Title

STROBE-metagenomics: a STROBE extension statement to guide the reporting of metagenomics studies

Journal

The Lancet Infectious Diseases, 20(PLoS One 8 2013)

ISSN

1473-3099

Authors

Bharucha, Tehmina
Oeser, Clarissa
Balloux, Francois
et al.

Publication Date

2020-10-01

DOI

10.1016/s1473-3099(20)30199-7

Peer reviewed

Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

STROBE-metagenomics: a STROBE extension statement to guide the reporting of metagenomics studies

Tehmina Bharucha, Clarissa Oeser, Francois Balloux, Julianne R Brown, Ellen C Carbo, Andre Charlett, Charles Y Chiu, Eric C J Claas, Marcus C de Goffau, Jutte J C de Vries, Marc Eloit, Susan Hopkins, Jim F Huggett, Duncan MacCannell, Sofia Morfopoulou, Avindra Nath, Denise M O'Sullivan, Lauren B Reoma, Liam P Shaw, Igor Sidorov, Patricia J Simner, Le Van Tan, Emma C Thomson, Lucy van Dorp, Michael R Wilson, Judith Breuer, Nigel Field

The term metagenomics refers to the use of sequencing methods to simultaneously identify genomic material from all organisms present in a sample, with the advantage of greater taxonomic resolution than culture or other methods. Applications include pathogen detection and discovery, species characterisation, antimicrobial resistance detection, virulence profiling, and study of the microbiome and microecological factors affecting health. However, metagenomics involves complex and multistep processes and there are important technical and methodological challenges that require careful consideration to support valid inference. We co-ordinated a multidisciplinary, international expert group to establish reporting guidelines that address specimen processing, nucleic acid extraction, sequencing platforms, bioinformatics considerations, quality assurance, limits of detection, power and sample size, confirmatory testing, causality criteria, cost, and ethical issues. The guidance recognises that metagenomics research requires pragmatism and caution in interpretation, and that this field is rapidly evolving.

Background

The term metagenome was coined in 1998 to describe the collection of genomes from microbes present in environmental soil samples by using approaches previously used to study single genomes.[1] The sequencing of genetic material from clinical samples has become common practice in research on clinical microorganisms. In this context, metagenomics refers to the application of sequencing methods that can identify coexistent genomic material from any organism present in patient samples (ie, microorganism and host nucleic acid), usually with the aim of pathogen identification for clinical diagnosis or research.[2–4] Examples of practical applications include pathogen detection and discovery, species characterisation or subtyping, antimicrobial resistance detection, virulence profiling, and studies of the microbiome and microecological drivers of health and disease.[5–12] Metagenomics is also being introduced as a diagnostic tool for causal studies of clinical syndromes (such as encephalitis),[13,14] for exploring the microbiome,[15,16] and for tracking disease outbreaks.[17,18] A current example of the transformational effect of direct sequencing of clinical samples has been the application for rapid investigation and dissemination of information on severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes COVID-19.[11,12]

Metagenomics data are generated using high-throughput sequencing methods, also referred to as deep, next-generation, massively parallel, or shotgun sequencing. In this Review, for simplicity, we refer to all these approaches as sequencing. We also include capture probe enrichment-based sequencing methods that use nucleotide probes to increase sensitivity[4] and targeted amplicon sequencing—eg, sequencing the 16S ribosomal ribonucleic acid (rRNA) gene to identify bacteria.[19] Capture probe enrichment-based sequencing and targeted amplicon sequencing might not be considered true examples of metagenomics and are not the focus of our Review; however, some similar considerations about reporting of results apply.

Metagenomic sequencing has advantages for pathogen identification over conventional methods, such as culture or targeted PCR, because many or most microbial species present within a sample can be detected simultaneously with high taxonomic resolution. Detailed characterisation of microbial communities and population dynamics also enables the study of ecological interactions. Furthermore, this method does not require culture techniques, and

Key messages

- The term metagenomics refers to the use of sequencing methods to simultaneously identify genomic material from all organisms present in a sample, with the advantage of greater taxonomic resolution than culture or other methods.
- Applications include pathogen detection and discovery, species characterisation, antimicrobial resistance detection, virulence profiling, and study of the microbiome and microecological factors affecting health.
- Metagenomics involves complex and multistep processes and there are important technical and methodological challenges that require careful consideration to support valid inference.
- We co-ordinated a multidisciplinary, international expert group to establish reporting guidelines that address specimen processing, nucleic acid extraction, sequencing platforms, bioinformatics considerations, quality assurance, limits of detection, power and sample size, confirmatory testing, causality criteria, cost, and ethical issues.
- The guidance recognises that metagenomics research requires pragmatism and caution in interpretation, and that this field is rapidly evolving. Reporting standards should support clarity, consistency, and robustness of research.

Lancet Infect Dis 2020; 20: e251–60

Published Online August 5, 2020 https://doi.org/10.1016/S1473-3099(20)30199-7

This online publication has been corrected. The corrected version first appeared at thelancet.com/infection on October 23, 2020

Department of Biochemistry (T Bharucha MRCP) **and Nuffield Department of Medicine** (L P Shaw PhD), **University of Oxford, Oxford, UK; Lao-Oxford-Mahosot Hospital-Wellcome Trust Research Unit, Microbiology Laboratory, Mahosot Hospital, Vientiane, Laos** (T Bharucha); **Centre for Molecular Epidemiology and Translational Research** (C Oeser PhD, N Field PhD) **and UCL Genetics Institute** (F Balloux PhD, L van Dorp PhD) **and Division of Infection and Immunity** (S Morfopoulou PhD, Prof J Breuer MD), **University College London, London, UK; Microbiology, Virology and Infection Prevention and Control** (J R Brown PhD), **Great Ormond Street Hospital for Children, London, UK** (J Breuer); **Department of Medical Microbiology, Leiden University Medical Center, Leiden, Netherlands** (E C Carbo MSc, E C H Claas PhD, J J C de Vries PhD, I Sidorov PhD); **Statistics, Modelling and Economics Department, Public Health England, London, UK** (A Charlett PhD); **Department of Laboratory Medicine, University of California San Francisco, San Francisco, CA, USA** (Prof C Chiu PhD); **Wellcome Sanger Institute, Hinxton, UK** (M C de Goffau PhD); **Department of Veterinary Medicine, University of Cambridge, Cambridge, UK** (M C de Goffau); **Pathogen Discovery Laboratory, Institut Pasteur, Paris, France** (Prof M Eloit PhD);

Healthcare-Associated Infection and Antimicrobial Resistance, Public Health England, London, UK (S Hopkins PhD)**; Infectious Diseases Unit, Royal Free Hospital, London, UK** (S Hopkins)**; National Measurement Laboratory, LGC, Teddington, UK** (J F Hugget PhD, D M O'Sullivan PhD)**; School of Biosciences & Medicine, Faculty of Health & Medical Sciences, University of Surrey, Guildford, UK** (J F Hugget)**; Office of Advanced Molecular Detection, Centers for Disease Control and Prevention, Atlanta, GA, USA** (D MacCannell PhD)**; Section of Infections of the Nervous System, National Institutes of Health, Bethesda, MD, USA** (Prof A Nath MD, L B Reoma MD)**; Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, USA** (P J Simner PhD)**; Emerging Infections Group, Oxford University Clinical Research Unit, Ho Chi Minh city, Vietnam** (Le Van Tan PhD)**; MRC-University of Glasgow Centre for Virus Research, University of Glasgow, Glasgow, UK** (Prof E C Thomson FRCP)**; and Weill Institute for Neurosciences and Department of Neurology, University of California, San Francisco, CA, USA** (M R Wilson MD)

Correspondence to: Dr Tehmina Bharucha, Institute of Glycobiology, Department of Biochemistry, University of Oxford, Oxford, OX1 3RQ, UK
t.bharucha@doctors.org.uk

For more on **protocol sharing** see http://www.protocols.io/

therefore can be used for microbial species that are difficult or time consuming to grow. This is particularly relevant for diagnostic applications, where routine culture is in decline.[20,21]

However, appropriate study design for metagenomics research is not well defined and metagenomic technologies pose important technical challenges. These challenges include methodological artefacts introduced by wet laboratory methods and the effect that different computational approaches have on the analysis of multivariate and complex data. Furthermore, the ethical implications of sequencing are substantial and data privacy considerations are increasingly recognised. The multiple steps and different expertise required to generate and analyse metagenomic sequence data involves numerous decision points, which could introduce bias and affect downstream inference about the presence and abundance of microbial species in the sample.

A metagenome result should therefore be interpreted as one of many possible representations of the true sample composition of a given microbiome. Understanding and reporting sources of bias and limitations to valid inference should improve protocol performance and enable metagenomic research to proceed with transparent recognition of the limitations. However, existing reporting statements for epidemiology studies, including STROBE (STrengthening the Reporting of OBservational studies in Epidemiology)[22] and its infectious disease molecular epidemiology extension, STROME-ID (Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases),[23] do not fully address issues specific to metagenomics. For this reason, scientific journals, and their readers, might not be adequately equipped with a standardised set of guidelines to evaluate and critically appraise clinical and epidemiological studies applying metagenomics. We aimed to improve the clarity and consistency of metagenomics research reporting, ranging from clinical diagnostics to microbiome studies, with suggestions for optimal practice and recommendations for robust and accurate reporting.

Titles and abstracts

The term metagenomics should be included in the title or abstract, and the keywords of the study when these methods contribute substaintially to the results reported

Clear and concise language incorporating standardised terminology, with references if appropriate, enables the accurate indexing of published studies in recognised databases. This is crucial for easy information retrieval and knowledge dissemination. For example, a systematic literature review of studies applying metagenomics in encephalitis using medical subject headings and keyword searches for the terms sequencing or metagenomics in four databases (PubMed, Embase, Web of Science, and Cochrane)[13] failed to identify two relevant studies that did not report the terms.[25,26] These studies were identified by experts in the field who were directly involved with the studies.

Describing methods and study design

Describe specimen collection, handling and storage processes, and nucleic acid extraction methods

Steps involved in sample collection, handling, and processing are frequently poorly reported in publications and yet they will have considerable effect on the results and reproducibility of a study and could introduce variability artefacts.[27–30] In particular, many studies use material banked and collected originally for other purposes. In this Review, we describe important potential sources of error and their contribution to bias.

Nucleic acids, particularly RNA, are labile. Consequently, the collection methods, addition of nucleic acid stabilisers, and time to processing can affect the results obtained.[31] To address these issues, reporting should include durations, volumes, temperatures, and methods used before, during, and after the storage of samples.[32,33] Extraction methods contribute to another major source of method-induced variation—eg, by being DNA or RNA specific, or tailored to specific organism types—so should be described.[34] Other details of sample preparation methods should also be reported including filtration, centrifugation, DNA digestion, rRNA depletion, separation in RNA or DNA, and random amplification. Standardised protocols of sample preparation methods should also be followed, if available and appropriate, and documented clearly in the publication methods. Authors should also consider submitting to standardised protocol repositories to provide transparency in the study design and methodology.

Describe sequencing methods, including sequencing depth

Different metagenomic sequencing platforms might produce different types of reads—eg, single versus paired-end, and short (100–300 bp) versus long (>1000 bp). Sequencing platforms have different error rates, with the probability of a nucleic base being read incorrectly ranging from less than 0·01% for Illumina sequencers to 5–10% for Oxford Nanopore Technologies sequencers (current figures as of February, 2020).[35] Additionally, sequencers often read a base incorrectly when processing samples with large homopolymer repeats, GC-rich, structurally repetitive, and other complex regions of the genome. Consequent false-positive and false-negative errors need consideration when reporting species composition.[36]

Sequencing depth refers to the number of times a particular nucleic base is represented within reads or the redundancy of coverage,[37] and has implications for identification of low abundant transcripts and confidence in sequencing data. However, sequencing depth must be balanced according to the research question and the available resources. There are several factors that affect sequencing depth, including the sequencing platform

and the sequence that is being read (eg, species diversity of the sample).[37–39]

Describe methods used for bioinformatics analysis

For the purposes of this statement, the term bioinformatics applies to all analysis steps involving raw sequencing data, including base calling, de-multiplexing, trimming and removal of reads (eg, reads of low quality, low complexity, adapters and indexes, or of human origin), read normalisation, alignment of sequence reads to reference databases, de-novo assembling of genomes, and taxonomic assignment of reads, assembled contigs, or both. There are multiple viable options for many of these tasks, with ongoing debate in the community about optimal methods, which can depend on the scientific question at hand. The field of metagenomics is developing rapidly and methods once considered best practice can be superseded following new analytical advances.

There should be clear descriptions of the bioinformatics methods used, including, at a minimum, the software name, version, and the main commands run with values for the essential parameters or flags. It is also advisable to make data and programming code open access, whether as supplementary files or shared online—eg, via Github or Figshare. Where possible, a version-controlled container, package, or easily installable version of the complete analytical pipeline (including all dependencies and required databases) could be made available for download and review. The open source release of bioinformatics workflows should be encouraged wherever possible to improve transparency and reproducibility, and should include adequate validation datasets, meaningful documentation, and examples of expected outputs and reports (appendix pp 1–2).

Describe quality assurance methods, including internal and external quality controls

An important strength of metagenomics analyses is their ability to detect any genomic material present within one sample. However, detection applies equally to true sample material and to any contaminating nucleic acids present in a sample, which can be introduced at any stage from sample collection to processing. For example, contamination could come from the extraction kit, the so-called kitome,[40] or at the point of specimen collection. Sampling is rarely done under completely sterile conditions, and tissues obtained from tissue banks are therefore often contaminated. Low biomass and low abundance sites (for example tumours, the brain, and fetal tissues such as the placenta) are particularly prone to the risk of misclassifying contaminants.

To show attempts to ensure internal validity and reproducibility and identify potential contamination, internal controls for all extraction and sequencing processes should be reported as part of standard operating procedures.[4,27] Positive controls are usually spiked with DNA or RNA—eg, synthetic nucleic acid standards such as sequins[47]—and negative controls are usually a blank (eg, water) sample or ideally a similar or identical matrix (tissue, body fluids, etc) that are expected to contain no microorganism genomic material based on patient factors and test results. For clinical metagenomics, formal laboratory implementation involves a system of external controls. Arranging this system of external controls is difficult; however, publicly and commercially available controls and mock community samples are now available and we recommend that their use should be reported.[48,49]

Describe use of orthogonal methods to confirm pathogen identity, function, and viability

The conventional methods in microbiology for confirming the presence of a pathogen are culture or growth of the pathogen from a clinical sample and immunohistochemistry, the histological localisation of candidate species in tissue biopsies. However, traditional culture can be difficult when antibiotics have been administered before sampling or for pathogens that are slow growing, fastidious, present in low-concentration, or currently undescribed. Sequencing has high discriminative power and could have higher sensitivity than culture-based methods. For example, in a polymicrobial sample, growth can be affected by presence of other competing bacteria or by inadequate growth conditions. Metagenomics methods have consistently shown higher classification accuracy when comparing taxonomic profiles of synthetic polymicrobial samples obtained from extended quantitative culture with non-selective media.[50]

Confirmatory assays appropriate to the study setting, justification for the methods used, and a description of their limitations should be reported. For cases in which confirmatory assays are not possible (eg, because of high cost or low volume of samples) an explanation should be provided. Rigorous validation of the method used, particularly for pathogens and proficiency testing, especially in clinical laboratories should be described (appendix pp 2–3).

See Online for appendix

Describe the criteria used to assess the role of pathogens in disease aetiology

Confirming the presence of microbial DNA or RNA in association with disease is an important step in establishing a causal relationship between a microorganism and disease.[51,52] A major challenge for metagenomics research and diagnostics is distinguishing pathogens from commensals or contaminants.[53,54] Interpretation of microbiome investigations can be further complicated if a misbalance in variation and abundance of different bacteria—sometimes referred to as dysbiosis—is suspected to be the cause of the condition.[55] It is also worth considering that the cause of some diseases might involve multiple sequential or interacting species, which can be collectively important.[56,57] Furthermore, sequencing investigations can identify novel organisms, for which the clinical significance will be unknown.

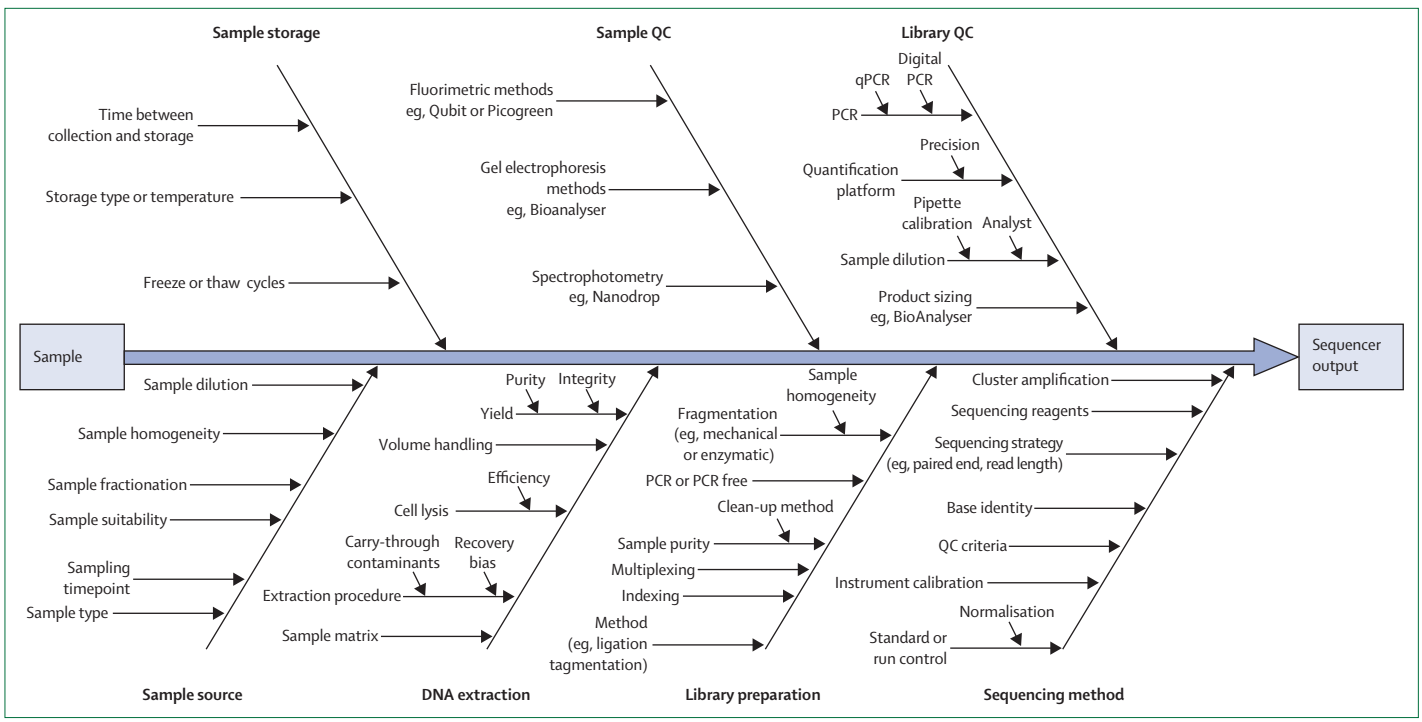

Figure 1: **Sources of uncertainty diagram highlighting potential contributing sources**
For simplicity, this figure considers the sequencing of DNA from an environment and does not consider the process beyond the data output from the sequencer. The arrows pointing towards the central black arrow show the experimental process from left to right and the sources of variability that could contribute uncertainty. Conceptually it is clear how some of these factors contribute to systematic effects (bias). However, in addition these factors also contribute to the random error (variance) that will influence the precision of a potential finding. QC=quality control.

These issues are particularly relevant in the investigation of the cause of CNS infections.

Several criteria to establish causality have been proposed over the past century, including the incorporation of metagenomic technologies (appendix 7–9).[58,59]

State the time from collection to results and cost consideration

The time from sample collection to processing (transport time), including cold-chain transportation and transit, can affect the compositional profile of microorganisms inferred from metagenomics. Overgrowth or degradation can occur during the period between collection and (cryo) storage with the result that the sequencing profile may not accurately reflect the composition of the sample at the time of collection. An extended duration of storage can result in a shift in the relative representation of bacterial taxa and substantial variability in metagenomics data. For example, faecal samples stored for longer than 3 months at –80°C experience selective loss of *Bacteroides* spp.[6,60,61]

If the sample is obtained post mortem, it is essential to report the time from death to sample acquisition given extravasation of gut bacteria into the bloodstream that can complicate interpretation of metagenomic data. For some applications, it might be relevant to report the overall turnaround time of the bioinformatic analyses—ie, including computational time for bioinformatics analysis. For example, Oxford Nanopore technology may be deployed in the field or at point of need, allowing sequencing to be done rapidly in near real-time; still, actionable results are also dependent on the time required for computational analysis.[62,62] The turnaround time of bioinformatic analyses is crucial in the context of clinical applications, when metagenomics is used to help to guide or tailor patient treatment. Variables such as sequencing run time and total computational analysis time (with system specifications—eg, number of cores and amount of memory used) should be stated clearly, as should the sequencing depth.[64]

Setting

State whether sample collection was retrospective or prospective

As described in the STAndards for Reporting of Diagnostic accuracy (STARD) guidelines, clarity is needed regarding the sequence of events in diagnostic testing to ensure that sources of bias are addressed.[65] The analyte can degrade if there is a long time in between sample collection and the metagenomics assay. Retrospective sampling might also lead to bias in the samples tested. For instance, when comparing studies of unidentified encephalitis, samples retrospectively selected for metagenomics might be those that are difficult to diagnose (eg, with a low titre) or taken at later timepoints in the course of infection, and therefore more likely to be non-infectious.[66]

Participants

Consider factors influencing microbiota compositions when selecting participants

Most diagnostic and public health laboratories do not yet use metagenomic technologies routinely. As such, patients included in metagenomics studies are often from tertiary referral or specialist centres, which are unlikely to be representative of the wider population, as discussed in STROBE and STROME-ID.[22,23] This limitation can introduce challenges for appropriate selection of controls for case-control studies and for studies assessing the strength of disease associations.

Species composition of human microbiomes are affected by various host factors, including age, sex, behaviour (eg, diet and lifestyle), and environment.[67,68] Exposure to pharmacological substances can also profoundly influence microbiome composition. For example, a single standard course of antibiotics has been shown to alter species composition of the gut and oral microbiomes for over a year.[69,70] Matching of cases and controls is particularly challenging for metagenomics studies given the broad range of microbes considered.[71] Metagenomics studies should aim to minimise and statistically control for host confounders or, at a minimum, list those confounders that might affect interpretation of results.

Bias

Bias is a source of error that remains constant with replication affecting trueness;[72] it is separate to random error, which affects the precision of an experiment. Together, these sources of error contribute to measurement uncertainty that, when conducting metagenomics sequencing, has many potential sources (figure 1). Replication, including replication of the whole process, provides a means to estimate random error, which can vary when using different sequencing strategies.[72] Adherence to strictly described laboratory protocols can improve random error and reproducibility,[21] but it cannot be used alone to remove bias.

Address potential sources of bias (sampling, transport, storage, library preparation, and sequencing)

Bias can occur at each step of a diagnostic sequencing pipeline (panel 1) and is more difficult to evaluate than random error. For metagenomics studies, microbiological contamination of samples can introduce bias. Experimental bias that is caused at different stages of a metagenomics experiment is more challenging to control for than selection bias or contamination. The fact that the microbiome is composed of many different microorganisms means that a given protocol could lead to certain groups being over-represented in the processed samples. For example, enrichment protocols can introduce bias for pathogen detection.[73] Capture probe-targeted sequencing will limit detection to targeted sequences, and 16S rRNA gene sequencing has limitations with regard to the level of taxonomic classification. This precise form of bias does not exist in untargeted metagenomics; however, other experimental bias can occur at different protocol stages, including during sampling, nucleic acid extraction,[74] or post-extraction steps.[75] Studies using 16S should consider that different primers amplify different bacterial families with varying degrees of success because of mismatches, resulting in potential bias in abundance and diversity metrics,[76] which cannot be completely corrected bioinformatically.[77]

By reporting the potential sources of bias for a given study (figure 1) their potential influence can be considered with mitigation or compensation strategies or caveats made to improve interpretation. The complexity and multistep nature of microbiome measurement means that any metagenomics experiment should be considered and reported as a representative

Panel 1: **Examples of potential sources of bias in metagenomics studies and implications for result interpretation***

Specimen collection methods

Collection without a cold chain, or nucleic acid stabilising agents, can cause nucleic acid degradation and potential false-negative results or overgrowth of selected organisms, which leads to misinterpretation of abundance. Multiple freeze-thaw cycles can also cause nucleic acid degradation.

Nucleic acid extraction method

The absence of a bead-beating step could make the detection of some bacteria difficult (ie, bacteria do not lyse properly so their DNA is not released and will not be sequenced). Small specimen volumes can reduce the ability to detect low-level organisms.

Sequencing library preparation

Poly-A tail enrichment of RNA will not include fragmented pathogen genomes; DNA sequencing alone will not detect RNA viruses.

Targeting of sequences

Capture probe-targeted sequencing will limit detection to targeted, known sequences. 16S targeted sequencing, as opposed to whole genome sequencing, will have limitations for the level of taxonomic classification.

Sequencing methods

High-level sample multiplexing can lead to insufficient read depth to detect organisms present at low levels. Computational contamination can occur between samples pooled on the same sequencing run due to a sample barcode for a sequence being misread and misassigned to another sample on the same run.[82] This is termed barcode bleed-through; dual barcodes drop the rate of bleed through dramatically compared with single barcodes. Unique molecular identifiers are an even more powerful way to identify this phenomenon when compared with dual barcodes.

Processing controls

Negative controls allow some contaminating organisms to be identified. Internal positive controls, reference standards such as sequins, reduce bias introduced by experimental variability and can improve recognition of low-level organisms.

Analysis methods

A small curated database, or highly stringent criteria might not include novel or unexpected organisms, leading to false negative results. An uncurated database or lenient criteria might also identify organisms incorrectly.

*This list is not comprehensive, but illustrates how results can be affected by collection, processing, and analysis methods.

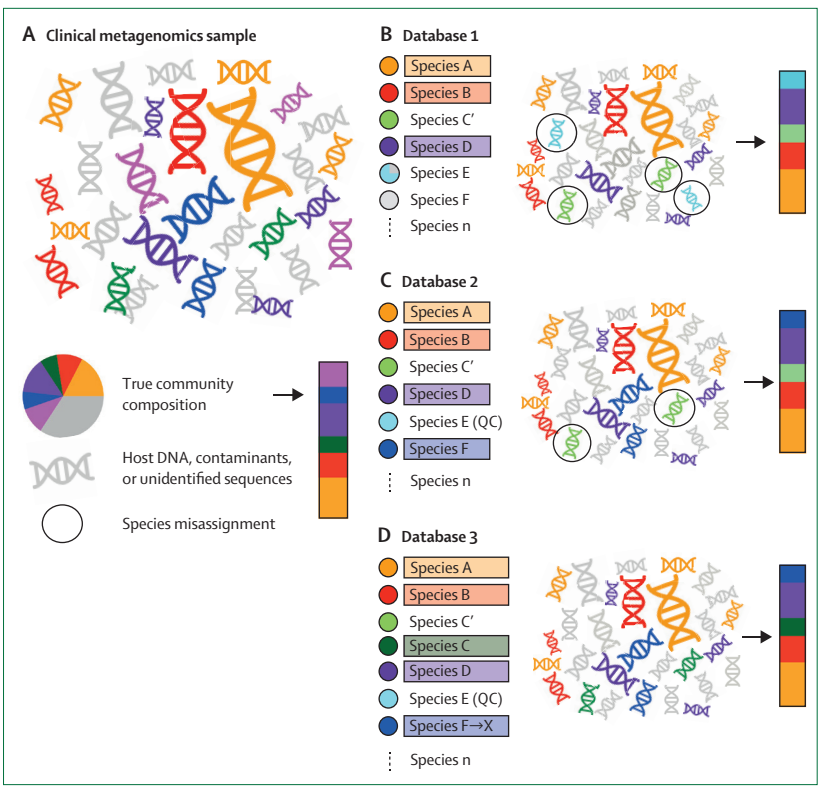

Figure 2: **The importance of reference database choice, design, and versioning in taxonomic profiling of clinical metagenomics samples**

(A) Schematic representation of a typical clinical metagenomics sample with species assigned as coloured DNA and grey denoting DNA deriving from the host, contaminants, unidentified taxa, or taxa sequenced at low depth. The pie chart provides the full metagenomic composition with the bar providing the species composition excluding host DNA and contaminants. (B) Taxonomic profiling based on database 1. Species confidently assigned are highlighted by colours with unassigned species shown in grey. Using database 1, species A, B, and D are correctly assigned. Species that are misassigned are outlined with a circle. In this instance, sequences from species C are assigned to the closely related species C′ because of the lack of a representative of species C in the reference database. Additionally, the reference database contains a partially contaminated sequence from species E, which is misassigned to contaminant sequences in the test clinical metagenomics sample. This affects the inference of species composition shown in the bar. (C) The addition of species F to database 2 allows assignment of a greater proportion of the species present in the original clinical metagenomics sample. Quality control and improvement of reference species E, now species E (QC), removes the spurious assignment of contaminant species. Species C is still misassigned to species C′, its closest representative in the database. (D) Updating the reference database to include species C results in the correct assignment of sequences to species C rather than species C′. Species F is taxonomically reassigned to species X, leading to a change in the assigned species name despite no change in the data in the reference or query datasets. In all cases the pink sequences present in the original clinical metagenomics sample are not assigned as this species is not present in any of the three reference databases.

result, rather than assuming that it perfectly reflects the microbes present and their abundance. It is also why the term unbiased, which is often used when describing metagenomic experiments that do not use enrichment, should be used with caution (or not at all). The term untargeted metagenomics could be used instead (appendix pp 3–4).

Address potential bias introduced by bioinformatics analysis

Classification algorithms rely on alignment of sequencing reads and contigs obtained from overlapping reads against reference genomes. In the case of the alignment of assembled contigs, reads that cannot be built into contigs (unassigned reads) are discarded, which can lead to a potential loss of information.[78] Classification of reads might be slow and a smaller database could be built with unique sequences representing certain taxa.[79] However, this can lead to bias in the assignment of homologous sequences and should be clearly reported.

Samples containing low abundance pathogens might produce false-negative results by not classifying sequencing reads as relevant or produce false-positive results if reads are non-specific.[80] Subsequent alignment of sequence reads against a reference genome of the candidate pathogen(s) identified by the metagenomics analysis can provide necessary validation—wide and distributed coverage of the reference genome and high mapping identity is unlikely to result in a false positive. The level of coverage might be limited in samples with low pathogen load but still can be a true-positive result. Sufficient read depth is not always available for metagenomics data from clinical samples, which often contain a large proportion of reads derived from the host. Additionally, high read depth can generally be achieved only for microbes present at high-copy number. Authors should report where these considerations are relevant.

Assessing the quality of reads before downstream classification is crucial for ensuring accuracy of taxonomic assignment. This quality control usually includes removal of adapters, background sequences (human, host, or known), low-complexity sequence reads, trimming of low-quality bases at the ends of reads, and removal of primer sequences. The total number of reads in each sample can be affected by factors including DNA extraction methods, sample handling, library preparation, differences in sequencing depth. As such, it is generally advisable to normalise read abundance between samples before any analysis and report where this is done.[81] Sophisticated statistical modelling approaches can deal with variation in read numbers between samples without loss of data (eg, DESeq2).[82]

Describe or address limitations of reference databases

The use of reference databases should be clearly described. It is crucial that the reference database, genomic data download date, and a description of the procedures behind the inclusion and indexing of reference sequences are clearly presented. Limitations of reference databases can interfere with correct assignment of sequences (figure 2). Curated reference databases might not include all the relevant microbial diversity. Conversely, non-curated databases can comprise incorrectly named, incomplete, low sequencing quality, or artefactual sequences.[83] Studies have shown that sequences arising from sample contamination or incompleteness (eg, an incomplete region of a genome that contains an important mutation) are frequent features of reference databases, particularly when draft

genomes are included. For example, over 1000 published microbial genome sequences have been identified as contaminated with phiX174, a bacteriophage used as a control in Illumina sequencing,[24] and 2250 NCBI GenBank draft bacterial and archaeal genomes contain spurious human sequences.[84] Additionally, false-negative results might be due to a focal species missing taxonomic representation in the databases, which have an inherent curatorial bias to known human associated pathogens (appendix pp 4–5).[85]

Study size

Describe clearly how power calculations were made

Whenever comparisons in metagenomic species composition between two or more groups are made, authors should report relevant parameters such as significance level, power threshold, sequencing depth, effect size, number of comparisons, methods used to correct for multiple comparisons, and details of the statistical methods used for power calculations. It should be clearly stated how an effect size was derived and a rationale for the clinical relevance of the specific effect size should be given. If no power calculation was made, an explanation should be given about why this was not considered feasible or useful (appendix pp 5–6).

Statistical methods

State the limit of detection, including analytical sensitivity and specificity

The limit of detection (LOD) refers to the minimum quantity of genomic material from an organism required for its detection and should be stated in metagenomics studies. Determination of the LOD for a metagenomics study is dependent on the sequencing technology, sequencing depth, read length, representation of genomes related to the taxa of interest in the reference database, and the complexity of the community and amount of host nucleic acid in the sample. Simple calculations give estimates for the LOD (eg, for 10^6 reads per sample, the LOD is one read per sample), which corresponds to a relative abundance of the order of magnitude of 10^{-6} (ie, ~0·0001%). Formal calculations of LOD that are needed for clinical validation should be done using probit analysis.[86] In practice, the LOD will be considerably higher than that derived from these calculations because a single read from a taxon is very likely to be due to contamination or misclassification. Rather than trusting such calculations, the use of positive (spiked) controls and negative controls in the sequencing run allows assessment of sensitivity and specificity. With a single infection, the number of on-target reads will be correlated with the signal in the sample but mixed infections and coinfections will influence sensitivity.[87] Experimentally validating these for model organisms that represent the specific pathogens of interest (eg, a DNA virus, an RNA virus, Gram-negative and Gram-positive bacteria, etc) is recommended, particularly for diagnostic tests.

Discussion

Attempt or acknowledge the need for functional or phenotypic validation

Genotypic data do not always correlate with clinical phenotype; for example, mechanisms that involve inducible resistance, gene expression and regulation, or post-translational modifications. In studies investigating mixed microbial communities it may not always be possible to determine which taxon a particular gene belongs to.[88,89] This is also relevant in the establishment of causality.

Efforts should be made to undertake phenotypic and functional validation to assess the inferred results. If this is not possible, or beyond the scope of the study, the limitations of inferring results solely from genotypic data should be acknowledged and discussed, including known caveats and restrictions on making key assumptions.

Consider the need for species or strain resolution

Different strains or lineages within a species can differ widely in their phenotypic characteristics. For example, sequencing with strain-level resolution enabled identification of specific strains of *Escherichia coli* associated with necrotising enterocolitis in preterm newborns[90] and lineages of *Salmonella enterica* associated with varying clinical phenotypes.[91] Therefore, profiling microbial communities with sub-species resolution can be useful, although de novo assembly of metagenomic reads remains a methodological challenge.

The strain and species resolution capacity of the assay used should be clearly stated with consideration for how the resolution applies to the study in question. In particular, microbial community profiling using 16S rRNA gene sequencing cannot identify individual species within some genera and should never be used to identify to the strain level. As recommended in STROME-ID, a definition or reference to published definitions of a strain should be provided.[23]

Other information

Report any ethical considerations with specific implications for metagenomics

Metagenomics produces a vast amount of host and pathogen data, which are untargeted and sometimes not of immediate interest.[92] Molecular methods to deplete human genomic material exist; however, they remain imperfect. It might be sufficient to detail in a protocol that the host data will be removed, and not analysed, although this approach could lead to bias in microbial reads caused by the in silico host-depletion method—host genomes can contain viable viral genomes and non-viable genetic material derived from or shared with micro-organisms. In these cases, the method used to identify and exclude host reads—eg, through mapping of all reads to the host reference genome—should be reported. including the choice of mapping algorithm and programme parameters.

Search strategy and selection criteria

In 2018, a STROBE-metagenomics working group was established, identified through notable researchers in the field, including a geographically diverse group of epidemiologists, statisticians, bioinformaticians, neurologists, virologists, microbiologists, and specialists in public health and infectious diseases. Participants met to agree the structure and content of the statement, and the proposal was registered with the Equator Network.[24] Specific issues to be covered were identified (panel 2).
A systematic approach was taken to gather evidence to support the recommendations, with literature searches performed in PubMed, searching references of articles, and supplemented by expert opinion. Literature searches were done in PubMed using medical subject headings terms and keywords "(?sequenc* OR metagenom* OR Illumina OR RNA-seq OR RNASeq OR (Roche 454) OR (Ion torrent) OR (Proton / PGM) OR MiSeq OR HiSeq OR NextSeq OR MinION OR Nanopore OR PacBio) AND (infectio* OR microorganism OR microorganisms OR pathogen OR pathogens OR bacteria* OR virus OR viral OR fungus OR fungi OR parasite OR parasites OR parasitic)", searching references of articles, and supplemented by expert opinion from within the group. Articles were limited to those in English language published between January, 2000, and June, 2019. Areas that were adequately addressed in existing STROBE[22] and STROME-ID[23] statements were not covered. Iterative versions of the guidelines and manuscript were circulated to develop a consensus. The STROBE-metagenomics extension has been developed to complement the STROBE and STROME-ID statements, with the new recommendations organised alongside the existing table. The guidelines discussed therefore cover only the new proposals for reporting.

Panel 2: Key issues to be addressed in publications applying metagenomics

- Specimen collection, handling, preservation, and storage
- Nucleic acid extraction
- Sequencing instrumentation and processing, including library preparation
- Bioinformatic analysis method, including workflow, database composition, and parameterisation
- Quality assurance measures, including internal quality control, such as the use of adequate internal and external controls
- Limits of detection, including analytical sensitivity, and specificity for clinical testing
- Power and sample size calculations
- Use of orthogonal methods to confirm sequencing results
- Criteria to confirm the role of pathogen(s) in disease aetiology
- Turnaround time
- Cost
- Ethical considerations
- Specific issues related to applications, such as in the diagnosis of CNS infections, and investigation of antimicrobial resistance

Even if data analysis is restricted to non-human reads, it could still unveil potentially sensitive information,[93] such as a new diagnosis of HIV. It has also been shown that more than 80% of individuals can be identified from populations of hundreds using their gut microbiome profile.[94] These issues pose real concerns, particularly with the increasing requirement for data to be made publicly available. For all these reasons, specific ethical implications relating to metagenomics data and corresponding approvals should be stated, and appropriate ethical approval should be obtained.

Conclusions

Metagenomics has already made a significant impact on pathogen detection and characterisation, and we probably still underestimate its full potential. Increasing use of metagenomics has been accompanied by recognition of complex issues at every stage in the pipeline—ie, sample collection, sequencing, and analysis. Standards for reporting are therefore needed to ensure clarity, consistency, and robustness of research. The guidance given in this paper constitutes a set of recommendations and we recognise that research studies need to be pragmatic and use available resources. Nonetheless, reporting known and potential limitations should minimise misrepresentation. It is inevitable that the field of metagenomics will continue to advance steadily and these guidelines will need to be updated.

Contributors

TB and NF conceived the idea and, together with CO, co-ordinated the Review. DOS and JH designed figure 1 and LvD and FB designed figure 2. All authors were involved in the study design, literature review, writing the manuscript, and editing successive drafts.

Declaration of interests

ME reports personal fees and other financials from PATHOQUEST, none received during the conduct of the study. MRW has a patent issued for Depletion of Abundant Sequences by Hybridization. All other authors declare no competing interests.

Acknowledgments

AN and LBR are National Institute of Health (NIH) employees and are in receipt on an NIH grant (NS003130). TB is supported by the University of Oxford and the Medical Research Council (grant number MR/N013468/1). MW is funded by a National Institute of Neurological Disorders and Stroke (grant number K08NS096117). LVT is a Wellcome Research Fellow (grant number 204904/Z/16/Z).

References

1 Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* 1998; **5:** R245–49.
2 Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnol* 2017; **35:** 833–844.
3 Forbes JD, Knox NC, Peterson C-L, Reimer AR. Highlighting clinical metagenomics for enhanced diagnostic decision-making: a step towards wider implementation. computational and structural biotechnology journal. *Comput Struct Biotechnol J* 2018; **16:** 108–20.
4 Chiu CY, Miller SA. Clinical metagenomics. *Nat Rev Genet* 2019; **20:** 341–55.
5 Forbes JD, Knox NC, Ronholm J, Pagotto F, Reimer A. Metagenomics: the next culture-independent game changer. *Front Microbiol* 2017; **8:** 1069.
6 Gorzelak MA, Gill SK, Tasnim N, Ahmadi-Vand Z, Jay M, Gibson DL. Methods for improving human gut microbiome data by reducing variability through sample processing and storage of stool. *PLoS One* 2015; **10:** e0134802.
7 Simner PJ, Miller S, Carroll KC. Understanding the promises and hurdles of metagenomic next-generation sequencing as a diagnostic tool for infectious diseases. *Clinical Infect Dis* 2018; **66:** 778–88.

8 Nakamura S, Yang C-S, Sakon N, et al. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS One* 2009; **4:** e4219.

9 van der Helm E, Imamovic L, Hashim Ellabaan MM, van Schaik W, Koza A, Sommer MOA. Rapid resistome mapping using nanopore sequencing. *Nucleic Acids Res* 2017; **45:** e61.

10 Kim D, Hofstaedter CE, Zhao C, et al. Optimising methods and dodging pitfalls in microbiome research. *Microbiome* 2017; **5:** 52.

11 Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020; **395:** 497–506.

12 Zhu N, Zhang D, Wang W, et al. a novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* 2020; **382:** 727–733.

13 Brown JR, Bharucha T, Breuer J. Encephalitis diagnosis using metagenomics: application of next generation sequencing for undiagnosed cases. *J Infect* 2018; **76:** 225–240.

14 Wilson MR, Sample HA, Zorn KC, et al. Clinical metagenomic sequencing for diagnosis of meningitis and encephalitis. *N Engl J Med* 2019; **380:** 2327–40.

15 Zhernakova A, Kurilshikov A, Bonder MJ, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* 2016; **352:** 565–69.

16 Wirbel J, Pyl PT, Kartal E, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med* 2019; **25:** 679–89.

17 Greninger AL, Zerr DM, Qin X, et al. Rapid metagenomic next-generation sequencing during an investigation of hospital-acquired human parainfluenza virus 3 infections. *J Clin Microbiol* 2017; **55:** 177–82.

18 Loman NJ, Constantinidou C, Christner M, et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *JAMA* 2013; **309:** 1502–10.

19 Brooks JP, Edwards DJ, Harwich MD, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol* 2015 ; **15:** 66.

20 Ruppé E, Lazarevic V, Girard M, et al. Clinical metagenomics of bone and joint infections: a proof of concept study. *Sci Rep* 2017; **7:** 7718.

21 Schmidt K, Mwaigwisya S, Crossman LC, et al. Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J Antimicrob Chemother* 2017; **72:** 104–14.

22 von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Int J Surg* 2008; **61:** 344–49.

23 Field N, Cohen T, Struelens MJ, et al. Strengthening the reporting of molecular epidemiology for infectious diseases (STROME-ID): an extension of the STROBE statement. *Lancet Infect Dis* 2014; **14:** 341–52.

24 Equator Network. EQUATOR (Enhancing the QUAlity and Transparency of health Research) Network home page. 2009. https://www.equator-network.org/ (accessed Jan 1, 2020).

25 Duncan CJ, Mohamad SM, Young DF, et al. Human IFNAR2 deficiency: lessons for antiviral immunity. *Sci Transl Med* 2015; **7:** 307ra154.

26 Morfopoulou S, Brown JR, Davies EG, et al. Human Coronavirus OC43 associated with fatal encephalitis. *N Engl J Med* 2016; **375:** 497–98.

27 Bustin SA, Benes V, Garson JA, et al. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* 2009; **55:** 611–22.

28 Yohe S, Hauge A, Bunjer K, et al. Clinical validation of targeted next-generation sequencing for inherited disorders. *Arch Pathol Lab Med* 2015; **139:** 204–10.

29 Jennings LJ, Arcila ME, Corless C, et al. Guidelines for validation of next-generation sequencing-based oncology panels: a joint consensus recommendation of the association for molecular pathology and college of american pathologists. *J Mol Diagn* 2017; **19:** 341–65.

30 Schlaberg R, Chiu CY, Miller S, Procop GW, Weinstock G. Validation of metagenomic next-generation sequencing tests for universal pathogen detection. *Arch Pathol Lab Med* 2017; **141:** 776–86.

31 Seelenfreund E, Robinson WA, Amato CM, Tan A-C, Kim J, Robinson SE. Long term storage of dry versus frozen RNA for next generation molecular studies. *PLoS One* 2014; **9:** e111827.

32 Panek M, Čipčić Paljetak H, Barešić A, et al. Methodology challenges in studying human gut microbiota—effects of collection, storage, DNA extraction and next generation sequencing technologies. *Sci Rep* 2018; **8:** 5143.

33 Wu WK, Chen CC, Panyod S, et al. Optimization of fecal sample processing for microbiome study—the journey from bathroom to bench. *J Formos Med Assoc* 2019; **118:** 545–55.

34 Ali N, Rampazzo RCP, Costa ADT, Krieger MA. Current nucleic acid extraction methods and their implications to point-of-care diagnostics. *Biomed Res Int* 2017; **2017:** 9306564.

35 Minervini CF, Cumbo C, Orsini P, et al. Nanopore sequencing in blood diseases: a wide range of opportunities. *Front Genet* 2020; **11:** 76.

36 Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res* 2015; **43:** e37.

37 Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* 2014; **15:** 121–32.

38 Clark MJ, Chen R, Lam HYK, et al. Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol* 2011; **29:** 908–14.

39 Jennings LJ, Arcila ME, Corless C, et al. Guidelines for validation of next-generation sequencing-based oncology panels: a joint consensus recommendation of the association for molecular pathology and college of american pathologists. *J Mol Diagn* 2017; **19:** 341–65.

40 Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 2014; **12:** 87.

41 Branton WG, Ellestad KK, Maingat F, et al. Brain microbial populations in HIV/AIDS: alpha-proteobacteria predominate independent of host immune status. *PLoS One* 2013; **8:** e54673.

42 Singer E, Andreopoulos B, Bowers RM, et al. Next generation sequencing data of a defined microbial mock community. *Sci Data* 2016; **3:** 160081.

43 Rinke C, Low S, Woodcroft BJ, et al. Validation of picogram- and femtogram-input DNA libraries for microscale metagenomics. *PeerJ* 2016; **4:** e2486.

44 Bowers RM, Clum A, Tice H, et al. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genomics* 2015; **16:** 856.

45 Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012; **486:** 207–14.

46 Branton WG, Lu JQ, Surette MG, et al. Brain microbiota disruption within inflammatory demyelinating lesions in multiple sclerosis. *Sci Rep* 2016; **6:** 37344.

47 Hardwick SA, Chen WY, Wong T, et al. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. *Nat Commun* 2018; **9:** 3096.

48 The Integrative HMP (iHMP) Research Network Consortium. The integrative human microbiome project. *Nature* 2019; **569:** 641–48.

49 Hornung BVH, Zwittink RD, Kuijper EJ. Issues and current standards of controls in microbiome research. *FEMS Microbiol Ecol* 2019; **95:** fiz045.

50 Cummings LA, Kurosawa K, Hoogestraat DR, et al. Clinical next generation sequencing outperforms standard microbiological culture for characterizing polymicrobial samples. *Clin Chem* 2016; **62:** 1465–73.

51 Lipkin WI. Microbe hunting. *Microbiol Mol Biol Rev* 2010; **74:** 363–77.

52 Granerod J, Cunningham R, Zuckerman M, et al. Causality in acute encephalitis: defining aetiologies. *Epidemiol Infect* 2010; **138:** 783–800.

53 Fischbach MA. Microbiome: focus on causation and mechanism. *Cell* 2018; **174:** 785–90.

54 Langelier C, Kalantar KL, Moazed F, et al. Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. *Proc Natl Acad Sci USA* 2018; **115:** e12353–62.

55 Singh VP, Proctor SD, Willing BP. Koch's postulates, microbial dysbiosis and inflammatory bowel disease. *Clin Microbiol Infect* 2016; **22:** 594–99.

56 Gyarmati P, Kjellander C, Aust C, Song Y, Öhrmalm L, Giske CG. Metagenomic analysis of bloodstream infections in patients with acute leukemia and therapy-induced neutropenia. *Sci Rep* 2016; **6:** 23532.

57 Grumaz S, Stevens P, Grumaz C, et al. Next-generation sequencing diagnostics of bacteremia in septic patients. *Genome Med* 2016; **8:** 73.

58 Fredricks DN, Relman DA. Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clin Microbiol Rev* 1996; **9:** 18–33.

59 Lipkin WI. The changing face of pathogen discovery and surveillance. *Nature Rev Microbiol* 2013; **11:** 133–41.

60 Cuthbertson L, Rogers GB, Walker AW, et al. Time between collection and storage significantly influences bacterial sequence composition in sputum samples from cystic fibrosis respiratory infections. *J Clin Microbiol* 2014; **52:** 3011–16.

61 Cardona S, Eck A, Cassellas M, et al. Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiol* 2012; **12:** 158.

62 Senol Cali D, Kim JS, Ghose S, Alkan C, Mutlu O. Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. *Brief Bioinform* 2019; **20:** 1542–1559.

63 Balloux F, Brønstad Brynildsrud O, van Dorp L, et al. From theory to practice: translating whole-genome sequencing (WGS) into the clinic. *Trends Microbiol* 2018; **26:** 1035–48.

64 Miller RR, Montoya V, Gardy JL, Patrick DM, Tang P. Metagenomics for pathogen detection in public health. *Genome Med* 2013; **5:** 81.

65 Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003; **49:** 7–18.

66 Ambrose HE, Granerod J, Clewley JP, et al. Diagnostic strategy used to establish etiologies of encephalitis in a prospective cohort of patients in England. *J Clin Microbiol* 2011; **49:** 3576–83.

67 Shaw L, Ribeiro ALR, Levine AP, et al. The human salivary microbiome is shaped by shared environment rather than genetics: evidence from a large family of closely related individuals. *mBio* 2017; **8:** e01237–17.

68 Lassalle F, Spagnoletti M, Fumagalli M, et al. Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. *Mol Ecol* 2018; **27:** 182–95.

69 Shaw LP, Bassam H, Barnes CP, Walker AS, Klein N, Balloux F. Modelling microbiome recovery after antibiotics using a stability landscape framework. *ISME J* 2019; **13:** 1845–56.

70 Zaura E, Brandt BW, Teixeira de Mattos MJ, et al. Same exposure but two radically different responses to antibiotics: resilience of the salivary microbiome versus long-term microbial shifts in feces. *mBio* 2015; **6:** e01693–15.

71 Schenk T, Enders M, Pollak S, Hahn R, Huzly D. High prevalence of human parvovirus B19 DNA in myocardial autopsy samples from subjects without myocarditis or dilative cardiomyopathy. *J Clin Microbiol* 2009; **47:** 106–10.

72 Sullivan DM, Laver T, Temisak S, et al. Assessing the accuracy of quantitative molecular microbial profiling. *Int J Mol Sci* 2014; **15:** 21476–91.

73 Pettengill JB, McAvoy E, White JR, Allard M, Brown E, Ottesen A. Using metagenomic analyses to estimate the consequences of enrichment bias for pathogen detection. *BMC Res Notes* 2012; **5:** 378.

74 Velásquez-Mejía EP, de la Cuesta-Zuluaga J, Escobar JS. Impact of DNA extraction, sample dilution, and reagent contamination on 16S rRNA gene sequencing of human feces. *Appl Microbiol Biotechnol* 2018; **102:** 403–11.

75 Huggett JF, Laver T, Tamisak S, et al. Considerations for the development and application of control materials to improve metagenomic microbial community profiling. *Accred Qual Assur* 2013; **18:** 77–83.

76 Cai L, Ye L, Tong AHY, Lok S, Zhang T. Biased diversity metrics revealed by bacterial 16S pyrotags derived from different primer sets. *PLoS One* 2013; **8:** e53649.

77 Edgar RC. UNBIAS: an attempt to correct abundance bias in 16S sequencing, with limited success. *bioRxiv* 2017; published online April 4. https://www.biorxiv.org/content/10.1101/124149v1.full.pdf (preprint).

78 Paez-Espino D, Pavlopoulos GA, Ivanova NN, Kyrpides NC. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nat Protoc* 2017; **12:** 1673–82.

79 Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* 2016; **26:** 1721–29.

80 Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol* 2018; **19:** 198.

81 Pereira MB, Wallroth M, Jonsson V, Kristiansson E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics* 2018; **19:** 274.

82 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014; **15:** 550.

83 Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand Genomic Sci* 2015; **10:** 18.

84 Breitwieser FP, Pertea M, Zimin AV, Salzberg SL. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res* 2019; **29:** 954–60.

85 Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* 2012; **9:** 811–14.

86 Burd EM. Validation of laboratory-developed molecular assays for infectious diseases. *Clin Microbiol Rev* 2010; **23:** 550–76.

87 Greninger AL. The challenge of diagnostic metagenomics. *Expert Rev Mol Diagn* 2018; **18:** 605–15.

88 Miller S, Naccache S, Samayoa E, et al. Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome Res* 2019; **29:** 831–42.

89 Relman DA. Actionable sequence data on infectious diseases in the clinical workplace. *Clin Chem* 2015; **61:** 38–40.

90 Ward DV, Scholz M, Zolfo M, et al. Metagenomic sequencing with strain-level resolution implicates uropathogenic *E coli* in necrotizing enterocolitis and mortality in preterm infants. *Cell Rep* 2016; **14:** 2912–24.

91 Khajanchi BK, Yoskowitz NC, Han J, Wang X, Foley SL. Draft genome sequences of 27 *Salmonella enterica* serovar schwarzengrund isolates from clinical sources. *Microbiol Resour Announc* 2019; **8:** e01687–18.

92 Charalampous T, Kay GL, Richardson H, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat Biotechnol* 2019; **37:** 783–92.

93 Ruppé E, Schrenzel J. Messages from the third international conference on clinical metagenomics (ICCMg3). *Microbes Infect* 2019; **21:** 273–277.

94 Franzosa EA, Huang K, Meadow JF, et al. Identifying personal microbiomes using metagenomic codes. *Proc Natl Acad Sci USA* 2015; **112:** e2930–08.