, misassignments, bacteriophages, and common (retro)viral endogenous sequences should not be reported to the requesting clinician [9] (**Recommendation 13**).

Pathogenic viruses detected as bystander, though not associated with the clinical syndrome at presentation, such as hepatitis C and HIV, can be detected by mNGS and should be reported. At the moment of clinical request of (viral) mNGS, the clinician should be informed about the potential to detect bystander pathogens [73]. This information can be available for example at the (digital) request form or in the diagnostic information booklet, and it should be made clear to the clinician that by performing the request for mNGS, virus identification in the broadest sense is agreed upon [74].

Viruses of unknown pathogenicity or uncommonly detected viruses may not have been associated with a specific disease before but at a later point in time may turn out to be associated with a specific syndrome, as seen with astrovirus encephalitis and thus reporting of these viruses is recommended. The interpretation of an unknown or potential association of the metagenomic finding in the particular patient can be discussed subsequently with the clinician or commented on at the report, for example in the case of low level detection of herpes viruses.

In case of the discovery of an exotic or novel agent, a literature review, personal discussion with the clinician and further virological testing may be required.

6. Conclusions

For some clinical syndromes, there is a need to extend the diagnostic

portfolio with mNGS. The recommendations provided here are intended to guide clinical diagnostics and Public Health laboratories on the implementation of viral mNGS bioinformatic pipelines and workflows. Bioinformatic software tools and platforms will develop very fast, and it is it is anticipated that these future developments will support the progressive and broad introduction of viral metagenomic sequencing into clinical diagnostics and Public Health laboratories.

Author contributions

Conceptualization: JV, XL, ECJC
Original draft: JV, XL, AP, IS, MHe, KT, SM, ECC, EPS
Review and editing: JRB, NC, MB, PLM, IS, AP, NF, BBOM, CR, MZ, AS, MH, APC, ECC, CB, JK, DS, KT, SM, DH, MH, EPS, AL, MH, PS, ECJC, XL

Declaration of Competing Interest

The authors declare no conflict of interest.

References

[1] E.C.B.E. Carbo, E. Karelioti, I. Sidorov, M.C.W. Feltkamp, P.A. von dem Borne, J.G. M. Verschuuren, A.C.M. Kroes, E.C.J. Claas, J.J.C. de Vries, Improved diagnosis of viral encephalitis in adult and pediatric hematological patients using viral metagenomics, bioRxiv. (2020), 06-05.
[2] C.Y. Chiu, S.A. Miller, Clinical metagenomics, Nat. Rev. Genet. 20 (6) (2019) 341–355.
[3] V. Kufner, A. Plate, S. Schmutz, L.D. Braun, F.H. GÃ¼nthard, R. Capaul, et al., Two years of viral metagenomics in a tertiary diagnostics unit: evaluation of the first 105 cases, Genes 10 (9) (2019).
[4] E.C. Carbo, I.A. Sidorov, J.C. Zevenhoven-Dobbe, E.J. Snijder, E.C. Claas, J.F. J. Laros, et al., Coronavirus discovery by metagenomic sequencing: a tool for pandemic preparedness, J. Clin. Virol. 131 (2020), 104594.
[5] P. Zhou, X.L. Yang, X.G. Wang, B. Hu, L. Zhang, W. Zhang, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature. 579 (7798) (2020) 270–273.
[6] T. Junier, M. Huber, S. Schmutz, V. Kufner, O. Zagordi, S. Neuenschwander, et al., Viral metagenomics in the clinical realm: lessons learned from a swiss-wide ring trial, Genes (Basel). 10 (9) (2019).
[7] S. van Boheemen, A.L. van Rijn, N. Pappas, E.C. Carbo, R.H.P. Vorderman, I. Sidorov, 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. 22 (2) (2020) 196–207.
[8] J. Chen, J. Huang, Y. Sun, TAR-VIR: a pipeline for TARgeted VIRal strain reconstruction from metagenomic data, BMC Bioinformatics 20 (1) (2019) 305.
[9] S. Miller, S.N. Naccache, E. Samayoa, K. Messacar, S. Arevalo, S. Federman, et al., Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid, Genome Res. 29 (5) (2019) 831–842.
[10] D. Paez-Espino, G.A. Pavlopoulos, N.N. Ivanova, N.C. Kyrpides, Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data, Nat. Protoc. 12 (8) (2017) 1673–1682.
[11] Y. Li, H. Wang, K. Nie, C. Zhang, Y. Zhang, J. Wang, et al., VIP: an integrated pipeline for metagenomics of virus identification and discovery, Sci. Rep. 6 (2016) 23774.
[12] S. Nooij, D. Schmitz, H. Vennema, A. Kroneman, M.P.G. Koopmans, Overview of Virus Metagenomic Classification Methods and Their Biological Applications, Front. Microbiol. 9 (2018) 749.
[13] F.X. Lopez-Labrador, J.R. Brown, N. Fischer, H. Harvala, S. Van Boheemen, O. Cinek, et al., Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure, J. Clin. Virol. 134 (2021), 104691.
[14] B. Giardine, C. Riemer, R.C. Hardison, R. Burhans, L. Elnitski, P. Shah, et al., Galaxy: a platform for interactive large-scale genome analysis, Genome Res. 15 (10) (2005) 1451–1455.
[15] V. Borges, M. Pinheiro, P. Pechirra, R. Guiomar, Jo P. Gomes, INSaFLU: An Automated Open Web-based Bioinformatics Suite â€œfrom-readsâ€ for Influenza Whole-genome-sequencing-based Surveillance, 2018, p. 46.
[16] H. Harstad, R. Ahmad, A. Bredberg, Nanopore-based DNA Sequencing in Clinical Microbiology: Preliminary Assessment of Basic Requirements, 2019, p. 382580.
[17] L.E. Kafetzopoulou, S.T. Pullan, P. Lemey, M.A. Suchard, D.U. Ehichioya, M. Pahlmann, et al., Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak, Science 363 (6422) (2019) 74.
[18] J. Quick, N.J. Loman, S. Duraffour, J.T. Simpson, E. Severi, L. Cowley, et al., Real-time, portable genome sequencing for Ebola surveillance, Nature 530 (2017) 228.
[19] A.B. Carter, Considerations for genomic data privacy and security when working in the cloud, J. Mol. Diagn. 21 (4) (2019) 542–552.
[20] T. Barrett, K. Clark, R. Gevorgyan, V. Gorelenkov, E. Gribov, I. Karsch-Mizrachi, et al., BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata, Nucleic Acids Res. 40 (Database issue) (2012) D57–63.
[21] U.H. Trivedi, T. Cezard, S. Bridgett, A. Montazam, J. Nichols, M. Blaxter, et al., Quality control of next-generation sequencing data without a reference, Front. Genet. 5 (2014) 111.
[22] P. Ewels, M. Magnusson, S. Lundin, M. Kaller, MultiQC: summarize analysis results for multiple tools and samples in a single report, Bioinformatics 32 (19) (2016) 3047–3048.
[23] R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets, Bioinformatics 27 (6) (2011) 863–864.
[24] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (15) (2014) 2114–2120.
[25] M. M, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnetjournal (2011).
[26] W.B. Langdon, Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks, BioData Min. 8 (1) (2015) 1.
[27] R. Schmieder, R. Edwards, Fast identification and removal of sequence contamination from genomic and metagenomic datasets, PLoS One 6 (3) (2011), e17288.
[28] S.J. Bush, T.R. Connor, T.E.A. Peto, D.W. Crook, A.S. Walker, Evaluation of methods for detecting human reads in microbial sequencing datasets, Microb. Genom. 6 (7) (2020).
[29] M.D. Czajkowski, D.P. Vance, S.A. Frese, G. Casaburi, GenCoF: a graphical user interface to rapidly remove human genome contaminants from metagenomic datasets, Bioinformatics 35 (13) (2019) 2318–2319.
[30] D.S. Horner, G. Pavesi, T. Castrignano, P.D. De Meo, S. Liuni, M. Sammeth, et al., Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing, Brief Bioinform. 11 (2) (2010) 181–197.
[31] A. Oulas, C. Pavloudi, P. Polymenakou, G.A. Pavlopoulos, N. Papanikolaou, G. Kotoulas, et al., Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies, Bioinform. Biol. Insights 9 (2015) 75–88.
[32] V. D'Argenio, Human microbiome acquisition and bioinformatic challenges in metagenomic studies, Int. J. Mol. Sci. 19 (2) (2018).
[33] T.D.S. Sutton, A.G. Clooney, F.J. Ryan, R.P. Ross, C. Hill, Choice of assembly software has a critical impact on virome characterisation, Microbiome 7 (1) (2019) 12.
[34] de Vries JJCea. Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples – submitted.
[35] B. Gruning, R. Dale, A. Sjodin, B.A. Chapman, J. Rowe, C.H. Tomkins-Tinch, et al., Bioconda: sustainable and comprehensive software distribution for the life sciences, Nat. Methods 15 (7) (2018) 475–476.
[36] J. Koster, S. Rahmann, Snakemake-a scalable bioinformatics workflow engine, Bioinformatics. 34 (20) (2018) 3600.
[37] P. Di Tommaso, M. Chatzou, E.W. Floden, P.P. Barja, E. Palumbo, C. Notredame, Nextflow enables reproducible computational workflows, Nat. Biotechnol. 35 (4) (2017) 316–319.
[38] J.D. Blischak, E.R. Davenport, G. Wilson, A Quick Introduction to Version Control with Git and GitHub, PLoS Comput. Biol. 12 (1) (2016), e1004668.
[39] S.H. Ye, K.J. Siddle, D.J. Park, P.C. Sabeti, Benchmarking metagenomics tools for taxonomic classification, Cell. 178 (4) (2019) 779–794.
[40] A. Morgulis, G. Coulouris, Y. Raytselis, T.L. Madden, R. Agarwala, A.A. Schaffer, Database indexing for production MegaBLAST searches, Bioinformatics. 24 (16) (2008) 1757–1764.
[41] D.E. Wood, S.L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments, Genome Biol. 15 (3) (2014) R46.
[42] D. Kim, L. Song, F.P. Breitwieser, S.L. Salzberg, Centrifuge: rapid and sensitive classification of metagenomic sequences, Genome Res. 26 (12) (2016) 1721–1729.
[43] R. Ounit, S. Wanamaker, T.J. Close, S. Lonardi, CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers, BMC Genomics 16 (2015) 236.
[44] B. Buchfink, C. Xie, D.H. Huson, Fast and sensitive protein alignment using DIAMOND, Nat. Methods 12 (1) (2015) 59–60.
[45] P. Menzel, K.L. Ng, A. Krogh, Fast and sensitive taxonomic classification for metagenomics with Kaiju, Nat. Commun. 7 (2016) 11257.
[46] M. Vilsker, Y. Moosa, S. Nooij, V. Fonseca, Y. Ghysens, K. Dumon, et al., Genome Detective: an automated system for virus identification from high-throughput sequencing data, Bioinformatics 35 (5) (2019) 871–873.
[47] S.N. Naccache, S. Federman, N. Veeraraghavan, M. Zaharia, D. Lee, E. Samayoa, et al., A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples10.1101/gr.171934.113, Genome Res. 24 (7) (2014) 1180–1192.
[48] M. Scheuch, D. Hoper, M. Beer, RIEMS: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets, BMC Bioinformatics 16 (2015) 69.
[49] D.T. Truong, E.A. Franzosa, T.L. Tickle, M. Scholz, G. Weingart, E. Pasolli, et al., MetaPhlAn2 for enhanced metagenomic taxonomic profiling, Nat. Methods 12 (10) (2015) 902–903.
[50] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410.
[51] D. Hoper, C. Wylezich, M. Beer, Loeffler 4.0: Diagnostic Metagenomics, Adv. Virus Res. 99 (2017) 17–37.

[52] A.L. van Rijn, S. van Boheemen, I. Sidorov, E.C. Carbo, N. Pappas, H. Mei, et al., The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease, PLoS One 14 (10) (2019) e0223952.
[53] J.M. Marti, Correction: Recentrifuge: Robust comparative analysis and contamination removal for metagenomics, PLoS Comput. Biol. 15 (6) (2019), e1007131.
[54] A. Dulanto Chiang, J.P. Dekker, From the pipeline to the bedside: advances and challenges in clinical metagenomics, J. Infect. Dis. (2019).
[55] D.A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, et al., GenBank, Nucleic Acids Res. 41 (Database issue) (2013) D36–42.
[56] N.A. O'Leary, M.W. Wright, J.R. Brister, S. Ciufo, D. Haddad, R. McVeigh, et al., Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, Nucleic Acids Res. 44 (D1) (2016) D733–45.
[57] A. Gleizes, F. Laubscher, N. Guex, C. Iseli, T. Junier, S. Cordey, et al., Virosaurus A reference to explore and capture virus genetic diversity, Viruses 12 (11) (2020).
[58] R. Schlaberg, C.Y. Chiu, S. Miller, G.W. Procop, G. Weinstock, Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection, Arch. Pathol. Lab. Med. 141 (6) (2017) 776–786.
[59] J.M. Marti, Recentrifuge: Robust comparative analysis and contamination removal for metagenomics, PLoS Comput. Biol. 15 (4) (2019), e1006967.
[60] N.M. Davis, D.M. Proctor, S.P. Holmes, D.A. Relman, B.J. Callahan, Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data, Microbiome. 6 (1) (2018) 226.
[61] M.B. Pereira, M. Wallroth, V. Jonsson, E. Kristiansson, Comparison of normalization methods for the analysis of metagenomic gene abundance data, BMC Genomics 19 (1) (2018) 274.
[62] P. Li, Y. Piao, H.S. Shon, K.H. Ryu, Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data, BMC Bioinformatics 16 (2015) 347.
[63] A. Brinkmann, A. Andrusch, A. Belka, C. Wylezich, D. Hoper, A. Pohlmann, et al., Proficiency testing of virus diagnostics based on bioinformatics analysis of simulated in silico high-throughput sequencing data sets, J. Clin. Microbiol. 57 (8) (2019).
[64] D. Hoper, J. Grutzke, A. Brinkmann, J. Mossong, S. Matamoros, R.J. Ellis, et al., Proficiency testing of metagenomics-based detection of food-borne pathogens using a complex artificial sequencing dataset, Front. Microbiol. 11 (2020), 575377.
[65] A. Fritz, P. Hofmann, S. Majda, E. Dahms, J. DrÃ¶ge, J. Fiedler, et al., CAMISIM: simulating metagenomes and microbial communities, Microbiome. 7 (1) (2019) 17.
[66] M. Zhao, D. Liu, H. Qu, Systematic review of next-generation sequencing simulators: computational tools, features and perspectives, Brief. Funct. Genomics 16 (3) (2017) 121–128.
[67] S.N. Naccache, A. Greninger, E. Samayoa, S. Miller, C.Y. Chiu, Clinical utility of unbiased metagenomic next-generation sequencing in diagnosis of acute infectious diseases: a prospective case series, Open Forum Infect. Dis. 2 (Suppl 1) (2015). DO - 10.1093/ofid/ofv131.23):103.
[68] S.C. Inzaule, R.L. Hamers, M. Noguera-Julian, M. Casadella, M. Parera, C. Kityo, et al., Clinically relevant thresholds for ultrasensitive HIV drug resistance testing: a multi-country nested case-control study, Lancet HIV 5 (11) (2018) e638-e46.
[69] J.R. Brown, S. Morfopoulou, J. Hubb, W.A. Emmett, W. Ip, D. Shah, et al., Astrovirus VA1/HMO-C: an increasingly recognized neurotropic pathogen in immunocompromised patients, Clin. Infect. Dis. 60 (6) (2015) 881–888.
[70] K. Mongkolrattanothai, S.N. Naccache, J.M. Bender, E. Samayoa, E. Pham, G. Yu, et al., Neurobrucellosis: Unexpected Answer From Metagenomic Next-Generation Sequencing, J. Pediatric Infect. Dis. Soc. 6 (4) (2017) 393–398.
[71] T. Junier, M. Huber, S. Schmutz, V. Kufner, O. Zagordi, S. Neuenschwander, et al., Viral metagenomics in the clinical realm: lessons learned from a swiss-wide ring trial, Genes 10 (9) (2019).
[72] M.R. Wilson, H.A. Sample, K.C. Zorn, S. Arevalo, G. Yu, J. Neuhaus, et al., Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis, N. Engl. J. Med. 380 (24) (2019) 2327–2340.
[73] R.J. Hall, J.L. Draper, F.G. Nielsen, B.E. Dutilh, Beyond research: a primer for considerations on using viral metagenomics in the field and clinic, Front. Microbiol. 6 (2015) 224.
[74] S.B. Johnson, I. Slade, A. Giubilini, M. Graham, Rethinking the ethical principles of genomic medicine services, Eur. J. Hum. Genet. 28 (2) (2020) 147–154.
[75] S.Sea Minot, One Codex: a sensitive and accurate data platform for genomic microbial identification, bioRxiv. (2015).
[76] S. Flygare, K. Simmon, C. Miller, Y. Qiao, B. Kennedy, T. Di Sera, et al., Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling, Genome Biol. 17 (1) (2016) 111.
[77] S. Morfopoulou, V. Plagnol, Bayesian mixture analysis for metagenomic community profiling, Bioinformatics 31 (18) (2015) 2930–2938.
[78] H. Sichtig, T. Minogue, Y. Yan, C. Stefan, A. Hall, L. Tallon, L. Sadzewicz, S. Nadendla, W. Klimke, E. Hatcher, M. Shumway, A. Lebron Dayanara, et al., FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science, Nature Communications 10 (2019) 3313.
[79] E.J. Lefkowitz, D.M. Dempsey, R.C. Hendrickson, R.J. Orton, S.G. Siddell, D. B. Smith, Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV), Nucleic Acids Res. 46 (2018) D708–D717.
[80] N. Goodacre, A. Aljanahi, S. Nandakumar, M. Mikailov, A.S. Shan, A Reference Viral Database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection, mSphere 3 (2018) e00069–18.
[81] T. Bigot, S. Temmam, P. Perot, M. Eliot, RVDB-prot, a reference viral protein database and its HMM profiles, F1000Res 8 (2019).
[82] The UniPrpt Consortium, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res 47 (2019) D506–D515.
[83] B. Pickett, D. Greer, Y. Zhang, L. Stewart, L. Zhou, G. Sun, et al., Virus Pathogen Database and Analysis Resource (ViPR): A Comprehensive Bioinformatics Database and Analysis Resource for the Coronavirus Research Community, Viruses 4 (11) (2012) 3209–3226.