STROBE-Metagenomics: A STROBE/STROME-ID Extension Statement to Guide the Reporting of Studies Applying Metagenomics

Authors

Tehmina Bharucha*[1,2] MRCP

Clarissa Oeser[3] PhD

Francois Balloux[4] PhD, Professor

Julianne R. Brown[5] PhD

Ellen Carbo[6] PhD

Andre Charlett[7] PhD

Charles Chiu[8] PhD, Professor

Eric C. J. Claas[6] PhD

Marcus de Goffau[9] PhD

Jutte de Vries[6] PhD

Marc Eloit[10] PhD, Professor

Susan Hopkins[11, 12] PhD

Jim F. Huggett[13, 14] PhD

Duncan MacCannell[15] PhD

Sofia Morfopoulou[4] PhD

Avindra Nath[16] MD, Professor

Denise M. O'Sullivan[13] PhD

Lauren B. Reoma[16] MD

Liam Shaw[17] PhD

Igor Sidorov[6] PhD

Patricia J. Simner[18] PhD

Le Van Tan[19] PhD

Emma C. Thomson[20] FRCP, Professor

Lucy van Dorp[4] PhD

Michael R. Wilson[21] MD

Judith Breuer[4,22] MD, Professor

Nigel Field*[3] PhD

Affiliations:

1. Department of Biochemistry, University of Oxford, Oxford, U.K.
2. Lao-Oxford-Mahosot Hospital-Wellcome Trust Research Unit, Vientiane, Laos
3. Centre for Molecular Epidemiology and Translational Research, University College London, London, U.K.
4. UCL Genetics Institute, University College London, London, U.K.
5. Microbiology, Virology and Infection Prevention and Control, Great Ormond Street Hospital for Children NHS Foundation Trust, London, U.K.
6. Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands
7. Statistics, Modelling and Economics Department, Public Health England, London, U.K.
8. Department of Laboratory Medicine, University of California San Francisco, San Francisco, U.S.A.
9. Wellcome Sanger Institute, Hinxton, U.K.
10. Pathogen discovery laboratory, Institut Pasteur, Paris, France
11. Healthcare-Associated Infection and Antimicrobial Resistance, Public Health England, London, U.K.
12. Infectious diseases unit, Royal Free Hospital, London, U.K.
13. National Measurement Laboratory, LGC, Teddington, U.K.
14. School of Biosciences & Medicine, Faculty of Health & Medical Sciences, University of Surrey, Guildford, UK
15. Office of Advanced Molecular Detection, Centers for Disease Control and Prevention, Atlanta, U.S.A.
16. Section of Infections of the Nervous System, National Institutes of Health, Bethesda, U.S.A.
17. Nuffield Department of Medicine, University of Oxford, Oxford, UK
18. Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, USAEmerging Infections Group, Oxford University Clinical Research Unit, Ho Chi Minh city, Vietnam
19. MRC-University of Glasgow Centre for Virus Research, Sir Michael Stoker Building, 464 BearsdenRoad, Glasgow, U.K.
20. Weill Institute for Neurosciences and Department of Neurology, University of California, San Francisco, U.S.A.

21. Division of Infection and Immunity, University College London, London, U.K.

*Co-corresponding authors. Correspondence details:

Dr Tehmina Bharucha. Institute of Glycobiology, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3RQ. 01865275723. t.bharucha@doctors.org.uk

Dr Nigel Field. Centre for Molecular Epidemiology and Translational Research, Mortimer Market Centre, Off Tottenham Court Road, UCL, London WC1E 6JB. 020 3108 2092. nigel.field@ucl.ac.uk

Background

The term "metagenome" was coined in 1998 to describe the collection of genomes from microbes present in environmental soil samples, using approaches previously employed to study single genomes (1). The sequencing of genetic material recovered directly from samples has now become common practice in research on 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 (i.e. microorganism as well as host nucleic acid), usually with the aim of pathogen identification for clinical diagnostics or research (2-4). Examples of real-world applications include pathogen detection and discovery, species characterisation or subtyping, antimicrobial resistance (AMR) detection, virulence profiling, and studies of the microbiome and microecological factors of health and disease (5-12). Metagenomics is also being introduced as a diagnostic tool in aetiological studies of clinical syndromes such as encephalitis (13, 14), for exploring the microbiome (15, 16), and tracking disease outbreaks (17, 18). A topical example of the value of sequencing has been the recent application for rapid investigation and dissemination of information on 2019-ncov (11, 12).

Metagenomics data are generated using high-throughput sequencing methods, also referred to as deep, next-generation, massively-parallel or shotgun sequencing. Here, for simplicity, we refer to all these approaches as sequencing. For the purposes of this statement, we also include capture probe enrichment-based sequencing methods that use nucleotide ‘baits’ to increase sensitivity (4), as well as targeted amplicon sequencing (e.g. sequencing the 16S rRNA gene to identify bacteria (19)). Although these latter two approaches might not be viewed as ‘true’ examples of metagenomics, and is not the focus of this review, 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 may be detected simultaneously with greater taxonomic resolution. More detailed characterisation of microbial community composition and population dynamics also enable the study of ecological interactions. Furthermore, this avoids culture, thereby including species that are difficult or time consuming to culture. This is particularly relevant for diagnostic applications, where routine culture is seeing a decline (20, 21).

However, appropriate study design remains in a state of relative infancy and metagenomic technologies raise important technical challenges, ranging from methodological artefacts introduced by wet laboratory methods to the impact of different computational approaches used to analyse and interpret multi-dimensional and complex data. Furthermore, the ethical implications of sequencing are significant, and human genomics and genetic privacy considerations are increasingly recognised as important. The multiple steps and different expertise required to generate and interogate metagenomic sequence data involves numerous decision points, each of which may introduce bias and affect downstream inference about the presence and abundance of species.

A metagenome result should therefore be considered with caution as one possible representation of the true sample composition of a given microbiome. Understanding the sources of bias and the limitations to valid inference should improve protocol performance, and enable metagenomic research to proceed with transparent recognition of the real-world 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, 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.

Aims of the STROBE-metagenomics extension statement

We aimed to improve the clarity and consistency of metagenomics research, ranging from clinical diagnostics to microbiome studies with suggestions for optimal practice and recommendations for robust and accurate reporting.

Search strategy

A STROBE-metagenomics core working group was established in 2018, 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 that needed to be covered were identified, and a systematic approach was taken to gather evidence to support recommendations, with literature searches performed in PubMed using Medical Subject Headings (MESH) 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. Subsequently, 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 recommendations organised to fit into the existing table. The guidelines discussed below therefore cover only the new proposals for reporting.

Issues identified by the STROBE-metagenomics working group that need 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 (AMR)

STROBE-metagenomics extension st\atement

1 Title and abstract

1.1 The term metagenomic/s should be included in the title or abstract and the keywords of the study when these methods contribute substantially 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 conducted for studies applying metagenomics in encephalitis utilising MESH and keyword searches for “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). In this case, the studies were identified by experts in the field who were directly involved with the studies.

4 Methods/ study design

4.1 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 impact on the results and reproducibility of a study and may introduce method-induced variability artefact (27). Many studies use material banked and collected originally for other purposes. Important potential sources of error are described below and their contribution to bias detailed in section 9.1.

Nucleic acids, particularly ribonucleic acid (RNA), are labile. Consequently, collection methods, the addition of nucleic acid stabilisers, and the time to processing can impact results obtained (28). To address these issues, reporting should include durations, volumes, temperatures and methods used before, during and after the storage of samples (section 9.1) (29, 30). Extraction methods are another major source of method-induced variation, and can be deoxyribonucleic acid (DNA)- or RNA-specific, or tailored to specific organism types, so should be described (31). Other details of sample preparation methods should also be reported, including filtration, centrifugation, DNA or nuclease digestion, rRNA depletion, separation in RNA/DNA or viral/ bacterial workflows or random amplification. Standardized protocols should also be followed, if available and appropriate, and documented clearly in the

publication methods. Authors should also consider submitting to standardized protocol repositories (e.g. www.protocols.io/).

4.2 Describe sequencing methods, including sequencing depth

Different metagenomic sequencing platforms may produce different types of reads, for example, single vs. paired-end, and short (100-300 bp) vs. long (>1,000 bp). Sequencing platforms have different error rates, with the probability of a base being read incorrectly ranging from <0.01% for Illumina® (Illumina, Inc.) to >10% for Oxford Nanopore Technologies (current figures as of May 2019 (32)). Additionally, sequencers often cope poorly 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 (33) (section 9.1 and 9.2).

Additionally, an important facet of the sequencing method is sequencing depth. This refers to the number of times a particular base is represented within reads, or the redundancy of coverage (34). This has implications for identification of low abundant transcripts, and improving confidence in sequencing data. However increasing sequencing depth must be balanced with available resources. There are a number of factors that affect sequencing depth, including the sequencing platform and the sequence that is being read (e.g. GC content and presence of pseudogenes) (34-36).

4.3 Describe methods used for bioinformatics analysis

For the purposes of this statement, the term ‘bioinformatics’ applies to all analysis steps downstream from the raw sequencing data, including base calling, de-multiplexing, trimming and removal of reads (e.g. those of low quality, low complexity, adapters and indexes, or of human origin), as well as read normalisation, alignment/mapping of sequence reads to reference databases, *de novo* assembling of genomes and taxonomic assignment of reads and/or assembled contigs. There are multiple viable options for many of these tasks, with ongoing debate in the community about optimal methods, which may depend on the scientific question at hand. The field is under rapid development and methods considered ‘best practice’ at one stage 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/flags. It is also advisable to make data and programming code open access, whether as supplementary files or shared, e.g. via Github or Figshare. Where possible, a version-controlled container, package or easily installable version of the complete analytic 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.

Taxonomic profiling and read classification methods fall into two categories, either using sequence composition (alignment-free) or sequence identity (similarity-based) compared against user-provided reference databases or aligning to marker genes (37). Similarity search-based methods, using algorithms such as megaBLAST, Bowtie2 (nucleotide) and DIAMOND, RAPSearch (translated nucleotide or protein), are considered the most sensitive methods for read classification, but require considerable computational power (38). Protein level taxonomic classification, based on translated nucleotide alignments is often considered the most appropriate for novel pathogen discovery and/or RNA virus detection (39), but may be prone to lower specificity than nucleotide-based classification for bacteria and eukaryotes due to higher sequence conservation in these organisms.

Methods using exact *k*-mer (sequences of *k* nucleotides) matching allow efficient analysis of samples and are typically accurate to at least the genus level but may identify false positive hits, for example with dinucleotide or homopolymer repeat regions, even if those can be removed prior to species-assignment, together with other low-complexity reads. To resolve ambiguous matches, several methods work either by partitioning reads with multiple possible assignments to the highest taxonomic level containing all matching species or their lowest common ancestor (LCA) (40), or consider the unique *k*-mer content relative to the number of reads assigned (41). Other similarity-based methods use probabilistic read classification to formally resolve ambiguous matches (42, 43).

All taxonomic profiling methods are limited by false positive and false negative results (22, 23). These can be reduced through appropriate read filtering before classification, for example, by removing low complexity and/or host DNA and rRNA gene fragments, and

depending on context, by penalising the addition of taxa in the results. Finally, a universal limitation of similarity-based methods is their reliance on public reference databases, even though masked databases can be used (44), which we describe in section 9.3. Several studies, including community-driven initiatives (45), have set out to benchmark the performance of the different classification methods. However, the necessary reliance on simulated data affects the evaluation of overall performance. The selection of appropriate classifiers is affected by the dataset, resource and research question. For instance, in microbiome research, accurate taxonomic classification to the species (or strain level) in a diverse community is critical, whereas lower resolution with more robust assignments may be required for clinical testing.

Rather than species composition, bioinformatic analyses can be targeted towards the detection of features of interest such as known AMR mutations/genes for one or multiple strains present in a metagenomic sample. An obvious application is the prediction of phenotype from sequence data, which is increasingly used in diagostics. However, implementation of phenotypic diagnostics tests based on whole-genome sequencing remains challenging even when applied to single species samples (46). The mixed nature of metagenomic data adds the further challenge of localising AMR genes and mutations to a particular species (47).

4.4 Describe quality assurance methods, including internal and external quality controls

An important strength of metagenomics studies is their ability to detect any genomic material present in a sample. However, this applies equally to ‘true’ sample material and to contaminating extraneous nucleic acids present in a sample, which may be introduced at any stage from sample collection to processing. For example, contamination may come from the extraction kit, the ‘kitome’ (48), or at the point of specimen collection. Sampling is rarely performed under completely sterile conditions, and tissues obtained from tissue banks are therefore often contaminated. Low biomass, low abundance sites (for example tumors, the brain, and fetal tissues such as the placenta) are particularly prone to risk of misclassification of contamination. However, their study is of potential great relevance and significance, necessitating thoughtful consideration of approach. A repreresentive balanced and unbiased view of such methodologies and their conclusions is outside the scope of the current review and panel expertise (49-54).

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-in DNA/RNA e.g. synthetic nucleic acid standards such as sequins (55), and negative controls are usually a blank/water sample or ideally a similar or identical matrix (tissue, body fluids, etc) that are expected to lack microorganism nucleic acid based on patient factors and test results. For clinical metagenomics, formal laboratory implementation involves a system of external controls. This is more difficult to arrange, however, publicly and commercially available controls and mock community samples are now available, and we recommend these are employed whenever possible (56, 57).

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

Broadly, culture, or growth of the pathogen from a clinical sample, and immunohistochemistry, the histological localisation of candidate species in tissue biopsies, are the conventional methods in microbiology for confirming the presence of a pathogen. Traditional culture, however, can be difficult when antibiotics have been administered before sampling, or for pathogens which are slow-growing, fastidious, present in low-concentration or as-yet undescribed. Sequencing has high discriminative power and may 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 demonstrated higher classification accuracy when comparing taxonomic profiles of synthetic polymicrobial samples obtained from extended quantitative culture with nonselective media (58).

Confirmatory assays appropriate to the study setting should be used. A justification for the methods employed and a description of their limitations should be provided. For cases in which confirmatory assays are not possible (e.g. due to 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.

Orthogonal methods should be considered for pathogen confirmation where immunohistochemistry or culture are not available or feasible (for example, species-specific qPCR, fluorescent in situ hybridisation ‘FISH’, and convalescent serology). Important tools

include measuring a sample twice, preferably with a different extraction method in separate runs, or subsequent verification with a different sequencing setup. However, contamination and incompleteness of reference databases can yield repeated false positive and false negative results (44) (section 9.3).

Sequence data, if applied only to DNA, should not automatically be assumed to indicate viability of an organism and/or functional expression of a gene, discussed in 4.3. In theory, follow-up proteomics could be used to confirm gene expression, although challenges and inconsistencies currently exist in correlating multi-omics data and this is rarely performed (59). Alternative methods to confirm functionality may also be useful, such as the use of animal models to confirm findings of microbiome studies (60).

Detection of DNA (or proteins) alone does not establish the presence of a viable organism. The half-life of DNA and proteins is typically long, and DNA can be detected days, if not weeks, after clearance of an infection (61, 62). If the primary objective of a study is to detect viable organisms from tissue, sequencing of RNA is preferable due to its shorter half-life, i.e. in the order of minutes, however convincing evidence of viability would depend upon further experiments (63). An approach to detect metabolically active organisms (both dormant and replicating) is to treat the sample with Ethidium-monoazide or Propidium-monoazide (PMA) to selectively remove DNA from dead cells during downstream sequencing process (64), though these protocols remain challenging to optimise, especially for clinical use (65).

4.6 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 (66, 67). A major challenge for metagenomics-based research and diagnostics is distinguishing pathogens from commensals or contaminants (68, 69). 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 (70). It is also worth considering that the aetiology of several diseases may involve multiple sequential or interacting species, which may be collectively important (71, 72). Furthermore, sequencing investigations may identify novel organisms, for which the clinical significance will be unknown. These issues are particularly relevant in the investigation of aetiology of central nervous system (CNS) infections due to the low biomass.

A number of criteria to establish causality have been proposed over the years, detailed in Supplementary Data A, including the incorporation of metagenomic technologies (73, 74).

4.7 State the time from collection to results and cost considerations

The time from sample collection to processing ('transport time'), discussed in section 4.1, including cold-chain and transit, can also affect the compositional profile of microorganisms inferred from metagenomics. Overgrowth and degradation may occur during the initial 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 its collection. An extended duration of storage can also lead to a shift in the relative representation of bacterial taxa and result in significant variability in metagenomics data. For example, faecal samples stored for longer than three months at -80°C experience selective loss of *Bacteroides spp.* (6, 75, 76).

If the sample is obtained post-mortem, another critical time period to report is the time from death to sample acquisition given the extravasation of gut bacteria into the bloodstream that can complicate interpretation of any metagenomic data. For some applications, it may also be relevant to report the overall turnaround time of the bioinformatic analyses i.e. 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 performed rapidly in near 'real-time'; still, actionable results are also dependent on the time required for computational analysis (77, 78). The turnaround time of bioinformatic analyses is critical in the context of clinical applications, when metagenomics is used to help guide or tailor patient treatment. Variables such as sequencing run time and total computational analysis time (with system specifications e.g. number of cores and amount of memory utilised) should be stated clearly (79).

Currently, the cost of sequencing is a limiting factor in research and may determine the sequencing depth used for a given study. While it is not suggested that a detailed breakdown of the costs of sequencing is reported, it is recommended that authors detail the sequencing depth, and acknowledge possible compromises in study design due to resource limitations.

5 Setting

5.1 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 (80). Samples analysed long after collection, as well as the performance of the reference standard assays, may lead to degradation of the analyte. Retrospective sampling may also lead to bias in the samples tested. For instance, when comparing studies of unidentified encephalitis, samples retrospectively selected for metagenomics may be those that are 'difficult to diagnose', for example with a low titre, be taken at later time points in the course of infection, and be more likely to be non-infectious (81).

6 Participants

6.1 Consider factors influencing microbiota compositions when selecting participants

At the time of writing, most routine 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 can also introduce challenges around 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 a variety of host factors, including age and sex, as well as behaviour (e.g. diet and lifestyle) and environment (82, 83). Exposure to pharmaceutical agents may 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 (84). Matching of cases and controls is particularly challenging for metagenomics studies given the broad range of microbes considered (85). Metagenomics studies should aim to minimise and/or statistically control for host-confounders, or at a minimum, list those that may impact the interpretation of results.

9 Bias

Bias is a source of error that remains constant with replication affecting trueness (86); it occurs along with random error, which affects the precision of an experiment. Together, these sources of error contribute to measurement uncertainty which, when conducting metagenomics sequencing, can be derived from a large number of potential sources (see

Figure 1). Replication, including replication of the whole process, provides a means to estimate random error, which can even differ when using different sequencing strategies (86). Adherence to strictly described laboratory protocols may improve random error and reproducibility (21), but it cannot be used alone to remove bias.

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

Bias can occur at each step of a sequencing pipeline (Table 2) and is more difficult to evaluate than random error. When conducting metagenomics studies, different categories of bias may manifest. Selection bias, where researchers unconciously treat groups of samples differently, can affect many areas of research, and investigator blinding is a simple way to mitigate against this. Multiple water and reagent blank controls (see section 4.4) should be included, alongside checking for contamination from current and previous runs and cycle indexing (11).

Experimental bias that is caused at different stages of the multistep process of a metagenomics experiment is more challenging to control for than sources caused by selection bias or contamination. The fact that the microbiome is a mixture of different microorganisms means that a given protocol may favour certain groups being over-represented in the processed samples. Experimental bias can occur at different protocol stages, including at extraction (87) or during post-extraction steps (88).

Validation procedures applying mock communities (50-53) or synthetic spike-ins (4, 55) prior to library preparation can assist in understanding experimental performance and identifying bias. Spike-in controls have also been shown to be useful for discriminating whether detected microbes likely represent reagent contaminants versus true infections given that the abundance of the former tends to be inversely correlated to the input mass of the nucleic acid (i.e., contaminants form a higher proportion of datasets from low input samples) (89-91). Spike-in controls are also useful for evaluating reduced analytic sensitivity, as the number of recovered reads was shown to be an indicator for nucleic acid background (92). Enrichment protocols may also introduce bias in the data (93).

By considering the many potential sources of bias, outlined in Figure 1, their likely influence can be considered and, where possible, mitigation or compensation strategies applied. Due to the complexity of the microbiome and the multistep process required to evaluate it, it is

advisable that a metagenomics experiment should be considered as a representative result, rather than assuming it is a perfect reflection of the microbes present. It is also why the term ‘unbiased’, which is a popular term when describing metagenomic experiments that do not employ enrichment, should be used with caution (or not at all). ‘Untargeted metagenomics’ could be used instead.

Table 2. Examples of potential sources of bias in metagenomics studies and implications for result interpretation. This list is not comprehensive, but illustrates how results may be affected by collection, processing and analysis methods.

Potential source of bias	Example of implication for results interpretation
Specimen collection methods	Collection without a cold chain or nucleic acid stabilising agents may cause nucleic acid degradation thus risking false negative results or overgrowth of selected organisms leading to misinterpretation of abundance; multiple freeze-thawing may also cause nucleic acid degradation.
Nucleic acid extraction method	A lack of bead-beating step may limit the detection of difficult-to-lyse bacteria; small specimen volumes may 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.
Sequencing methods	High-level sample multiplexing may lead to insufficient read depth to allow detection of organisms present at low levels. Computational contamination may 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 (94). This is termed 'barcode bleed-through', and dual barcodes drop the rate of bleed through dramatically compared to single barcodes. Unique molecular identifiers (UMIs) are an even more powerful way to identify this phenomenon.
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 may improve recognition of low-level organisms.
Analysis methods	A small curated database, or highly stringent criteria may limit the identification of novel or unexpected organisms, leading to false negative results; an un-curated database or lenient criteria may mis-identify organisms.

Figure 1. Uncertainty (fishbone) diagram describing potential contributing sources of uncertainty in the multistep protocol required for metagenomics experiment.

9.2 Address potential bias introduced by bioinformatics analysis package(s)

As discussed in section 4.3, classification algorithms rely on alignment of sequencing reads and/or contigs obtained from overlapping reads against reference genomes. In the case of the alignment of assembled contigs, in principle, reads that cannot be built into contigs are discarded, leading to a potential loss of information (95). Classification of reads may be slow, and a smaller database may be built with unique sequences representing certain taxa (96). However, this may lead to bias in the assignment of homologous sequences.

Samples containing low abundance pathogens may give rise to false negatives by not classifying something as relevant, or false positives if reads are non-specific (41). Subsequent alignment of sequence reads against a reference genome of the candidate pathogen(s) identified by the metagenomics analysis can provide necessary further validation: wide and evenly distributed coverage of the reference genome and high mapping identity is unlikely to

be a false positive. However, sufficient read depth is not always available for metagenomics data from clinical samples, which often contain a majority of reads from the host. Additionally, high read depth can generally be achieved only for microbes present at high-copy number within the samples analysed.

Assessing the quality of reads before downstream classification is also important for ensuring accuracy of taxonomic assignment. This quality control usually includes removal of adapters, human/host/known background sequences, 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, differences in sequencing depth, etc (see section 9.1). As such, it is generally advisable to normalise read abundance between samples prior to any analysis (97). More sophisticated statistical modelling approaches can deal with variation in read numbers between samples without loss of data (e.g. DESeq2 (98)).

9.3 Describe/address limitations of reference database(s)

Limitations of reference databases may interfere with correct assignment of sequences, as illustrated in Figure 2. Curated reference databases may not include all the relevant microbial diversity. Conversely, non-curated databases may comprise incorrectly named, incomplete, low sequencing quality or artefactual sequences (99). Studies have shown that contamination from various sources or incompleteness (e.g. of a region of a genome that contains an important mutation) are frequent features of reference databases, particularly when draft genomes are included. For example, over 1,000 published microbial genome sequences have been identified as contaminated with PhiX174, a bacteriophage used as a control in Illumina sequencing (24), and 2,250 NCBI GenBank draft bacterial and archaeal genomes contain spurious human sequences (100). Additionally, false negatives may occur due to a focal species missing taxonomic representation in the databases, which have an inherent curatorial bias to known human associated pathogens (37).

In cases where databases may be sufficiently comprehensive, the choice of protein or nucleotide reference database depends on the study setting and objectives. Protein (translated nucleotide) similarity searches recover more remote homology, allowing for the detection of divergent viruses that might be missed by nucleotide similarity searches (101), but incur increased risk of false-positives. In other cases, intraspecies or sub-strain differences may be

clinically important, requiring nucleotide similarity searches against comprehensive reference databases, often employing computationally sophisticated methods (102, 103). Similarly, metagenomics gene screening approaches for specific genotypic features, such as for known antimicrobial resistance (AMR) mutations, rely on nucleotide-scale resolution and updated databases facilitating rule-based prediction of phenotypes. Increasingly, nucleotide sequences are being used as comparative references, providing environment-specific taxonomic composition which may be useful to identify disease or host-specific niches (104).

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

Outputs of almost all metagenomics assigners are ranked as taxonomic hits relative to a reference database. These provide convenient classification units, however, it is acknowledged that within-taxon genetic diversity can generate misassignments. A well-known example is the apparent mis-classification of *Shigella* despite its high sequence similarity (>80%) to *Escherichia coli* compared to other *Escherichia* species (105). It follows that a single cut-off is not possible to delineate different taxonomic levels. Despite recent progress in the ability of metagenomic approaches to detect 'unculturable' bacteria (106), classification of the latter remains challenging (107). New microorganisms and viruses are being constantly described and naming conventions may change following progress in the taxonomy of previously described taxa (Figure 2d). It is also not uncommon for the same fungal organism to have been given multiple names due to the historic nomenclature of referring to different sexual phases with distinct names (108). As such, it is critical that reference databases are presented together with a genomic data download date, and a description of the procedures behind the inclusion of reference sequences and their indexing. Lastly, we acknowledge that nomenclature is complex, and a moving field, and for this reason, our conclusion is that it is beyond the scope of this review.

10 Study size

10.1 Describe clearly how power calculations were undertaken

Whenever comparisons in metagenomic species composition between two or more groups are performed, authors should report relevant parameters such as significance level and power

threshold, sequencing depth, effect size, number of comparisons and methods used to correct for multiple comparisons, and provide details of the statistical methods used to perform 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 performed, an explanation should be given about why this was not considered feasible or useful.

The need for *a priori* power calculations is exemplified by a recent meta-analysis of gut microbiome studies; despite significant overall associations, single studies with small sample sizes often had few or no significant associations, most likely due to being under-powered (109). Methods for different scenarios have been described. For example, Thompson's method has been suggested as appropriate for the calculation of power to detect α-diversity, as has a method to calculate the power of PERMANOVA to estimate β-diversity (110), whilst La Rosa *et al* (111) proposed a power method based on Dirichlet multinomial distribution for analysing relative abundance (112).

Power calculations depend on a number of variables including expected effect size, sequencing depth, level of taxonomical description, acceptable significance level (*p*-value) and power threshold, choice of the test statistics, how multiple comparisons were corrected for and the sample size (111). Parameters such as significance level and power threshold are usually chosen by convention as p-value ≤ 0.05 and power of 80% (111) and the level of taxonomical description will depend on the question asked. However, other parameters, such as effect size, require more careful consideration and a rationale for chosen options.

Published effect sizes are affected by other factors which differ between studies (Section 9). Large effect sizes can overcome this issue - for example, differences between infants' and adults' microbiome composition have been observed even in studies using very different processing techniques because of the large effect size of age (113). This was demonstrated in a meta-analysis of the association between differences in the microbiome and obesity, which found that most studies lacked the power to detect modest effect sizes of 0.9-6% changes in alpha diversity (114). Pilot studies are useful to determine the expected effect size, although small pilot studies may overestimate an effect, and lead to underestimation of the required sample (115). It may be challenging to anticipate a clinically or biologically relevant effect size, particularly for microbiota studies or studies with indeterminate endpoints, but some justification and elaboration should be attempted.

Multiple comparisons are inevitable in metagenomics studies (e.g. between hundreds of bacterial species and tens of clinical variables), and these increase the probability of false positive findings. For example, Vogtman *et al* conducted whole-genome shotgun sequencing on faecal samples from 52 cases with colorectal cancer and 52 matched controls and compared findings from a previously published 16S rRNA gene study. The study was underpowered to detect many statistically significant associations after correction for multiple testing (116). Power calculations need to incorporate correction for multiple testing, using approaches such as the Bonferroni method, or the less conservative Benjamini-Hochberg method (10).

12 Statistical methods

12.1 State the limit of detection, including analytical sensitivity and specificity for clinical work

The limit of detection (LOD) refers to the minimum quantity of genomic material from an organism required for its detection. 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, as well as the complexity of the community in the sample, e.g. amount of host nucleic acid. Simple calculations give estimates for the LOD e.g. 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} i.e. ~0.0001%. Formal calculations of LOD that are needed for clinical validation should be performed using probit analysis (117). In practice, the LOD will be considerably higher because a single read from a taxon (a 'singleton') is very likely to be due to contamination or misclassification. Rather than trusting such calculations, the use of positive spike-in controls and negative controls in the sequencing run allows assessment of sensitivity and specificity (also see section 9.1). With a single infection, the number of on-target reads will be correlated with the signal in the sample, but mixed infections and co-infections will influence sensitivity (118). Experimentally validating these for model organisms that represent the specific pathogens of interest (e.g. a DNA virus, an RNA virus, Gram-negative and Gram-positive bacteria, etc) is recommended, particularly for diagnostic tests.

19 Discussion/ limitations

19.1 Attempt, or acknowledge the need for functional and/or phenotypic validation

Genotypic data do not always correlate with clinical phenotype; for instance, 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 (119, 120). This is also relevant in the establishment of causality, as discussed in section 4.6.

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.

19.2 Consider the need for species or strain resolution

Different strains or lineages within a species can differ widely in regard to their phenotypic characteristics. For example, sequencing with strain-level resolution enabled identification of specific strains of *E. coli* associated with necrotizing enterocolitis in preterm newborns (121) and lineages of *S. enterica* associated with varying clinical phenotypes (122). 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 of how this level of resolution yields relevant epidemiological or clinical information for the study in question. In particular, microbial community profiling using 16S rRNA gene PCR 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).

23 Other information

23.1 Report any ethical considerations with specific implications for metagenomics

Metagenomics produces a vast amount of host and pathogen data, which is untargeted and sometimes unwanted and unneeded (123). Molecular methods to deplete human genomic

material exist, however they remain imperfect and may lead to bias. Alternatively, it may be sufficient to detail in a protocol that the host data will be removed, and not analysed, although this may also lead to bias in microbial reads caused by the *in silico* host-depletion method; host genomes can contain viable viral genomes as well as non-viable genetic material derived from or shared with microorganisms. In these cases, the method used to identify and exclude host reads, for example through mapping of all reads to the host reference genome, should be detailed, including the choice of mapping algorithm and program parameters.

Even if data analysis is restricted to non-human reads, it may still unveil potentially sensitive information (124), such as a new diagnosis of HIV. It has also been demonstrated that >80% of individuals can be identified using their gut microbiome alone (125). 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. 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, utilising resources available. Nonetheless, reporting known and potential limitations should minimise misrepresentation. It is inevitable that the field will continue to advance steadily, and these guidelines will need to be updated.

Contributors

TB and NF conceived the idea, co-ordinated the review and submitted the final draft. All authors were involved in the study design, writing the manuscript and editing successive drafts.

Acknowledgement

Authors Avindra Nath and Lauren B. Reoma are NIH employees, and are in receipt on an NIH grant (no. NS003130).

Conflicts of interest

Dr. ELOIT reports personal fees and other from PATHOQUEST, none received during the conduct of the study. Dr. Wilson has a patent issued for Depletion of Abundant Sequences by Hybridization. None of the other authors declare any conflict of interests .

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. Chemistry & biology. 1998;5(10):R245-9.
2. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nature Biotechnology. 2017;35:833.
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. 2018;16:108-20.
4. Chiu CY, Miller SA. Clinical metagenomics. Nature Reviews Genetics. 2019;20(6):341-55.
5. Forbes JD, Knox NC, Ronholm J, Pagotto F, Reimer A. Metagenomics: The Next Culture-Independent Game Changer. Frontiers in Microbiology. 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(8):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 Infectious Diseases. 2018;66(5):778-88.
8. Nakamura S, Yang C-S, Sakon N, Ueda M, Tougan T, Yamashita A, et al. Direct Metagenomic Detection of Viral Pathogens in Nasal and Fecal Specimens Using an Unbiased High-Throughput Sequencing Approach. PloS one. 2009;4(1):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(8):e61.
10. Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C, Clarke E, et al. Optimizing methods and dodging pitfalls in microbiome research. Microbiome. 2017;5(1):52.
11. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet.
12. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. The New England journal of medicine. 2020.
13. Brown JR, Bharucha T, Breuer J. Encephalitis diagnosis using metagenomics: application of next generation sequencing for undiagnosed cases. J Infect. 2018.
14. Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J, et al. Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. New England Journal of Medicine. 2019;380(24):2327-40.
15. Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science (New York, NY). 2016;352(6285):565-9.
16. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med. 2019;25(4):679-89.
17. Greninger AL, Zerr DM, Qin X, Adler AL, Sampoleo R, Kuypers JM, et al. Rapid Metagenomic Next-Generation Sequencing during an Investigation of Hospital-Acquired Human Parainfluenza Virus 3 Infections. Journal of Clinical Microbiology. 2017;55(1):177-82.
18. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ, Quick J, 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(14):1502-10.
19. Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiology. 2015;15(1):66.
20. Ruppé E, Lazarevic V, Girard M, Mouton W, Ferry T, Laurent F, et al. Clinical metagenomics of bone and joint infections: a proof of concept study. Scientific reports. 2017;7(1):7718.

21. Schmidt K, Mwaigwisya S, Crossman LC, Doumith M, Munroe D, Pires C, et al. Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. J Antimicrob Chemother. 2017;72(1):104-14.
22. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008;61(4):344-9.
23. Field N, Cohen T, Struelens MJ, Palm D, Cookson B, Glynn JR, et al. Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases (STROME-ID): an extension of the STROBE statement. The Lancet Infectious Diseases. 2014;14(4):341-52.
24. Network E. EQUATOR (Enhancing the QUAlity and Transparency of health Research) Network home page. 2009.
25. Duncan CJ, Mohamad SM, Young DF, Skelton AJ, Leahy TR, Munday DC, et al. Human IFNAR2 deficiency: Lessons for antiviral immunity. Sci Transl Med. 2015;7(307):307ra154.
26. Morfopoulou S, Brown JR, Davies EG, Anderson G, Virasami A, Qasim W, et al. Human Coronavirus OC43 Associated with Fatal Encephalitis. The New England journal of medicine. 2016;375(5):497-8.
27. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, et al. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem. 2009;55(4):611-22.
28. 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(11):e111827.
29. Panek M, Čipčić Paljetak H, Barešić A, Perić M, Matijašić M, Lojkić I, et al. Methodology challenges in studying human gut microbiota - effects of collection, storage, DNA extraction and next generation sequencing technologies. Scientific reports. 2018;8(1):5143-.
30. Wu WK, Chen CC, Panyod S, Chen RA, Wu MS, Sheen LY, et al. Optimization of fecal sample processing for microbiome study - The journey from bathroom to bench. J Formos Med Assoc. 2018.
31. Ali N, Rampazzo RdC, #xe1, Pontello s, Costa ADT, Krieger MA. Current Nucleic Acid Extraction Methods and Their Implications to Point-of-Care Diagnostics. BioMed Research International. 2017;2017:13.
32. Bowden R, Davies RW, Heger A, Pagnamenta AT, de Cesare M, Oikkonen LE, et al. Sequencing of human genomes with nanopore technology. Nature Communications. 2019;10(1):1869.
33. 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(6):e37.
34. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15(2):121-32.
35. Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. Nature biotechnology. 2011;29(10):908-14.
36. Jennings LJ, Arcila ME, Corless C, Kamel-Reid S, Lubin IM, Pfeifer J, 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. The Journal of molecular diagnostics : JMD. 2017;19(3):341-65.
37. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods. 2012;9:811.
38. Patil KR, Haider P, Pope PB, Turnbaugh PJ, Morrison M, Scheffer T, et al. Taxonomic metagenome sequence assignment with structured output models. Nature Methods. 2011;8:191.
39. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature Communications. 2016;7:11257.
40. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Research. 2007;17(3):377-86.
41. Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biology. 2018;19(1):198.

42. Morfopoulou S, Plagnol V. Bayesian mixture analysis for metagenomic community profiling. Bioinformatics (Oxford, England). 2015;31(18):2930-8.
43. Francis OE, Bendall M, Manimaran S, Hong C, Clement NL, Castro-Nallar E, et al. Pathoscope: Species identification and strain attribution with unassembled sequencing data. Genome Research. 2013.
44. Lu J, Salzberg SL. Removing contaminants from databases of draft genomes. PLoS Comput Biol. 2018;14(6):e1006277.
45. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. Nature Methods. 2017;14:1063.
46. Harris SR, Cole MJ, Spiteri G, Sánchez-Busó L, Golparian D, Jacobsson S, et al. Public health surveillance of multidrug-resistant clones of Neisseria gonorrhoeae in Europe: a genomic survey. The Lancet Infectious Diseases. 2018;18(7):758-68.
47. Leggett RM, Alcon-Giner C, Heavens D, Caim S, Brook TC, Kujawska M, et al. Rapid profiling of the preterm infant gut microbiota using nanopore sequencing aids pathogen diagnostics. bioRxiv. 2018:180406.
48. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC biology. 2014;12:87-.
49. Branton WG, Ellestad KK, Maingat F, Wheatley BM, Rud E, Warren RL, et al. Brain microbial populations in HIV/AIDS: alpha-proteobacteria predominate independent of host immune status. PloS one. 2013;8(1):e54673.
50. Singer E, Andreopoulos B, Bowers RM, Lee J, Deshpande S, Chiniquy J, et al. Next generation sequencing data of a defined microbial mock community. Scientific Data. 2016;3:160081.
51. Rinke C, Low S, Woodcroft BJ, Raina J-B, Skarshewski A, Le XH, et al. Validation of picogram- and femtogram-input DNA libraries for microscale metagenomics. PeerJ. 2016;4:e2486.
52. Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, et al. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. BMC Genomics. 2015;16(1):856.
53. The Human Microbiome Project C, Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207.
54. Branton WG, Lu JQ, Surette MG, Holt RA, Lind J, Laman JD, et al. Brain microbiota disruption within inflammatory demyelinating lesions in multiple sclerosis. Scientific reports. 2016;6:37344.
55. Hardwick SA, Chen WY, Wong T, Kanakamedala BS, Deveson IW, Ongley SE, et al. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. Nature Communications. 2018;9(1):3096.
56. Proctor LM, Creasy HH, Fettweis JM, Lloyd-Price J, Mahurkar A, Zhou W, et al. The Integrative Human Microbiome Project. Nature. 2019;569(7758):641-8.
57. Hornung BVH, Zwittink RD, Kuijper EJ. Issues and current standards of controls in microbiome research. FEMS Microbiol Ecol. 2019;95(5):fiz045.
58. Cummings LA, Kurosawa K, Hoogestraat DR, SenGupta DJ, Candra F, Doyle M, et al. Clinical Next Generation Sequencing Outperforms Standard Microbiological Culture for Characterizing Polymicrobial Samples. Clin Chem. 2016;62(11):1465-73.
59. Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Briefings in Bioinformatics. 2016;19(2):286-302.
60. Hugenholtz F, de Vos WM. Mouse models for human intestinal microbiota research: a critical evaluation. Cellular and Molecular Life Sciences. 2018;75(1):149-60.
61. Branger S, Casalta JP, Habib G, Collard F, Raoult D. Streptococcus pneumoniae Endocarditis: Persistence of DNA on Heart Valve Material 7 Years after Infectious Episode. Journal of Clinical Microbiology. 2003;41(9):4435-7.
62. Selenska S, Klingmuller W. DNA recovery and direct detection of Tn5 sequences from soil. Lett Appl Microbiol. 1991;13(1):21-4.

63. Bernstein JA, Lin PH, Cohen SN, Lin-Chao S. Global analysis of Escherichia coli RNA degradosome function using DNA microarrays. Proc Natl Acad Sci U S A. 2004;101(9):2758-63.
64. Nocker A, Camper AK. Novel approaches toward preferential detection of viable cells using nucleic acid amplification techniques. FEMS Microbiol Lett. 2009;291(2):137-42.
65. Li R, Tun HM, Jahan M, Zhang Z, Kumar A, Dilantha Fernando WG, et al. Comparison of DNA-, PMA-, and RNA-based 16S rRNA Illumina sequencing for detection of live bacteria in water. Scientific reports. 2017;7(1):5752-.
66. Lipkin WI. Microbe hunting. Microbiology and molecular biology reviews : MMBR. 2010;74(3):363-77.
67. Granerod J, Cunningham R, Zuckerman M, Mutton K, Davies NW, Walsh AL, et al. Causality in acute encephalitis: defining aetiologies. Epidemiol Infect. 2010;138(6):783-800.
68. Fischbach MA. Microbiome: Focus on Causation and Mechanism. Cell. 2018;174(4):785-90.
69. Langelier C, Kalantar KL, Moazed F, Wilson MR, Crawford ED, Deiss T, et al. Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. Proc Natl Acad Sci U S A. 2018;115(52):E12353-e62.
70. Singh VP, Proctor SD, Willing BP. Koch's postulates, microbial dysbiosis and inflammatory bowel disease. Clin Microbiol Infect. 2016;22(7):594-9.
71. 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. Scientific reports. 2016;6:23532.
72. Grumaz S, Stevens P, Grumaz C, Decker SO, Weigand MA, Hofer S, et al. Next-generation sequencing diagnostics of bacteremia in septic patients. Genome Medicine. 2016;8(1):73.
73. Fredricks DN, Relman DA. Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. Clin Microbiol Rev. 1996;9(1):18-33.
74. Lipkin WI. The changing face of pathogen discovery and surveillance. Nature reviews Microbiology. 2013;11(2):133-41.
75. Cuthbertson L, Rogers GB, Walker AW, Oliver A, Hafiz T, Hoffman LR, 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(8):3011-6.
76. Cardona S, Eck A, Cassellas M, Gallart M, Alastrue C, Dore J, et al. Storage conditions of intestinal microbiota matter in metagenomic analysis. BMC microbiology. 2012;12:158-.
77. 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. Briefings in Bioinformatics. 2018.
78. Balloux F, Brønstad Brynildsrud O, van Dorp L, Shaw LP, Chen H, Harris KA, et al. From Theory to Practice: Translating Whole-Genome Sequencing (WGS) into the Clinic. Trends in Microbiology. 2018;26(12):1035-48.
79. Miller RR, Montoya V, Gardy JL, Patrick DM, Tang P. Metagenomics for pathogen detection in public health. Genome Medicine. 2013;5(9):81.
80. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis Ca, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. Clinical Chemistry. 2003;49:7-18.
81. Ambrose HE, Granerod J, Clewley JP, Davies NW, Keir G, Cunningham R, et al. Diagnostic strategy used to establish etiologies of encephalitis in a prospective cohort of patients in England. J Clin Microbiol. 2011;49(10):3576-83.
82. Shaw L, Ribeiro ALR, Levine AP, Pontikos N, Balloux F, Segal AW, 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(5):e01237-17.
83. Lassalle F, Spagnoletti M, Fumagalli M, Shaw L, Dyble M, Walker C, et al. Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. Molecular Ecology. 2018;27(1):182-95.
84. Shaw LP, Bassam H, Barnes CP, Walker AS, Klein N, Balloux F. Modelling microbiome recovery after antibiotics using a stability landscape framework. The ISME journal. 2019;13(7):1845-56.

85. 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. Journal of Clinical Microbiology. 2009;47(1):106-10.
86. #039, Sullivan DM, Laver T, Temisak S, Redshaw N, Harris KA, et al. Assessing the Accuracy of Quantitative Molecular Microbial Profiling. International Journal of Molecular Sciences. 2014;15(11):21476-91.
87. 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. Applied Microbiology and Biotechnology. 2018;102(1):403-11.
88. Huggett JF, Laver T, Tamisak S, Nixon G, O'Sullivan DM, Elaswarapu R, et al. Considerations for the development and application of control materials to improve metagenomic microbial community profiling. Accreditation and Quality Assurance. 2013;18(2):77-83.
89. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome. 2018;6(1):226.
90. Wilson MR, O'Donovan BD, Gelfand JM, Sample HA, Chow FC, Betjemann JP, et al. Chronic Meningitis Investigated via Metagenomic Next-Generation Sequencing. JAMA Neurol. 2018;75(8):947-55.
91. Zinter MS, Mayday MY, Ryckman KK, Jelliffe-Pawlowski LL, DeRisi JL. Towards precision quantification of contamination in metagenomic sequencing experiments. Microbiome. 2019;7(1):62.
92. Miller S, Naccache SN, Samayoa E, Messacar K, Arevalo S, Federman S, et al. Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. Genome Res. 2019;29(5):831-42.
93. 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 Research Notes. 2012;5(1):378.
94. Wilson MR, Fedewa G, Stenglein MD, Olejnik J, Rennick LJ, Nambulli S, et al. Multiplexed Metagenomic Deep Sequencing To Analyze the Composition of High-Priority Pathogen Reagents. mSystems. 2016;1(4).
95. Paez-Espino D, Pavlopoulos GA, Ivanova NN, Kyrpides NC. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. Nat Protoc. 2017;12(8):1673-82.
96. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016;26(12):1721-9.
97. Pereira MB, Wallroth M, Jonsson V, Kristiansson E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. BMC Genomics. 2018;19(1):274.
98. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12):550.
99. Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. Standards in genomic sciences. 2015;10:18.
100. Breitwieser FP, Pertea M, Zimin AV, Salzberg SL. Human contamination in bacterial genomes has created thousands of spurious proteins. Genome Res. 2019;29(6):954-60.
101. Fancello L, Raoult D, Desnues C. Computational tools for viral metagenomics and their application in clinical research. Virology. 2012;434(2):162-74.
102. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Research. 2017;27(4):626-38.
103. Quince C, Delmont TO, Raguideau S, Alneberg J, Darling AE, Collins G, et al. DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biology. 2017;18(1):181.
104. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, curated metagenomic data through ExperimentHub. Nature Methods. 2017;14:1023.
105. Lan R, Reeves PR. Escherichia coli in disguise: molecular origins of Shigella. Microbes and Infection. 2002;4(11):1125-32.
106. Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. Nature. 2016;533:543.

107. Konstantinidis KT, Rosselló-Móra R. Classifying the uncultivated microbial majority: A place for metagenomic data in the Candidatus proposal. Systematic and Applied Microbiology. 2015;38(4):223-30.
108. Taylor JW. One Fungus = One Name: DNA and fungal nomenclature twenty years after PCR. IMA Fungus. 2011;2(2):113-20.
109. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017;8(1):1784.
110. Kelly BJ, Gross R, Bittinger K, Sherrill-Mix S, Lewis JD, Collman RG, et al. Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. Bioinformatics (Oxford, England). 2015;31(15):2461-8.
111. La Rosa P, Zhou Y, Sodergren E, Weinstock G, D. Shannon W. Hypothesis Testing of Metagenomic Data. 2015. p. 81-96.
112. Li H, Gao H, Meng H, Wang Q, Li S, Chen H, et al. Detection of Pulmonary Infectious Pathogens From Lung Biopsy Tissues by Metagenomic Next-Generation Sequencing. Front Cell Infect Microbiol. 2018;8:205.
113. Debelius J, Song SJ, Vazquez-Baeza Y, Xu ZZ, Gonzalez A, Knight R. Tiny microbes, enormous impacts: what matters in gut microbiome studies? Genome biology. 2016;17(1):217-.
114. Sze MA, Schloss PD. Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. mBio. 2016;7(4):e01018-16.
115. Odintsova V, Tyakht A, Alexeev D. Guidelines to Statistical Analysis of Microbial Composition Data Inferred from Metagenomic Sequencing. Current issues in molecular biology. 2017;24:17-36.
116. Vogtmann E, Hua X, Zeller G, Sunagawa S, Voigt AY, Hercog R, et al. Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. PloS one. 2016;11(5):e0155362.
117. Burd EM. Validation of laboratory-developed molecular assays for infectious diseases. Clin Microbiol Rev. 2010;23(3):550-76.
118. Greninger AL. The challenge of diagnostic metagenomics. Expert Rev Mol Diagn. 2018;18(7):605-15.
119. Miller S, Naccache S, Samayoa E, Messacar K, Arevalo S, Federman S, et al. Laboratory Validation of a Clinical Metagenomic Sequencing Assay for Pathogen Detection in Cerebrospinal Fluid. bioRxiv. 2018:330381.
120. Relman DA. Actionable sequence data on infectious diseases in the clinical workplace. Clinical chemistry. 2015;61(1):38-40.
121. Ward DV, Scholz M, Zolfo M, Taft DH, Schibler KR, Tett A, et al. Metagenomic Sequencing with Strain-Level Resolution Implicates Uropathogenic E. coli in Necrotizing Enterocolitis and Mortality in Preterm Infants. Cell Rep. 2016;14(12):2912-24.
122. 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 [Internet]. 2019 2019/03//; 8(12). Available from:
http://europepmc.org/abstract/MED/30938709
http://europepmc.org/articles/PMC6430326?pdf=render
http://europepmc.org/articles/PMC6430326
https://doi.org/10.1128/MRA.01687-18.
123. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nature Biotechnology. 2019;37(7):783-92.
124. Ruppe E, Schrenzel J. Messages from the third International Conference on Clinical Metagenomics (ICCMg3). Microbes Infect. 2019.
125. Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJ, et al. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci U S A. 2015;112(22):E2930-8.