Chapter 7
Metagenomic Applications for Infectious Disease Testing in Clinical Laboratories

Laura Filkins and Robert Schlaberg

7.1 Introduction

An explosion of technological advancements in clinical microbiology over the past two decades is rapidly transforming the laboratory diagnosis of infectious disease. Some of the most influential advancements include introduction of rapid organism identification by matrix-assisted laser desorption-ionisation time-of-flight (MALDI-TOF) mass spectrometry and DNA sequencing of marker genes, increased availability of direct-from-specimen nucleic acid detection tests (NAAT, including syndromic panels), targeted detection of genetic markers to rapidly predict antimicrobial resistance [1, 2]. These methods decrease time-to-results, provide accurate identification and improved sensitivity compared to classic methods, enable clinicians to select optimal antimicrobial therapy sooner, and reduce overuse of antibiotics [3, 4].

While the current clinical microbiology methods have greatly improved routine diagnostics, these approaches have limitations. Both culture-dependent and independent methods are only able to detect a limited repertoire of organisms. Utilising these methods, only targeted (pre-selected), viable, and/or culturable microorganisms will be detected. Additionally, strains exhibiting non-standard phenotypes (biochemical identification), altered protein expression profiles (MALDI-TOF), or genetic variation (NAAT) within the targeted micro-organism groups may lead to incorrect or false-negative results. For NAAT, frequent test redesign may be

L. Filkins
Departement of Pathology, University of Texas Southwestern Medical Center, Dallas, TX, USA

R. Schlaberg (✉)
Departement of Pathology, University of Utah, Salt Lake City, UT, USA
e-mail: robert.schlaberg@path.utah.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

J. Moran-Gilad, Y. Yagel (eds.), *Application and Integration of Omics-powered Diagnostics in Clinical and Public Health Microbiology*,
https://doi.org/10.1007/978-3-030-62155-1_7

necessary, especially when new pathogens emerge as has recently been highlighted by the need to design, manufacture, validate, and distribute new NAAT to detect the emerging SARS-CoV-2. Further, differentiating strains of the same species (strain typing) for diagnostic, surveillance, and infection prevention purposes usually requires additional testing, which limits availability and timeliness of results. Metagenomic next-generation sequencing (NGS) directly from patient specimens in clinical laboratories (clinical metagenomics) helps overcome these challenges as it provides a hypothesis-free, genome-based, high-resolution alternative to conventional testing. Clinical metagenomics enables detection of organisms that are difficult to culture, slow growing, genetically divergent, while also providing genotypic information for the purpose of strain-typing or prediction of antimicrobial resistance.

As clinical metagenomic testing is adopted by a rapidly growing number of laboratories the need for standardised, streamlined, high quality, and compliant workflows increases. In this chapter, we present an overview of current technologies, remaining challenges, and approaches to overcome them. We define metagenomic sequencing as the process of sequencing nucleic acid (RNA and/or DNA) directly from clinical specimens, including the use of workflows that apply target enrichment, host depletion or other pre-sequencing steps.

7.2 Clinical Need for Advanced Testing

The efficacy of conventional diagnostics varies based on the clinical syndrome, patient population, and breadth of available diagnostic resources. The most challenging clinical syndromes to diagnose are those that present with non-specific symptoms, have a broad differential, and are unresponsive to empiric therapy. Strong interest is placed on the application of metagenomic testing for the diagnosis of meningitis/encephalitis, pneumonia, fever of unknown origin (FUO), bone and joint infections, intraocular infections, and others. Glaser and colleagues reported that a likely aetiologic agent of encephalitis was identified in less than 40% of patients enrolled in the California Encephalitis Project [5]. Similarly, diagnosis of community acquired and healthcare associated pneumonias is challenging with current testing approaches returning negative results in 20–60% of cases [6–8]. Further, determining the true aetiologic agent of pneumonia when one or more potential pathogens are detected often requires additional scrutiny and clinical interpretation, especially with pathogens that are highly prevalent, can also be commensals, persist after an acute infection, or causes varying disease severity [9]. In prosthetic joint infections, conventional culture methods fail to identify the causative microorganism in about 40–50% of cases. Broad-range PCR or NGS can increase the diagnostic yield by 25% or more compared to culture [10, 11]. Sequencing of cell-free DNA (cfDNA) from plasma has recently been applied for the detection of micro-organisms associated with numerous clinical indications including sepsis, FUO, pneumonia, deep-seated infections, and others [12–15]. Finally, clinical metagenomics is a promising approach for the diagnosis of intraocular infections.

The very small specimen quantity that is obtainable from intraocular sources limits the number of NAAT and culture testing that can be performed. Using current molecular methods, fungi and viruses can be detected with >90% sensitivity and 75–90% sensitivity for bacterial detection from ocular sources, but achieving these sensitivities requires multiple assays and relatively large specimen volume [16]. Metagenomic sequencing provides a unified testing alternative that requires less specimen volume than a combination of bacterial, mycobacterial, fungal and viral culture, and multiple pathogen-specific NAAT [17].

Metagenomics can provide a diagnosis in many challenging diagnostic scenarios when conventional methods may be unsuccessful, as discussed below (Fig. 7.1). Additional applications of NGS in clinical microbiology include antimicrobial resistance (AMR) prediction, molecular epidemiology, and microbiome community profiling which are not covered here [18–21] but in other chapters of this book.

Clinical metagenomics can decrease time-to-results for slow growing or hard to diagnose micro-organisms, provide rapid, high-resolution micro-organism identification, resistance prediction to support optimal treatment choices, and reduce costs

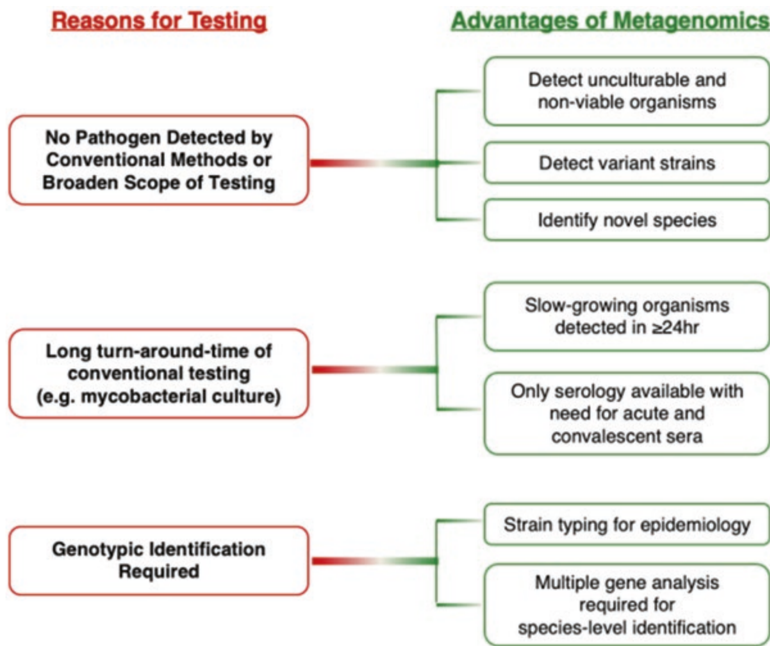

Fig. 7.1 Applications for metagenomic pathogen detection. Untargeted metagenomic next-generation sequencing is a culture-independent method that identifies pathogens by microbial nucleic acid detection directly from the patient specimen. This method detects nucleic acid from viable or non-viable cells and extracellular nucleic acids. Detection of variant strains and novel organisms is possible. Turnaround time for metagenomics test results is variable, typically ranging from 24 hours to 2 weeks, depending on the test design. Finally, whole genome or multiple gene sequencing enables specific classification of micro-organisms, even those that are phylogenetically closely related

by providing a comprehensive approach to answering multiple diagnostic questions [22]. However, for patients and clinicians to benefit from these advantages, significant changes in diagnostic algorithms and laboratory workflows will be required. While case reports have been published in diverse disease areas highlighting the power of clinical metagenomics, few clinical trials systematically comparing diagnostic yield and clinical benefit to standard of care have been conducted. Some of the most notable case reports and case series are based on cerebrospinal fluid (CSF) testing yielding unexpected aetiologies, including neuroleptospirosis in a pediatric patient with severe combined immunodeficiency, chronic meningoencephalitis due to Cache Valley Virus in a patient with X-linked agammaglobulinaemia, and neurobrucellosis in a female paediatric patient [23–25]. In a multicentre clinical trial for diagnosis of meningitis and encephalitis, 27.9% of enrolled patients were ultimately diagnosed with an infectious pathogen; of which, 33% of pathogens were detected by both conventional testing and metagenomic NGS on CSF, 45% by conventional testing only, and 22% by metagenomic NGS only [26]. The SEP-SEQ study employed metagenomic pathogen detection from plasma cfDNA and demonstrated detection of probable infectious causes of sepsis in an additional 15% of patients undiagnosed by conventional testing [15]. Similarly, Long et al. showed increased bacterial detection using plasma cfDNA compared to standard blood culture alone in patients with suspected sepsis in addition to detection of viral pathogens in 18% of the patients [27]. Testing of nasopharyngeal/oropharyngeal swabs from children hospitalised for community-acquired pneumonia by next-generation RNA sequencing identified previously missed putative pathogens in approximately 30% of patients [28].

Metagenomics is also a powerful tool to discover novel or emerging pathogens that escape detection by conventional methods. An early example was the identification of a novel rhabdovirus in serum from a patient with haemorrhagic fever [29]. Since then, pathogen (especially viral) discovery has been accelerated by the use of metagenomic sequencing and led to the detection of Henan Fever Virus, a novel bunyavirus in patients with fever, thrombocytopenia, and leukaemia syndrome; a novel arenavirus related to the lymphocytic choriomeningitis viruses in a cluster of fatal organ transplants; Lujo virus, an arenavirus first discovered from an outbreak of five cases of hemorrhagic fever in South Africa [30–32]; the recently discovered SARS-CoV-2; and many others.

These and other success stories were among the first evidence demonstrating the power of metagenomics-based pathogen detection for clinical diagnosis. However, until recently testing workflows and equipment were too slow, too expensive, required too much expert knowledge, and bioinformatics skills to be implemented in clinical laboratories. As these barriers are being removed, clinical metagenomics is increasingly being implemented in routine diagnostic algorithms. Further optimising, streamlining, and accelerating specimen preparation and sequencing technologies, standardising micro-organism identification, result interpretation, and quality control methods will facilitate adoption by clinical laboratories.

7.3 Test Design and Development

The potential benefits of utilising NGS technologies in clinical microbiology has been strongly demonstrated with case reports and initial clinical studies, including those described above. While the gains are substantial, as with any new technology, performance must be characterised for each clinical application and testing approach so that risks can be mitigated. Here, we describe technical and clinical challenges of metagenomic analyses for infectious disease diagnosis and suggest approaches to improve test characteristics while minimising sources of potential error.

7.3.1 Pre-Analytic Factors

As with any laboratory test, pre-analytic factors affect performance. Relevant factors include appropriate patient selection, defining relevant specimen types, specimen collection, preservation, transport, and storage conditions, and determining specimen stability. Pre-analytical steps need to be controlled and specimen rejection criteria need to be defined [33]. These factors are not unique to metagenomic testing but can affect the testing outcome differently than other microbiologic methods. For example, leaving a sputum specimen at room temperature for extended periods may result in reduced viability of fastidious pathogens, which can affect their recovery by culture. Results for that same specimen could be affected by over-growth of normal flora or degradation of pathogen nucleic acids limiting the sensitivity of metagenomics. Either scenario could cause decreased test sensitivity and would likely change the interpretation of results.

7.3.2 Specimen Preparation

At minimum, wet bench processes include nucleic acid extraction, library preparation, and sequencing. Additional wet bench procedures can enhance detection of pathogen-derived nucleic acids during sequencing, such as pathogen enrichment and removal of host nucleic acid or cells.

7.3.2.1 Nucleic Acid Extraction

Efficient, nucleic acid isolation is essential for producing high-quality sequencing libraries and reliable pathogen identification results. The target nucleic acid of interest (RNA versus DNA, or both) must be determined during test design to best address the needs of the test. DNA is useful when evaluating bacterial, fungal, and eukaryotic targets and for detection of DNA viruses. However, RNA extraction

sequencing is required for detecting RNA viruses. DNA enables whole-genome sequencing, whereas RNA sequencing is limited to those genes that are actively expressed. RNA can be advantageous for detection of pathogens that have high levels of gene expression, as the number of nucleic acids in the sequencing dataset is amplified compared to the amount of nucleic acid that would be present from a genome. Conversely, latent infections with quiescent micro-organisms may be more difficult to detect using RNA compared to DNA.

Specimen type and relevant pathogens guide selection of extraction methods [34]. Tissue and stool typically require more aggressive methods, such as mechanical lysis or bead beating, to release nucleic acids due to the physical composition of the specimen, whereas for other specimens, such as CSF or synovial fluid, chemical lysis is sufficient. Target micro-organisms with thick cell walls, including many fungi, usually require a mechanical, bead beating lysis method. Finally, the use of high purity plasticware and reagents (i.e. tubes, columns, buffers) with low levels of contaminating nucleic acids reduces detection of background organisms.

7.3.2.2 Pathogen Enrichment

A common challenge of untargeted metagenomic analysis for pathogen detection is the significant proportion of sequencing reads that are derived from host nucleic acid. Host cells or free nucleic acids compete with pathogen nucleic acid during sequencing and can reduce the analytical strength. Methods to enrich pathogen-derived and/or reduce host-derived nucleic acids can improve analytical sensitivity while reducing sequencing costs by reducing the depth of sequencing required to detect low abundance organisms. Target enrichment can be achieved by capture of pathogen nucleic acid or PCR-based enrichment. A common approach in microbiome studies focused specifically on bacterial or fungal communities is PCR amplification of marker genes followed by next-generation sequencing of PCR amplicons. Broad-range primers usual target conserved regions of the 16S rRNA gene (bacteria) or ITS region (fungi), or other highly conserved genes and are applied for amplification from total DNA (or less commonly cDNA) [35]. The resulting bacterial or fungal-enriched nucleic acid pool is then used for library preparation. This broadly targeted approach is also utilised for analysis of clinical specimens when suspicion for bacterial or fungal aetiology is strong, however detection of a causative micro-organism is limited to the selected category. Multiple primer enrichment can similarly be used to increase nucleic acid quantities for viral detection [36]. Capture-based enrichment methods have also been employed to select for sequences from organisms of interest [37]. However, bias is introduced by both broad targeting amplification methods, random amplification methods, and sequence capture [38–40]. Therefore, bias should be closely evaluated and characterised during clinical test development when these enrichment methods are used.

For untargeted metagenomic approaches, host-depletion is an important consideration and can increase detection of pathogens [41]. A variety of host-depletion methods exist and are applied at different steps within the sequencing workflow.

One approach is to deplete host nucleic acids before extraction. Allander et al. demonstrated improved detection of enveloped DNA viruses after treatment of serum with DNase to reduce extracellular host DNA [42]. For RNA sequencing, the removal of highly abundant host RNA includes ribosomal RNAs (most specimen types) or globin transcripts (whole blood specimens). Common methods of targeted RNA depletion include probe-based removal and target cleavage after nucleic acid extraction. See further discussion in Chap. 8.

7.3.2.3 Library Preparation

Sequencing libraries preparation methods have improved rapidly involving fewer and fewer steps, becoming faster to perform (often within a few hours or less), and can be automated on routine liquid handling instruments. Workflows are further streamlined by methods that limit the need for quality control and quantification of sequencing libraries. For optimal efficiency and to reduce costs, laboratories usually pool multiple barcoded libraries for sequencing on one sequencing run. Barcodes should be selected and demultiplexing parameters should be defined to limit misassociation of sequencing reads ("index cross-talk", "barcode hopping") as this can cause false-positive results [43, 44]. Strategies include dual indexing and design or selection of barcodes with maximal edit distance. Ideal sequencing datasets are diverse, containing numerous different reads mapping to each target microorganism. Therefore, library preparation methods that produce libraries with minimal duplication and increased diversity of reads typically yields higher-quality sequencing datasets and higher-confidence pathogen detection.

7.3.2.4 Sequencing

The selection of a sequencing platform is a critical step in the design phase of test development. Considerations should include resources already available at the institution, capital expense requirements, complexity of specimen preparation, reagent and sequencing run cost, desired read length, total read number per run, sequencing run time, sequencing error profile, and bioinformatics/analysis support. Prioritisation of these variables for individual applications in clinical microbiology will vary making one sequencing platform preferable over another given the precise needs of the test. Sequencing platform characteristics have been summarised elsewhere [35, 45, 46]. A comparison of Illumina sequencing platform (short read) versus Oxford Nanopore MinION platform (long read) of stool from pre-term infants demonstrated that long reads improved species-level detection for some bacteria, while the high error rate of the MinION prevented species-level identification for other bacterial genera that were successfully identified by Illumina [47]. This example highlights the challenges of either approach (long versus short read length) and the importance of tailoring the test design for the goal of an individual test.

The sensitivity of detection and specificity of micro-organism identification in specimens with low pathogen load is improved with increased sequencing depth, especially when de novo genome assembly is required for identification [48]. Unfortunately, increasing sequencing depth comes with increased cost and often longer run times. Speeding up the sequencing step of metagenomics workflows is a high priority for clinical applications as infectious disease testing requires a more rapid turn-around time than other genomics applications. In contrast to short read sequencing platforms, some long read technologies allow real time analysis which can be used to terminate sequencing when sufficient data has been generated. Using the Oxford Nanopore MinION, Greninger et al. demonstrated that sufficient sequencing data could be achieved to identify viral pathogens in high load serum specimens with <10 minutes of sequencing time, whereas moderate load specimens required 30–40 minutes [49].

7.3.3 Sequence Analysis

Clinical metagenomics presents unique challenges when compared to academic discovery applications. In research settings, the focus is often on comprehensive analyses (e.g. whole genome sequences) and increased time for computation and manual analysis by experts are acceptable. Additionally, multiple different analysis approaches are frequently used, often in a batched mode for all specimens that are part of a given study, to extract all pertinent genetic information and/or enable quantification of gene expression or organism abundance. In contrast, clinical testing requires testing and interpretation of results on a daily basis by a number of operators, strict adherence to pre-determined and validated procedures and interpretative criteria. Software used for data analysis needs to be diagnostic grade, version controlled, regularly updated, and meet data protection and privacy requirements. All procedures must be thoroughly vetted and turnaround time (TAT) for computational analysis steps are essential to the clinical utility of metagenomic tests. The selection of all analysis steps, including run quality pass/fail, read quality filtering, read classification (for organism detection) and/or contig assembly (for strain typing and de novo discovery), micro-organism determination, and reporting needs to be carefully determined based on the clinical application.

7.3.3.1 Sequence Analysis Tools

Preferred sequence data analysis methods may depend on the intended use of the test and the type of results that need to be generated [50–56]. Numerous bioinformatics tools have been published for research applications and vary in their approach to analysing sequences, accuracy and sensitivity of read classification, run time, and other characteristics [51, 52, 54, 56]. Requirements for data analysis software used in diagnostic workflows and need for bioinformatics support have to be taken into

consideration when determining sequence analysis strategies. General approaches for sequence data analysis and read classification include alignment-based and alignment-free methods (k-*mer* based), use of whole genome or marker gene-based approaches (e.g. *rRNA*s, other conserved genes) [57–66].

Analysis time is a critical characteristic for clinical NGS-based tests, as extended TAT limits clinical utility. General approaches to faster read classification include reducing the number of sequence comparisons by limiting the number of query (i.e. reads per specimen) or reference sequences (i.e. database size) and utilising faster sequence comparisons tools (i.e. faster alignment or alignment-free methods) [57]. Reducing the number of query sequences is most commonly achieved by removal of duplicate reads, binning or clustering of sequences before querying and subsequent querying of a single representative sequence for each cluster and assembly of sequences into longer contigs [67]. Database sizes can be reduced by limiting redundancy while representing as much sequence diversity as possible [68]. However, for clinical diagnostics reducing reference database sizes carries substantial risk for loss of performance via higher rates of false negative (pathogen-derived reads do not match the representative sequence closely enough to be identified) and false positive results (mis-assignment of reads to the next-closest reference sequence if a better, correct match is missing). Thus, database design is a critical component of clinical metagenomics tests. Many open source sequence analysis tools (e.g. Kraken) allow users to provide their own reference sequences, allowing customisation to specific requirements and applications [52]. However, extreme bias and limited quality of public reference databases pose substantial challenges when broad pathogen detection requirements necessitate comprehensive databases [69, 70]. In recent years, rapid read classification tools have been developed that reduce the need to limit the size of reference sequence databases. Analyses that took days or longer can now be performed within an hour or less [51, 60, 63]. In addition, to speed the ease of use, reliability and accuracy, independence of expertise of the user, and version control are other important features for data analysis tools to be used as part of clinical metagenomics workflows.

7.3.3.2 Organism Classification and Result Interpretation

Independent of the selected bioinformatic analysis tool, criteria for micro-organism classification and result interpretation must be defined. Important considerations include relative importance of sensitivity vs. specificity of pathogen detection, relevant micro-organism abundance (e.g. are low-positive results relevant?), composition and abundance of normal microbiota (e.g. do pathogens need to be differentiated from closely related commensals?, which ones?), expected biologic sequence diversity for relevant pathogens (e.g. RNA viruses), and prioritisation and interpretation of results (e.g. do certain commensals need to be excluded or high-impact pathogens be prioritised for reporting purposes?). In general, if the focus is on sensitivity, less stringent classification and interpretation criteria may be appropriate whereas applications that require high specificity will need to employ more stringent

classification and interpretation criteria. In addition, stringency may have to vary substantially between different taxa and require adjustment for given sequencing read lengths and sequencing error profiles. For example, classification of pathogens with divergent genomes (e.g. RNA viruses) may require laxer sequence comparison conditions (smaller k in k-*mer* based approaches, shorter seed length and higher tolerance for gaps and mismatches) or protein-level analyses (i.e. comparison of translated nucleotide query sequences against a protein or translated nucleotide database) to maximise sensitivity. While traditionally slow, these searches can now be performed at rapid speed [51, 58, 60, 63]. Final classification and interpretation criteria for a test as a whole or given micro-organism will impact test performance and should be acknowledged in the test information provided to clinicians.

7.3.3.3 Identifying Contamination

Sequencing artifacts (e.g. low-quality reads) and sequencing data representing contamination introduced during specimen collection or processing (e.g. reagent contamination) need to be anticipated, identified as such, and differentiated from relevant, specimen-derived sequences. Contamination may arise from reagents containing microbial DNA (e.g. due to environmental contamination, as part of recombinant enzymes, etc.), may be introduced during specimen collection, storage, or processing, mis-inoculation or impurities of barcode sequences, carry-over of within sequencing instruments, index hopping, and other mechanisms [44, 71, 72]. The use of ultra-pure reagents in well controlled molecular laboratory settings reduces but often cannot completely eliminate the risk of contamination. Therefore, carefully selected external (positive and negative) controls and internal (spike-in) controls are needed throughout the entire workflow to identify sequences not derived from the clinical specimen [73].

7.3.3.4 Result Interpretation

Some of the consideration for determining which detected micro-organisms should be included in a diagnostic report may include: (1) comparison of micro-organisms detected in patient specimens with those identified in external controls; (2) in shotgun metagenomic workflows, the detection level for a given micro-organism depends on the presence and abundance of other organisms and host nucleic acid; because those may differ between patient specimens and external controls, simply excluding micro-organisms found in external controls may not yield the optimal results; approaches that take the biomass and composition of the specimen into consideration have been developed [74, 75]; (3) as discussed above, *a priori* defining those organisms that are relevant for a given test and prioritising those for reporting may be beneficial; (4) adjusting confidence thresholds for reporting of organisms based on the intended use of the test, impact of a given detection, completeness of reference databases and/or genetic variability of relevant

micro-organisms; and (5) reporting only of those organisms that meet a validated minimum reporting detection thresholds. Thresholds may be based on a number of individual metrics or combinations of criteria including minimum total number of reads assigned to a given organism, establishing a minimum proportion of genes or genome that needs to be identified, minimum depth of coverage over a pre-determined region of the genome, and others. Thresholds may need to be customised for specific micro-organisms. In particular, taxa from dense parts of the phylogenetic tree (i.e. with genetically similar neighbours) may require particular attention. By tailoring detection and reporting criteria to individual micro-organisms, sensitivity and specificity can be maximised.

7.3.3.5 Approach to Test Validation

Ideally, validations would include clinical specimens with known results based on high-quality predicate tests, with known quantities, covering all relevant micro-organisms detectable by the sequencing test, in all relevant specimen matrices, combined with clinical specimens that contain micro-organisms that need to be differentiated from relevant organisms to avoid false-positive results. However, due to limited availability of well-characterised specimens, lack of a universal reference method, and the sheer scope of clinical metagenomics tests, this is generally not realistic. There is currently no consensus on how laboratories should strike a balance between sufficiently characterising test performance while using limited resources judiciously. Approaches often include a combination of positive and negative patient specimens (based on conventional tests), spiked patient specimens, reference materials (as individual positives or mock communities, with or without matrix), and *in silico* generated mock specimens (based on simulated micro-organism sequences with or without real or mock matrix sequences) [15, 28, 76]. Usually, positives at least for the most common pathogens and commensals can be sourced for the relevant specimen types. Mock specimens (laboratory spiked or *in silico* generated) can help assess performance for detection of clinically important but less widely available micro-organisms. Serial dilution studies (again, laboratory spiked or *in silico* generated) can be used to assess sensitivity while specificity can also be tested using negative patient specimens, blanks, and *in silico* generated specimens. As with other diagnostic tests, routine performance characteristics (accuracy, reproducibility, sensitivity, specificity, stability, etc.) should be considered.

Testing of *in silico* generated specimens enables assessment of a much larger number of relevant pathogens and commensals at low cost and with complete knowledge of the expected results. Sequencing data of the same size, read length, and error profile can be constructed computationally (*in silico*) and analysed with the diagnostic pipeline. As discussed above, sequencing data from real patient specimens often contains sequence artefacts and sequences that did not originate from the specimen. If using *in silico* generated specimens, this should be taken into consideration. Relevant sample composition can be recapitulated by generating

sequencing data from the host (human DNA sequences), common contaminants, and commensals [28]. A large number of metagenomics datasets are also available from public databases and may help avoid over-training when *in silico* data are generated from sequences contained in classification databases (i.e. perfect matches exist for simulated specimens) [77, 78]. *In silico* testing is especially important for validation of rare but clinically important pathogens, including emerging pathogens and biosafety level (BSL)-3 or BSL-4 agents that may not be practical to handle for spiking experiments. This approach can also be useful for studying closely related taxa that may be common but difficult to differentiate (e.g. *Streptococcus pneumoniae* and *S. mitis*) as specimen composition can be fully controlled, including their relative abundance.

7.3.4 Quality Management

Quality control and quality assurance must be implemented throughout the metagenomic testing process. All steps of testing, including pre-analytic, analytic, and post-analytic should be assessed through the laboratory's quality procedures [79]. There is no consensus yet on the specifics and extent of quality control measures. Some approaches are listed below.

7.3.4.1 Quality Control and Assessment

Analysis of specimen-level and run-level quality metrics is recommended throughout the specimen processing and data analysis workflow, including pre-analytic specimen checks, nucleic acid yield, assessment of library quantity and quality, evaluation of sequencing data quality and quantity for the entire run (including results for external controls) and for each specimen [28]. Sequencing error rate and base call quality are among the commonly used metrics to assess run performance. Pass/fail criteria should be defined to ensure high quality results without being overly stringent, resulting in unnecessary costs and delays. For positive control specimens, the expected identity and relative abundance of detectable organisms is known, and expected results need to be obtained. Negative controls can consist of matrix-matched or blank specimens and help identify contamination (see above). Matrix-matched controls can also identify problems that are dependent on specimen characteristics (e.g. viscosity, presence of inhibitors). Internal controls (e.g. whole micro-organisms also controlling for extraction, or synthetic nucleic acid) should be selected so that they can be readily differentiated from micro-organisms of interest and can be spiked into a master mix that is used for all specimens (e.g. lysis buffer) or be used as specimen-specific spike-in control with a unique sequence [28]. Depending on the specific strategy, internal controls can be used as processing controls, to monitor specimen composition, and identify specimen-to-specimen

contamination. The number of sequencing reads and/or sequence coverage of spike-in controls can also be used to assess specimen adequacy.

7.3.4.2 TAT

To be clinically actionable, results need to be reported in a timely manner. Longer TAT tests may have clinical utility for chronic infections. At least for short read platforms, sequencing library preparation and NGS contribute the most to the overall TAT. Often, host depletion or target enrichment steps can further increase processing times. When determining the need for automation, the rate of errors during sample processing, repeat rates, reproducibility, as well as impact on TAT should be taken into consideration. Time to result can also be highly variable for different sequencing platforms and throughput needs, ranging from less than an hour to multiple days [45]. Data analysis (even for diagnostic applications) can now be performed in well below an hour [51, 60, 63] and data analysis steps often do not significantly contribute to the overall TAT any more. Workflow management further impact TAT. To minimize TAT, organising workflows in at least two shifts may be required. Implementation of clear protocols including repeat algorithms and multiple pass/fail check points throughout testing and special considerations for specimens with short storage stability that may not support repeat testing is especially important for minimizing TAT during non-ideal testing situations.

7.4 Remaining Challenges for NGS in Clinical Diagnostics

Breakthroughs in specimen preparation, sequencing technology, and computational biology enabled introduction of the first clinical metagenomics tests at select reference and public health laboratories. Protocols and technologies evolve rapidly and implementing clinical metagenomics tests is becoming feasible for a growing number of laboratories. To further increase access, future workflow improvements will likely increase analytical sensitivity, reduce TAT and costs (both per sample costs and capital expense requirements), streamline test development. Clinical outcome and test utilisation studies are needed to establish guidelines for best ordering practices.

7.4.1 Sample Processing

An ongoing challenge for metagenomics-based testing is the fact that host nucleic acid and pathogen nucleic acid compete during library preparation and sequencing. Numerous methods for both pathogen target enrichment and host (nucleic acids or cell) depletion exist aiming at increasing sensitivity and decreasing the required

sequencing depth, and therefore cost [41, 42, 80]. However, most available methods have considerable limitations, requiring fresh specimens, high molecular weight nucleic acid, long incubation times, or off-target effects. For RNA-sequencing-based workflows, ribosomal RNA (rRNA) and globin depletion (for bloody specimens) are commercially available. In addition, greater ease-of-use and lower costs of customized depletion probes makes it feasible to also consider removal of other highly abundant transcripts. Similar technology has also made it possible to design target enrichment workflows that allow for broad pathogen detection [37, 81–84]. Potential cross-reactivity between host and pathogen sequences - that may be difficult to exclude or quantify - remains a challenge for hybridisation-based depletion methods. Another concern is loss of specimen nucleic acid and pathogen yield in additional processing steps. Commercially available depletion or enrichment methods are needed that reduce cost and workflow barriers for diagnostic laboratories and maximise analytical sensitivity of broad NGS-based pathogen detection tests.

Clean reagents that are free of contaminating nucleic acids and workflows that reduce the risk for environmental contamination are essential for molecular testing in general but problems are amplified for clinical metagenomics tests due to their broad scope [85, 86]. The impact of any improvements will be greatest on low biomass specimens that are most vulnerable to artifacts introduced by reagent and environmental contamination.

Complexity of metagenomics specimen preparation workflows provide a barrier for laboratories. Resources including laboratory space for unidirectional workflow, personnel training, and expertise for data analysis and interpretation have to anticipate and accommodate workflow complexities [87]. Future development will have to focus on simplifying workflows, minimising hands-on time, reducing expertise needed for post-sequencing steps, including quality control/quality assurance of metagenomics workflows. Many of these problems have been addressed in other areas of NGS testing already and lessons can be applied to clinical metagenomics, and the next years are likely to bring substantial improvements in ease-of-use and performance of metagenomics tests.

7.4.2 *Sequencing and Data Analysis*

In addition to user-friendly data analysis and reporting tools designed for use by clinical laboratories, the combination of fast (within approximately 4 h), reliable, and economical sequencing platforms will be essential for broad adoption of clinical metagenomics in clinical and public health laboratories. Decreased costs could also open NGS technology to a number of additional microbiology applications. For example, laboratories might consider more general use of whole genome sequencing for identification of clinical isolates.

Sequence data analysis, organism identification, and reporting will need to be further standardised [88]. Currently, most laboratories use customised analysis tools and criteria limiting reproducibility of results and external validity of published

studies [89]. Standardised data analysis will also reduce the effort needed for laboratories to develop metagenomics tests. NGS data analysis software should include user interfaces designed for laboratory staff (i.e. not requiring bioinformatics skills), reporting functionality, including interfacing of results with laboratory information systems, and support routine result review and release workflows [90].

7.4.3 Test Utilization

As with any new technology, optimal applications for clinical metagenomics need to be established. More clinical utility studies need to be performed with specific application, patient enrollment criteria, comprehensive predicate testing, defined specimen collection, preservation, and processing protocols, and clinical outcome data. Currently, the most common scenario for ordering clinical metagenomics tests is in critically ill patients in addition to standard diagnostic workup or after standard testing is unsuccessful. The use as a test of last resort has the disadvantages of further prolonging the time to diagnosis and limiting testing to patients with low pretest probability. In addition, current testing approaches often provide an incomplete picture of the potential pathogens detected. At least on some specimen types (e.g. respiratory specimens) identification of one potential pathogen does not exclude the possibility that additional, possibly more relevant pathogens may have gone undetected. Incorporating metagenomics tests earlier may benefit patients and reduce unnecessary testing but adequate patient selection criteria need to be defined. For example, in patients with risk factors or clinical presentations that lead to a long list of differential diagnoses, broad pathogen detection with a single test early on could shorten the time to diagnosis and reduce costs for unnecessary testing and inadequate treatment. Another application is specimens that usually have very limited volume available but require testing for a number of organisms (e.g. vitreous or intraocular fluid). Limited specimen volume may allow clinicians and laboratories to perform only a few pathogen-specific tests. Being able to test for a much larger number of potential pathogens with a single test provides an advantage to metagenomics tests [91]. Further clinical studies are required to identify high yield testing situations with positive clinical impact.

7.4.4 Incidental Findings

One potential consequence of untargeted testing is the inadvertent detection of host genomic variants, unexpected pathogens (e.g. sexually transmitted infections), or non-validated micro-organisms with confident detection and clear clinical significance. Thus, the question "should the additional information be disclosed to the patient?" becomes relevant. To avoid incidental detection of host genomic variants, human sequence data can be removed or not analysed further [92] and patient

privacy considerations or requirements may dictate methods for storing and processing data [93]. The possibility of generating incidental findings requires balancing best clinical care with patient privacy [94–96]. The American College of Medical Genetics and Genomics (ACMG) has published recommendations for reporting of specific conditions, genes, or variants when discovered incidentally [97]. Similar guidelines have not been published yet for incidental results generated by metagenomics tests.

7.5 Conclusions

The development of metagenomic tests for pathogen detection has the potential to change the face of laboratory testing for infectious diseases. Published cases and early clinical studies demonstrate the promise of detecting unexpected, uncommon, slow growing, co-infecting pathogens in difficult-to-diagnose patients [12–15]. This technology can be particularly useful for diagnosis of rare micro-organisms for which there is a lack of available clinical tests and detection of uncommon variants of common pathogens [26, 98]. The untargeted nature of testing enables broad pathogen detection from a single, low-volume specimen, which is especially important for testing in children, precious specimens (e.g. intraocular fluid, CSF), or those that are difficult to recollect (e.g. specimens collected before initiation of antimicrobial therapy). In addition to clinical diagnoses, metagenomics also has many important applications in public health testing and infection control (e.g. strain typing, profiling for molecular resistance determinants, or surveillance).

Understanding and defining appropriate clinical indications for metagenomics testing remains a challenge and clinical utility studies will be needed. Conducting those studies and continuously improving metagenomics tests will require a multidisciplinary approach, involving clinical, laboratory, computational biology, and data science teams. Because of the heavy dependence on sequencing and data analysis technologies, collaborations between laboratory experts and test developers will also be required. Analytic phase improvements include optimisation of wet-bench methods, sequencing technology, and data analysis procedures. Result analysis and reporting can be improved to better assist clinicians in interpretation of results. Test development and validation will likely continue to provide challenges to laboratories until methods are more standardised and guidance documents become available. In their absence, laboratories will have to use judgment, a risk-based approach, and consider a combination of the different validation strategies outlined above. Microbiology test results are generally reported as “detected” or “not detected”. Given the vast quantity and resolution of data acquired by metagenomic approaches, the laboratory has the opportunity to provide additional, clinically relevant information to assist result interpretation. Reporting may include not only an micro-organism name, but the quantity at which it was detected, genotypic information, genetic markers of drug resistance, and even gene expression activities of detected pathogens. By their sheer breadth, metagenomics tests also require a

paradigm that relies less on extensive expertise in a certain class of pathogens as the same workflow will produce results across all categories of pathogens.

References

1. Muldrew K (2009) Molecular diagnostics of infectious diseases. Curr Opin Pediatr 21(1):102–111
2. Tan KE, Ellis BC, Lee R, Stamper PD, Zhang SX, Carroll KC (2012) Prospective evaluation of a matrix-assisted laser desorption ionization-time of flight mass spectrometry system in a hospital clinical microbiology laboratory for identification of bacteria and yeasts: a bench-by-bench study for assessing the impact on time to identification and cost-effectiveness. J Clin Microbiol 50(10):3301–3308
3. Banerjee R, Teng CB, Cunningham SA, Ihde SM, Steckelberg JM, Moriarty JP et al (2015) Randomized trial of rapid multiplex polymerase chain reaction-based blood culture identification and susceptibility testing. Clin Infect Dis 61(7):1071–1080
4. Pavia AT (2011) Viral infections of the lower respiratory tract: old viruses, new viruses, and the role of diagnosis. Clin Infect Dis 52(Suppl 4):S284–S289
5. Glaser CA, Honarmand S, Anderson LJ, Schnurr DP, Forghani B, Cossen CK et al (2006) Beyond viruses: clinical profiles and etiologies associated with encephalitis. Clin Infect Dis 43(12):1565–1577
6. Choi SH, Hong SB, Ko GB, Lee Y, Park HJ, Park SY et al (2012) Viral infection in patients with severe pneumonia requiring intensive care unit admission. Am J Respir Crit Care Med 186(4):325–332
7. Jain S, Self WH, Wunderink RG, Fakhran S, Balk R, Bramley AM et al (2015) Community-acquired pneumonia requiring hospitalization among U.S. adults. N Engl J Med 373(5):415–427
8. Jain S, Williams DJ, Arnold SR, Ampofo K, Bramley AM, Reed C et al (2015) Community-acquired pneumonia requiring hospitalization among U.S. children. N Engl J Med 372(9):835–845
9. Self WH, Williams DJ, Zhu Y, Ampofo K, Pavia AT, Chappell JD et al (2016) Respiratory viral detection in children and adults: comparing asymptomatic controls and patients with community-acquired pneumonia. J Infect Dis 213(4):584–591
10. Tarabichi M, Shohat N, Goswami K, Alvand A, Silibovsky R, Belden K et al (2018) Diagnosis of Periprosthetic joint infection: the potential of next-generation sequencing. J Bone Joint Surg Am 100(2):147–154
11. Gallo J, Kolar M, Dendis M, Loveckova Y, Sauer P, Zapletalova J et al (2008) Culture and PCR analysis of joint fluid in the diagnosis of prosthetic joint infection. New Microbiol 31(1):97–104
12. Hong DK, Blauwkamp TA, Kertesz M, Bercovici S, Truong C, Banaei N (2018) Liquid biopsy for infectious diseases: sequencing of cell-free plasma to detect pathogen DNA in patients with invasive fungal disease. Diagn Microbiol Infect Dis 92(3):210–213
13. Farnaes L, Wilke J, Ryan Loker K, Bradley JS, Cannavino CR, Hong DK et al (2019) Community-acquired pneumonia in children: cell-free plasma sequencing for diagnosis and management. Diagn Microbiol Infect Dis 94(2):188–191
14. Hogan CA, Yang S, Garner OB, Green DA, Gomez CA, Dien Bard J et al (2020) Clinical impact of metagenomic next-generation sequencing of plasma cell-free DNA for the diagnosis of infectious diseases: a multicenter retrospective cohort study. Clin Infect Dis. https://doi.org/10.1093/cid/ciaa035
15. Blauwkamp TA, Thair S, Rosen MJ, Blair L, Lindner MS, Vilfan ID et al (2019) Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. Nat Microbiol 4(4):663–674

16. Taravati P, Lam D, Van Gelder RN (2013) Role of molecular diagnostics in ocular microbiology. Curr Ophthalmol Rep 1(4). https://doi.org/10.1007/s40135-013-0025-1
17. Doan T, Pinsky BA (2016) Current and future molecular diagnostics for ocular infectious diseases. Curr Opin Ophthalmol 27(6):561–567
18. Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, Garcia-Cobos S et al (2017) Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol 243:16–24
19. Dunne WM Jr, Jaillard M, Rochas O, Van Belkum A (2017) Microbial genomics and antimicrobial susceptibility testing. Expert Rev Mol Diagn 17(3):257–269
20. Kashyap PC, Chia N, Nelson H, Segal E, Elinav E (2017) Microbiome at the frontier of personalized medicine. Mayo Clin Proc 92(12):1855–1864
21. Koser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M et al (2012) Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog 8(8):e1002824
22. Motro Y, Moran-Gilad J (2017) Next-generation sequencing applications in clinical bacteriology. Biomol Detect Quantif 14:1–6
23. Mongkolrattanothai K, Naccache SN, Bender JM, Samayoa E, Pham E, Yu G et al (2017) Neurobrucellosis: unexpected answer from metagenomic next-generation sequencing. J Pediatric Infect Dis Soc 6(4):393–398
24. Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G et al (2014) Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med 370(25):2408–2417
25. Wilson MR, Suan D, Duggins A, Schubert RD, Khan LM, Sample HA et al (2017) A novel cause of chronic viral meningoencephalitis: Cache Valley virus. Ann Neurol 82(1):105–114
26. Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J et al (2019) Clinical metagenomic sequencing for diagnosis of meningitis and encephalitis. N Engl J Med 380(24):2327–2340
27. Long Y, Zhang Y, Gong Y, Sun R, Su L, Lin X et al (2016) Diagnosis of Sepsis with cell-free DNA by next-generation sequencing technology in ICU patients. Arch Med Res 47(5):365–371
28. Schlaberg R, Chiu CY, Miller S, Procop GW, Weinstock G, Professional Practice C et al (2017) Validation of metagenomic next-generation sequencing tests for universal pathogen detection. Arch Pathol Lab Med 141(6):776–786
29. Grard G, Fair JN, Lee D, Slikas E, Steffen I, Muyembe JJ et al (2012) A novel rhabdovirus associated with acute hemorrhagic fever in Central Africa. PLoS Pathog 8(9):e1002924
30. Palacios G, Druce J, Du L, Tran T, Birch C, Briese T et al (2008) A new arenavirus in a cluster of fatal transplant-associated diseases. N Engl J Med 358(10):991–998
31. Xu B, Liu L, Huang X, Ma H, Zhang Y, Du Y et al (2011) Metagenomic analysis of fever, thrombocytopenia and leukopenia syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus. PLoS Pathog 7(11):e1002369
32. Briese T, Paweska JT, McMullan LK, Hutchison SK, Street C, Palacios G et al (2009) Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. PLoS Pathog 5(5):e1000455
33. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL et al (2013) ACMG clinical laboratory standards for next-generation sequencing. Genet Med 15(9):733–747
34. Ali N, Rampazzo RCP, Costa ADT, Krieger MA (2017) Current nucleic acid extraction methods and their implications to point-of-care diagnostics. Biomed Res Int 2017:9306564
35. Lefterova MI, Suarez CJ, Banaei N, Pinsky BA (2015) Next-generation sequencing for infectious disease diagnosis and management: a report of the Association for Molecular Pathology. J Mol Diagn 17(6):623–634
36. Deng X, Achari A, Federman S, Yu G, Somasekar S, Bartolo I et al (2020) Metagenomic sequencing with spiked primer enrichment for viral diagnostics and genomic surveillance. Nat Microbiol 5(3):443–454
37. Wylie TN, Wylie KM, Herter BN, Storch GA (2015) Enhanced virome sequencing using targeted sequence capture. Genome Res 25(12):1910–1920

38. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M et al (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res 41(1):e1
39. Kim KH, Bae JW (2011) Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. Appl Environ Microbiol 77(21):7663–7668
40. Rosseel T, Van Borm S, Vandenbussche F, Hoffmann B, van den Berg T, Beer M et al (2013) The origin of biased sequence depth in sequence-independent nucleic acid amplification and optimization for efficient massive parallel sequencing. PLoS One 8(9):e76144
41. Hasan MR, Rawat A, Tang P, Jithesh PV, Thomas E, Tan R et al (2016) Depletion of human DNA in spiked clinical specimens for improvement of sensitivity of pathogen detection by next-generation sequencing. J Clin Microbiol 54(4):919–927
42. Allander T, Emerson SU, Engle RE, Purcell RH, Bukh J (2001) A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. Proc Natl Acad Sci U S A 98(20):11609–11614
43. Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L et al (2018) Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics 19(1):332
44. MacConaill LE, Burns RT, Nag A, Coleman HA, Slevin MK, Giorda K et al (2018) Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. BMC Genomics 19(1):30
45. Goldberg B, Sichtig H, Geyer C, Ledeboer N, Weinstock GM (2015) Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. MBio 6(6):e01888–e01815
46. Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17(6):333–351
47. Leggett RM, Alcon-Giner C, Heavens D, Caim S, Brook T, Kujawska M et al (2017) Rapid MinION metagenomic profiling of the preterm infant gut microbiota to aid in pathogen diagnostics. bioRxiv
48. Cheval J, Sauvage V, Frangeul L, Dacheux L, Guigon G, Dumey N et al (2011) Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. J Clin Microbiol 49(9):3268–3275
49. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V et al (2015) Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med 7:99
50. Breitwieser FP, Lu J, Salzberg SL (2019) A review of methods and databases for metagenomic classification and assembly. Brief Bioinform 20(4):1125–1136
51. Lindgreen S, Adair KL, Gardner PP (2016) An evaluation of the accuracy and speed of metagenome analysis tools. Sci Rep 6:19233
52. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Henaff E, Alexander N et al (2017) Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol 18(1):182
53. Nooij S, Schmitz D, Vennema H, Kroneman A, Koopmans MPG (2018) Overview of virus metagenomic classification methods and their biological applications. Front Microbiol 9:749
54. Peabody MA, Van Rossum T, Lo R, Brinkman FS (2015) Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. BMC Bioinf 16:363
55. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N (2017) Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35(9):833–844
56. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Droge J et al (2017) Critical assessment of metagenome interpretation-a benchmark of metagenomics software. Nat Methods 14(11):1063–1071
57. Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE (2013) Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics 29(18):2253–2260

58. Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12(1):59–60
59. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7(5):335–336
60. Flygare S, Simmon K, Miller C, Qiao Y, Kennedy B, Di Sera T et al (2016) Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. Genome Biol 17(1):111
61. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17(3):377–386
62. Huson DH, Beier S, Flade I, Gorska A, El-Hadidi M, Mitra S et al (2016) MEGAN Community edition - interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comput Biol 12(6):e1004957
63. Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E et al (2014) A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res 24(7):1180–1192
64. Ounit R, Wanamaker S, Close TJ, Lonardi S (2015) CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics 16:236
65. Rosen G, Garbarine E, Caseiro D, Polikar R, Sokhansanj B (2008) Metagenome fragment classification using N-mer frequency profiles. Adv Bioinforma 2008:205969
66. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR et al (2013) Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods 10(12):1196–1199
67. Leung HC, Yiu SM, Yang B, Peng Y, Wang Y, Liu Z et al (2011) A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. Bioinformatics 27(11):1489–1495
68. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9(8):811–814
69. Breitwieser FP, Pertea M, Zimin AV, Salzberg SL (2019) Human contamination in bacterial genomes has created thousands of spurious proteins. Genome Res 29(6):954–960
70. Edgar R (2018) Taxonomy annotation and guide tree errors in 16S rRNA databases. PeerJ 6:e5030
71. D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC et al (2016) A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics 17:55
72. Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J (2014) Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. PLoS One 9(4):e94249
73. Lusk RW (2014) Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. PLoS One 9(10):e110808
74. Zinter MS, Mayday MY, Ryckman KK, Jelliffe-Pawlowski LL, DeRisi JL (2019) Towards precision quantification of contamination in metagenomic sequencing experiments. Microbiome 7(1):62
75. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS (2019) Contamination in low microbial biomass microbiome studies: issues and recommendations. Trends Microbiol 27(2):105–117
76. Miller S, Naccache SN, Samayoa E, Messacar K, Arevalo S, Federman S et al (2019) Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. Genome Res 29(5):831–842
77. Bokulich NA, Rideout JR, Mercurio WG, Shiffer A, Wolfe B, Maurice CF et al (2016) Mockrobiota: a public resource for microbiome bioinformatics benchmarking. mSystems 1(5):e00062–16

78. Sichtig H, Minogue T, Yan Y, Stefan C, Hall A, Tallon L et al (2019) FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science. Nat Commun 10(1):3313
79. Aziz N, Zhao Q, Bry L, Driscoll DK, Funke B, Gibson JS et al (2015) College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. Arch Pathol Lab Med 139(4):481–493
80. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H et al (2016) Depletion of abundant sequences by hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. Genome Biol 17:41
81. Paskey AC, Frey KG, Schroth G, Gross S, Hamilton T, Bishop-Lilly KA (2019) Enrichment post-library preparation enhances the sensitivity of high-throughput sequencing-based detection and characterization of viruses from complex samples. BMC Genomics 20(1):155
82. Wylie KM, Wylie TN, Buller R, Herter B, Cannella MT, Storch GA (2018) Detection of viruses in clinical samples by use of metagenomic sequencing and targeted sequence capture. J Clin Microbiol 56(12): e01123–18
83. O'Flaherty BM, Li Y, Tao Y, Paden CR, Queen K, Zhang J et al (2018) Comprehensive viral enrichment enables sensitive respiratory virus genomic identification and analysis by next generation sequencing. Genome Res 28(6):869–877
84. Briese T, Kapoor A, Mishra N, Jain K, Kumar A, Jabado OJ et al (2015) Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. mBio 6(5):e01491–e01415
85. Chiu CY, Miller SA (2019) Clinical metagenomics. Nat Rev Genet 20(6):341–355
86. Strong MJ, Xu G, Morici L, Splinter Bon-Durant S, Baddoo M, Lin Z et al (2014) Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. PLoS Pathog 10(11):e1004437
87. Miller S, Chiu C, Rodino K, Miller M. (2020) Point-Counterpoint: Should We be Performing Metagenomic Next-Generation Sequencing for Infectious Disease Diagnosis in the Clinical Laboratory? J Clin Microbiol 58(3):e01739–19
88. Mitchell SL, Simner PJ (2019) Next-generation sequencing in clinical microbiology: are we there yet? Clin Lab Med 39(3):405–418
89. White DJ, Wang J, Hall RJ (2017) Assessing the impact of assemblers on virus detection in a De Novo metagenomic analysis pipeline. J Comput Biol 24(9):874–881
90. Allcock RJN, Jennison AV, Warrilow D (2017) Towards a universal molecular microbiological test. J Clin Microbiol 55(11):3175–3182
91. Doan T, Acharya NR, Pinsky BA, Sahoo MK, Chow ED, Banaei N et al (2017) Metagenomic DNA sequencing for the diagnosis of intraocular infections. Ophthalmology 124(8):1247–1248
92. Hall RJ, Draper JL, Nielsen FGG, Dutilh BE (2015) Beyond research: a primer for considerations on using viral metagenomics in the field and clinic. Front Microbiol 6:224
93. Chiu C, Miller S (2016) Molecular microbiology: diagnostic principles and Practice. Persing DH, editor. ASM Press, Washington, DC
94. Rahimzadeh V, Avard D, Senecal K, Knoppers BM, Sinnett D (2015) To disclose, or not to disclose? Context matters. Eur J Hum Genet 23:279–284
95. Clarke AJ (2014) Managing the ethical challenges of next-generation sequencing in genomic medicine. Br Med Bull 111(1):17–30
96. Davey S (2014) Next generation sequencing: considering the ethics. Int J Immunogenet 41(6):457–462
97. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL et al (2013) ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med 15(7):565–574
98. Graf EH, Simmon KE, Tardif KD, Hymas W, Flygare S, Eilbeck K et al (2016) Unbiased detection of respiratory viruses by use of RNA sequencing-based metagenomics: a systematic comparison to a commercial PCR panel. J Clin Microbiol 54(4):1000–1007