TECHNIQUES, TECHNOLOGY & ANALYSIS

A clinician's guide to microbiome analysis

Marcus J. Claesson[1,2], Adam G. Clooney[1–3]* and Paul W. O'Toole[1,2]*

Abstract | Microbiome analysis involves determining the composition and function of a community of microorganisms in a particular location. For the gastroenterologist, this technology opens up a rapidly evolving set of challenges and opportunities for generating novel insights into the health of patients on the basis of microbiota characterizations from intestinal, hepatic or extraintestinal samples. Alterations in gut microbiota composition correlate with intestinal and extraintestinal disease and, although only a few mechanisms are known, the microbiota are still an attractive target for developing biomarkers for disease detection and management as well as potential therapeutic applications. In this Review, we summarize the major decision points confronting new entrants to the field or for those designing new projects in microbiome research. We provide recommendations based on current technology options and our experience of sequencing platform choices. We also offer perspectives on future applications of microbiome research, which we hope convey the promise of this technology for clinical applications.

Microbiome
The collection of microbial genomes at a given site.

Biomarkers
A measurable indicator of disease, pharmacological response or normal biological function.

[1]School of Microbiology, University College Cork, Western Road, T12 Y337 Cork, Ireland.
[2]APC Microbiome Institute, University College Cork, Western Road, T12 Y337 Cork, Ireland.
[3]Department of Biological Sciences, Cork Institute of Technology, Rossa Avenue, Bishopstown, T12 P928 Cork, Ireland.

Correspondence to P.W.O.T. and M.J.C.
pwotoole@ucc.ie; m.claesson@ucc.ie

**These authors contributed equally to this work.*

doi:10.1038/nrgastro.2017.97
Published online 9 Aug 2017; corrected online 11 Aug 2017

For practising gastroenterologists, microbiology has traditionally been the remit of the infectious diseases department. However, over the past few decades, gut microbiota have been increasingly implicated in gastrointestinal disorders such as IBD[1] and IBS[2], adding impetus for more research on the gut microbiome. The outcomes of a number of pioneering studies that investigated gut microbiota in individuals or larger cohorts led to the realization that variation in the composition and function of the gut microbiota can also be correlated with conditions including atherosclerotic disease[3], metabolic disease[4] and colorectal cancer[5]. The strongest links between altered microbiome function and human health or disease arguably involve bacterial metabolites but, even in those cases, most of the evidence relies on correlation between an altered microbiota and pathophysiology, rather than a demonstrated microbial causative mechanism. Thus, the young field of microbiome research is still in a phase of investigating the earliest examples, and the growing number of pathophysiologies, with a microbiome connection.

Regardless of whether the altered microbiome is a cause or consequence of disease, or more likely an environmental risk factor or disease modulator, it is becoming clear that the microbiome provides biomarkers that could be tested for risk or presence of disease. Thus, even with the mechanisms remaining elusive, these biomarkers might still have great diagnostic or prognostic value. The clinician, whether engaged in clinical practice or clinical research, is increasingly required to have a working knowledge of what the microbiota are, how they are measured, what available data can and cannot reveal, and what the short-to-medium term applications of microbiota information are. To be in a position to do this, the clinician must have a working knowledge of bioinformatics terminology — computational analysis of biological data — and ideally a grasp of the rudiments of phylogenetics, the system for naming and grouping organisms based on their evolutionary history and relatedness. The gut microbiota are now being added to the list of readouts and clinical metadata being collected from existing or newly initiated cohort studies. Thus, it is also desirable for new entrants to microbiota research to appreciate the choices of technological approach and their respective limitations; topics which this Review will cover succinctly. We also share our experiences of designing experimental approaches and choosing sequencing platforms, to arm new practitioners for deciding how to implement current and future applications of this technology.

Terminology

Harmonized nomenclature prevents confusion and is particularly desirable in scientific discussions. Here, we use the term microbiota to refer to the collection of microorganisms in a specified location or sample, microbiome to refer to their collective coding capacity, and metagenome to refer to the experimentally

Key points

- Complex communities of microorganisms live on and in the human body, and variations in the composition and function of these communities are increasingly linked to various conditions and diseases
- Although it is not known if microbiome changes are causative or consequential in most pathophysiologies, they might provide biomarkers for disease detection or management
- Microbiome analysis is likely to become a routine component of secondary health care and is emerging as a modifiable environmental risk factor in multifactorial diseases that could be targeted by novel therapeutics
- Technology advancements are leading to a range of powerful methods for microbiome analysis becoming available and affordable for clinical studies
- Judicious choice of sample type and sequencing platform are required to maximize the clinical utility of microbiome data

Bioinformatics
The use of computer science, statistics and mathematics to analyse and interpret biological processes and molecular components.

Phylogenetics
Evolutionary relationships between organisms, genes or proteins.

Metagenome
The collective microbial genomes and genes in an environment or sample.

Shotgun sequencing
All extracted DNA is randomly sheered into desired fragment sizes for high-throughput sequencing, as opposed to targeting a specific marker gene.

Amplicon
A target gene or sequence that is amplified naturally or artificially.

Copy number
The number of copies of a particular section of DNA; some organisms have multiple copies of a targeted gene.

Taxa
A population of phylogenetically related organisms.

16S ribosomal RNA gene
A gene located in the 30S subunit of a prokaryotic ribosome, which contains nine variable regions that can be targeted for amplification and used for microbial taxonomic profiling of a sample.

18S ribosomal RNA gene
A gene located in the 40S ribosomal subunit found in eukaryotic cells, targeted in the analysis of fungal communities.

determined dataset from shotgun sequencing the genomes of microorganisms in a particular sample. However, we acknowledge that other authors use these terms differently, particularly whereby microbiome refers to the total microbial community present[6]. The most commonly used terms are also explained in the Glossary. The majority of studies of the gut microbiota currently focus on bacteria and essentially ignore other microorganisms present, but terminology does exist to describe the burgeoning studies of viruses (the virome), bacterial viruses called bacteriophages (the phageome) and fungi (the mycobiome).

Study design

An obvious stage in establishing a new project involving microbiome analysis is study design. Some of the parameters might be fixed, based on the hypothesis being tested, such as changes in the microbiome (longitudinal analysis) or whether microbiome differences correlate with clinical phenotypes (cross-sectional or cohort analysis). If testing the effect of a treatment (drug, dietary ingredient, physiological intervention), the magnitude of potential changes in the microbiome must be predicted at the appropriate phylogenetic level to enable statistical power calculations to be performed. Particular care must be given to confounding factors: BMI, age and pregnancy all influence the gut microbiota[7] and should be recorded and considered when selecting case-matched controls. Antibiotics also have a large effect on the human microbiome[8,9] and so an exclusion criterion of antibiotic use within 6 months is often used.

Another feature of the study design is whether to perform marker gene analysis or metagenomic sequencing. Marker gene analysis differs from metagenomic shotgun sequencing as it is based on targeting an amplicon of only one gene instead of attempting to sequence all or most genes in a sample. This type of analysis is very convenient for taxonomic classification and basic sample or cohort comparisons, while also being cheaper than metagenomics owing to an overall lower coverage that results in many more samples being sequenced on one instrument run. However, balancing these advantages are a number of limitations, including low-resolution at species level[10], issues with varying copy number[11] (microorganisms, in particular, might have multiple copies of the targeted gene) and difficulty in detecting low-abundant taxa, although low-abundant taxa are also potentially missed in shotgun sequencing data of insufficient coverage. However, for initial studies of the gut microbiota, this type of analysis is very powerful and offers important information on both microbial diversity and composition.

The marker gene most commonly used in prokaryotes is the ubiquitous 16S ribosomal RNA (rRNA) gene, which consists of nine variable regions, each flanked by highly conserved DNA that provides ideal primer sites for amplification[12]. Single or multiple adjacent variable regions can be sequenced, enabling the classification of prokaryotes and diversity analysis. Unfortunately, the lack of standardized usage of variable regions, primers and amplification parameters leads to substantial variation in results, even from the same sample. Thus, caution is warranted and comparisons between and within studies should only be trusted when the exact same experimental protocols have been adopted[13,14]. For fungi and other single-cell eukaryotes, the internal transcribed spacer (ITS) region and 18S ribosomal RNA gene are the preferred marker genes[15].

Sample types and their collection

Studying the microbiome of the gastrointestinal tract can involve sampling from a wide array of sites and so the sampling plan is ultimately dependent on the hypothesis in question. For example, the microbiome of the mucosal layers is generally accessed through pinch biopsy samples during endoscopy, although luminal brushing[16,17] and submucosal sampling[18] are sometimes applied to acquire more superficial and deeper sampling of tissues, respectively. The latter method requires manual excision of tissue biopsy samples. Colonic lavage is a proxy for biopsies, in which liquid remaining after the bowel preparation is aspirated through the colonoscope[19]. Rectal swabbing is a more targeted technique that can be applied outside the clinic[20] (see TABLE 1 for advantages and disadvantages of sampling types).

Notably, the bowel cleansing often associated with colonoscopy has been shown to affect the microbiota of both faecal and mucosal (biopsy) samples[21]. The direction and magnitude of change is known to vary depending on sample type and disease states, with bowel preparation both reducing microbiota diversity as well as introducing compositional differences in the mucosa and lumen microbiota in patients with IBD[21]. Information on bowel preparation should, therefore, be incorporated into the study design to facilitate statistical adjustments.

Faecal samples are the most convenient collection method for gut microbiota samples and often provide opportunities for large-scale and longitudinal studies. They have long been regarded as an acceptable proxy for distal gut microbiota, partly owing to their noninvasive collection nature and suitability over biopsy samples for accessible biomarker discovery. Although some studies have shown gradual microbiota changes along the length of the colon[22] and in stool, the intraindividual

Table 1 | **Advantages and disadvantages of sample types for gut microbiome analysis**

Sample	Advantages	Disadvantages
Faecal sample	Noninvasive; no bleeding or discomfort; no bowel cleansing; easier to sample frequently	A proxy for the gut microbiome; might contain dead bacteria and/or bacteria from unspecified gastrointestinal tract compartments; less controlled sampling variables
Luminal brush	Captures host–microbe interactions; increased mucosal coverage; no bleeding; greater proportion of bacterial to host DNA than biopsies	Requires endoscopy; less biomass for host studies; affected by bowel cleansing
Rectal swab	No bleeding; greater proportion of bacterial to host DNA than biopsies; no bowel cleansing; can be administered at home; easier to sample frequently	No visual aid to pinpoint areas of interest; limited biomass for host studies; more discomfort than stool sampling; potential contamination with skin bacteria
Colonic lavage	Provides more DNA than biopsy samples; no bleeding	A proxy for the gut microbiome; requires endoscopy; affected by bowel cleansing
Pinch biopsy	Captures host–microbe interactions; can target exact areas of interest	Requires endoscopy; disrupts epithelium; affected by bowel cleansing
Sub-mucosal biopsy	Captures host–microbe interactions and bacterial translocations through epithelial layers; can target exact areas of interest	Requires endoscopy; disrupts epithelium; requires extensive sample processing; affected by bowel cleansing

changes are generally smaller than the interindividual differences[23]. If investigating a new clinical question, the difference between faecal and mucosal microbiota should be kept in mind, especially if aiming at microbiome-related diagnostics or biomarkers. For instance, faecal samples are more suitable for frequent monitoring and biomarker screening, whereas biopsy samples could be more meaningful than faecal samples for patients undergoing diagnostic endoscopy. Intestinal biopsy samples are preferred when investigating host–microorganism interactions, due to the requirement of obtaining associated host DNA or RNA and when visual gastroenterological assessment is required.

Compared with samples acquired in the clinic, many more potential variables such as time of day collected and storage regimes are associated with the sampling of stool, which often happens at home under varying conditions. Appropriate storage is crucial to preserve the integrity of the microbial DNA and RNA. Unfortunately, recommended storage conditions often vary, with transfer to −20 °C suggested within 15 min[24] to 24 h of defecation[25], before long-term storage at −80 °C. Some evidence suggests long-term storage at −80 °C can affect microbiota composition, in particular, increasing the Firmicutes to *Bacteroides* ratio[26], although this finding has not been reproducible in all cases[27] and remains controversial. A general recommendation for large-scale studies in which immediate DNA extraction is not feasible is to adhere to a consistent freezing protocol[28]. However, storage preservation kits[29,30], including OMNIgene (DNA Genotek, Canada) and faecal occult blood test cards, along with preservative buffers such as RNAlater (Invitrogen, USA)[31], can extend the room temperature restriction period beyond a week, which is useful in situations in which prompt sample delivery to the laboratory is difficult[32]. The latter buffer also has the advantage of providing high yields of host and microbial RNA in a wide range of samples[33]. All samples should be collected and stored in the same manor to minimize bias.

Extraction of nucleic acids

The DNA and/or RNA extracted from collected samples must be of sufficient quantity, of high quality, and must contain a faithful representation of the microbial community present in the sample. Nucleic acid extraction from most human body sites including the gut is a complex procedure owing to the occasionally high proportion of (unwanted) host DNA or RNA, along with other substances including food and cellular metabolites. Unfortunately, no gold-standard method suitable for all sample and cell types exists.

Most extraction methods are based on chemical and mechanical cell lysis through the use of buffers followed by washing and elution of pure DNA or RNA from an immobilization matrix[34]. Several studies involving both mock microbial communities (that is, artificially constructed from DNA or cells and of known composition) and human faecal samples provide evidence supporting the inclusion of a mechanical lysis step (for example, by mechanical bead beating) to yield a more accurate representation of the microbiota composition[35–37]. A wide variety of commercial extraction kits are available for DNA or RNA extractions, which are recommended for new entrants to the field because they offer documented reproducibility and stringent quality control. Nonetheless, numerous studies have reported substantial differences in DNA yield and apparent microbiota composition when different extraction kits were compared[38–43]. Notably, even commercial kit reagents have been demonstrated to contain contaminating bacterial DNA, which could be a confounding issue for low-biomass tissue biopsy samples[44,45]. Thus, it is advisable, at a minimum, to bioinformatically screen results for known contaminants, such as environmental microorganisms or those present in reagents[45,46]. Although all known and unknown causes and consequences of nucleic acid extraction differences are not resolved to date, it is of utmost importance to consistently apply a single suitable extraction (and sampling) protocol throughout a study, and ideally across multiple studies with which comparisons are desired.

Composition versus function

In addressing biological questions, sequence-based microbiome studies are either aimed at investigating microbiota composition or function. Compositional studies generally involve comparing amplicon sequences of the 16S rRNA gene, which is ubiquitous in prokaryotes. The 16S rRNA gene is composed of regions of conserved sequence (near-identical across most bacteria) and of variable sequence or regions, which are phylogenetically distinct for a particular genus and species. Sequencing amplicon pools derived from one or more variable regions reveals 'who is there' in terms of relative abundances of bacterial taxa, along with comparisons of alpha diversity (within-sample diversity; one value per sample) and beta diversity (between-sample diversity; pairwise values for all sample combinations). Although this method is affected by PCR bias and uneven 16S rRNA gene copy numbers across different bacterial species, it is generally 20–30 times cheaper than shotgun metagenomic sequencing per sample, even excluding the additional computational costs and requirement for a skilled bioinformatician for performing metagenomic analyses[14].

On the other hand, metagenomic shotgun sequencing, in which the total extracted DNA is fragmented and randomly sequenced, can answer 'what can they do', as it reveals the encoded functions of the sequenced microbial DNA. The phylogenetic origins of microorganisms can be determined from shotgun sequencing data analysis by comparison with previously annotated genes, thereby offering data that is similar to compositional information derived from 16S rRNA gene sequencing, although this process is reliant on the completion and accuracy of databases. This type of analysis does, however, require high levels of expertise, computational overheads and high sequencing costs. Even more complex and expensive is metatranscriptomics, whereby high-throughput cDNA sequencing (so-called RNAseq) is utilized to sequence transcribed microbial genes (mRNA), answering 'what are they doing'. This method is very challenging for intestinal biopsy samples because microbial mRNA is vastly under-represented compared with host tissue. Metatranscriptomics has, therefore, been applied mainly to the high biomass available in stool samples, albeit with questionable interpretational value because of the long bowel transit time that potentially affects RNA integrity, which is less stable than DNA[47]. Thus, owing to the shorter half-life of RNA, such analysis will also reflect the bacterial response to being voided from the colon. As a so metimes useful intermediate, rRNA (the transcript, not its corresponding gene) can be transcribed to cDNA and amplified. Compositional analysis of these transcript reveals the relative abundances of metabolically active bacteria, thereby answering 'who is active'.

Sequencing platforms

The substantial progress of next-generation sequencing (NGS) technologies in the past decade has broadened and deepened our knowledge of the microbiome. Although a comprehensive review of NGS is beyond the scope of this Review (see elsewhere[48]), it is worth highlighting the major technologies routinely used in microbiome studies. Both Illumina and ThermoFisher have an array of instruments with varying capabilities and costs. Ion PGM and especially Illumina MiSeq are often used for amplicon sequencing as their longer read lengths (300–400 bp) cover several variable regions of the 16S rRNA gene, which improves the accuracy of phylogenetic assignment. Shotgun sequencing instead requires increased sequencing reads to cover larger genomic regions and is more suitably performed by other approaches. For example, the Illumina HiSeq and Ion Proton can deliver an output of up to 750 gigabase (Gb) (150 bp read length; $22/Gb) and 10 Gb (200 bp read length; $80/Gb) of sequence per instrument run, respectively[48]. Newer technologies developed by Pacific BioScience (20 kb read legnth; $1000/Gb) and Oxford Nanopore (≤20 kb; $750/Gb) offer amplification-free sequencing with longer reads, albeit with lower throughput, to date, and increased error rates that, in some cases, can be mitigated through applying various algorithms.

Bioinformatic analysis

Microbiome data analysis is usually tailored to the question and data type and there are many methods and protocols, even for analysing the same type of data. A large repertoire of free and open-source software is available for the various analytical steps, from quality-filtering of the raw data to the final stages of visualizing results. The application of these methods is largely dependent upon the study design and dataset, the most notable constraint being marker gene amplicons or shotgun metagenome sequencing (see FIG. 1 for an outline of bioinformatic analysis). The value of a study is ultimately circumscribed by the quality of the data to be analysed so it is essential to quality check and assess all sequencing reads before proceeding with the downstream analysis and interpretation, as low-quality reads will inhibit the assembly of contigs and reduce annotation efficiency. A key starting point is the assessment of the data quality. Software tools such as FastQC can be applied to assess the overall quality of the sequencing runs, while Trimmomatic[49] and tools in the FASTX-Toolkit can be used to filter and trim sequencing reads based on PHRED quality scores, by setting desired lengths or thresholds.

Metagenomic shotgun sequencing

Assembly. A valuable feature of shotgun sequencing is that it enables assessment of the functional potential of the microbiome because (ideally) a majority of the genetic material is sequenced, as opposed to targeting one particular gene. After sequencing, shorter read fragments are often assembled into longer continuous sequences (contigs; see FIG. 2), potentially covering complete protein coding genes and operons. Metagenome assembly also permits simpler analysis by decreasing the number of overall sequences (reads are collapsed into contigs), whose abundance can still be quantitatively analysed.

Assemblies are either based on reference genomes[50] or performed *de novo*[51,52]. Reference-based assemblers are less computationally demanding and map sequencing reads onto reference genomes to construct contigs joined *in silico* from individual reads. However, they

Alpha diversity
Microbiota diversity within an individual site or sample diversity; one value per sample.

Beta diversity
Intervariability, diversity between separate samples.

PCR bias
Unequal amounts of amplification across DNA sequences that leads to a skewed distribution of PCR products.

Metatranscriptomics
The study of RNA copies of the collective microbial genes in a community or sample.

Assembly
The process in which short DNA fragments are aligned and merged to form longer DNA fragments.

Contigs
Contiguous DNA sequences assembled from shorter, overlapping sequencing reads.

Annotation
Assigning functions or functional categories to gene or protein.

PHRED quality scores
A measure of the quality of base calling in a sequenced strand of DNA.

de Bruijn graphs
Consist of nodes (*k*-mers) and edges (overlaps between *k*-mers). The graph is constructed using *k*-mer overlaps leading to an assembled sequence.

Scaffolds
The product of aligning and merging contigs to form longer continuous DNA sequences.

Binning
Grouping DNA sequences based on particular attributes such as GC content or similarity with other genes.

k-mer
Short DNA sequence with fixed length *k*.

Homology
Shared ancestry or degree of relationship between sequences or genes.

are limited by the quality and availability of reference genomes and rely on previously characterized species. *De novo* methods, on the other hand, do not require reference genomes and instead use graph theory algorithms such as de Bruijn graphs to assemble sequencing reads[53]. Metagenome assembly was initially carried out using single-genome assemblers[54], but now tools have been adapted to handle multiple genomes with varying abundances. Such tools include MetaVelvet[51], IDBA-UD[55] and SOAPdenovo2 (REF. 52). MetaVelvet, and its latest update MetaVelvet-SL[56], increases accuracy by using a streamlined assembly approach, while also orientating the contigs into longer assemblies called scaffolds.

Sequence binning. Most metagenomic shotgun tools are aimed at phylogenetic binning, which is the process of grouping sequencing reads or assembled contigs by their likely host genomes and subsequently assigning taxonomy, much like grouping words in a sentence into verbs, nouns or adjectives. Binning is performed either through similar DNA compositions, such as distinguishable nucleotide patterns like *k*-mer (DNA sequence with fixed length *k*) frequency or GC content, or by gene homology (sequence similarity to known genes). Each method has its limitations; compositional-based approaches perform poorly on short reads, whereas homology-based approaches struggle to be discriminative when the dataset contains many closely related species. Some of the most frequently used binning software are composition-based and include TETRA[57], PhyloPythiaS[58] and Kraken[59], which compares unique distributions of *k*-mers across the metagenomic sequences. In contrast, the homology-based MetaPhlAn2[60] software uses a set of core marker genes to differentiate between microbial taxa. A hybrid approach (implemented by PhymmBL[61], MetaCluster-TA[62], AMPHORA2 (REF. 63) and MaxBin[64]) can also be performed, combining DNA compositional and homology-based methods. However, considerable differences between the performance of binning tools have been observed[14,65] and for some datasets, CLARK[66]

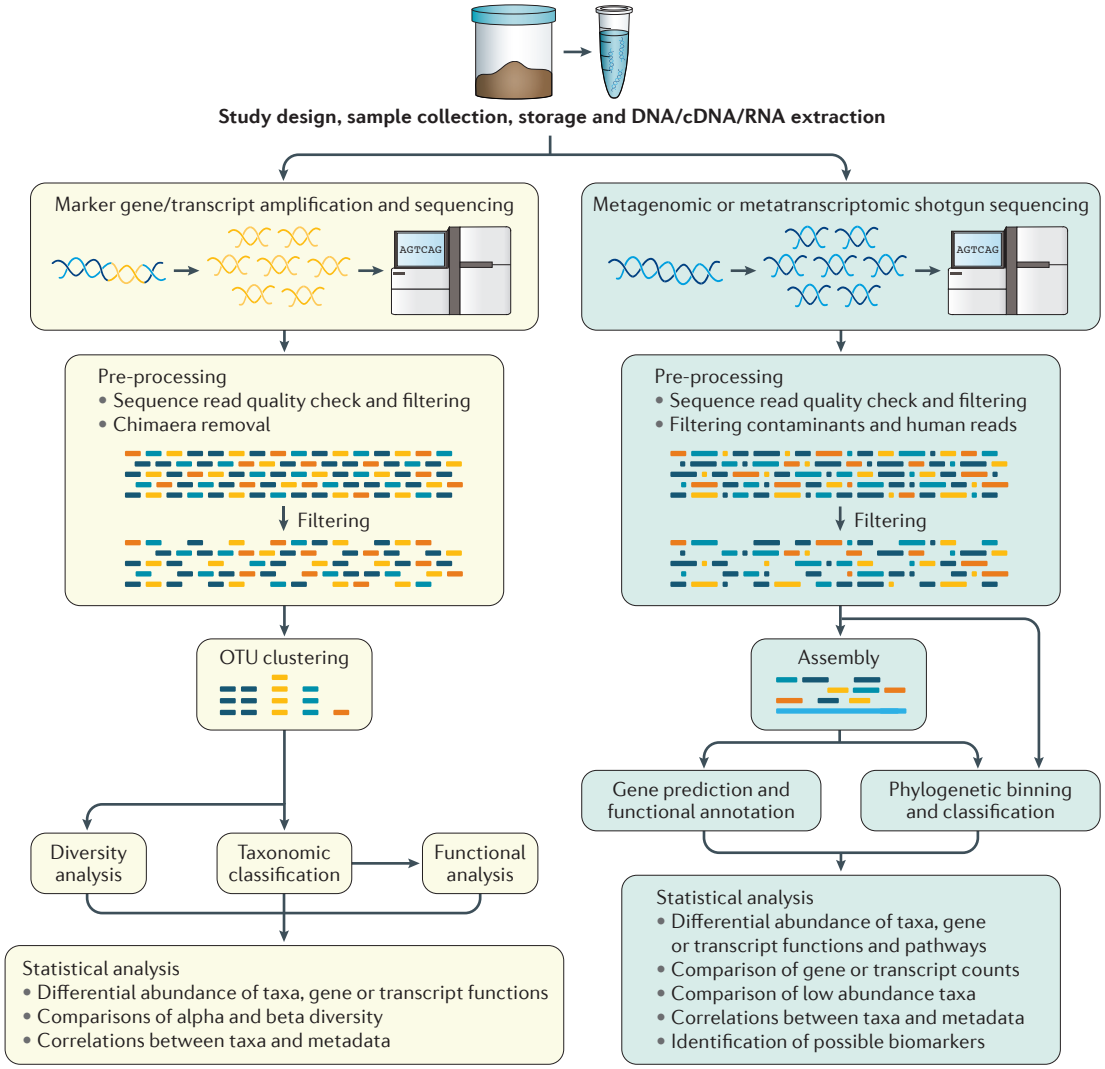

Figure 1 | **Flowchart of the major steps involved in bioinformatic analysis of the microbiome.** The analysis is divided into two sections depending on the type of sequencing. This schematic describes the basic steps and might vary depending on the aim of the analysis. OTU, operational taxonomic unit.

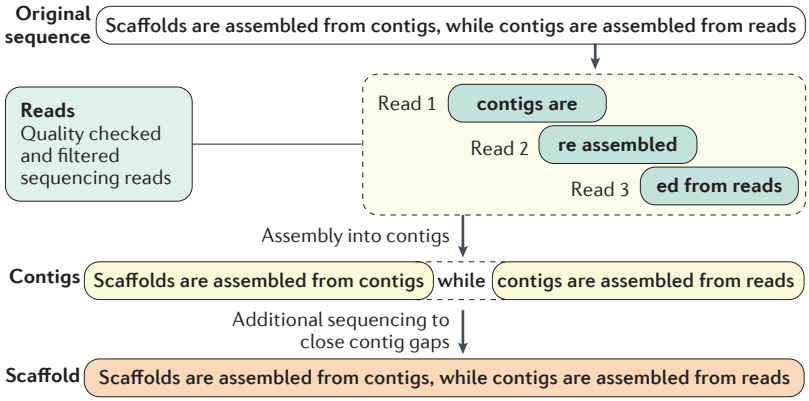

Figure 2 | **Sequence read assembly.** A mock example explaining bioinformatic sequence assembly along with the terms sequence, reads, contigs and scaffolds.

and Kraken were the best performers[67]. The accuracy of binning methods will increase as new sequencing technologies offer longer reads combined with a greater number of reference genomes becoming available in databases.

Functional analysis. Once a metagenome is assembled, genes within the assembled contigs are then identified, which is a prerequisite for deducing the functional potential (annotation) of a microbiome. Although a number of tools exist for this task, such as MetaGeneMark[68], FragGeneScan[69] and MetaProdigal[70], the computational challenges they address (with variable success) are greater than for single genomes. As the metagenome assemblies are generally of poorer quality than for single genomes, even when performed at an optimal level, many genes still remain fragmented and incomplete. Following gene identification (gene calling), functional annotation is carried out using computationally demanding homology-based searches (often BLAST[71]) against databases of orthologues (EggNOG[72], COG or KOG[73]), enzymes (KEGG[74]) and protein domains and families (Pfam[75], TIGRFAMs[76], InterPro[77]).

Owing to the vast number of possible combinations of tools, along with their various computational prerequisites, a number of publicly available pipelines have been developed. These tools have particular concatenations of methods or software with predefined parameters, and include MG-RAST[78], IMG/M, CAMERA[79] and the European Bioinformatics Institute's Metagenomics[80] service. These pipelines provide quality-filtering, gene calling, functional annotation, basic statistics and visualization in easy to use online platforms. Although these approaches suit many users without relevant experience in command-line usage of bespoke methods, they have a level of less-controlled analysis and the services often become user-saturated.

Marker gene analysis

A number of pipelines and software packages are available for analysing 16S rRNA gene amplicons, most notably QIIME[81], Mothur[82], and UPARSE[83], which are all highly flexible and versatile. One study concluded that QIIME, Mothur and MG-RAST produced comparable results when analysing stool samples, and similar microbial composition was found irrespective of the pipeline chosen[84]. A typical pipeline starts with quality-filtering and de-multiplexing of raw sequencing reads into appropriate sample bins. This process is possible as barcode sequences of short DNA sequences, unique to each sample, have been added to the primers before amplification. A typical approach following this step is to cluster reads into operational taxonomic units (OTUs), which are grouped on the basis of DNA sequence similarity. A 97% similarity level is usually applied as this is considered suitable for binning reads derived from the same species. The main methods for this are *de novo*, closed-reference and open-reference clustering, each with their own set of advantages and disadvantages. *De novo* methods cluster reads without the use of a reference database, whereas the opposite is true for the closed-reference clustering (exclusively reference based). The latter method is much faster because reads with no alignment to the database are discarded. However, some evidence suggests *de novo* clustering is the optimal method[85]. Open-reference clustering combines both *de novo* and closed reference methods, in which any reads discarded during closed reference are subjected to *de novo* clustering, leading to some performance improvements[86]. One comparison of older and newer clustering methods found the newer methods superior to the more traditional methods[87].

Another important step in the analysis of marker gene sequences is the removal of chimeric sequences, a known artefact of PCR amplification in which hybrid sequences are produced from unrelated parent sequences. The majority of chimaeras are relatively easy to detect using existing reference databases, and software applications for detection (integrated into the commonly used pipelines QIIME and Mothur) include UCHIME[88] and ChimeraSlayer[89], although all methods might not be 100% effective. QIIME also provides a framework for taxonomic assignment to the created OTUs. The most commonly used classifiers are Mothur[82] and the RDP-classifier[90]. Taxonomic classification requires a reference database of previously classified sequences, such as SILVA[91] (16S and 18S rRNA genes), Greengenes[92] (16S rRNA gene), Ribosomal Database Project[93] (16S rRNA gene) and UNITE[94] (ITS region). These classifiers are very accurate to family taxonomic ranks and are often accurate to genus level with some occasional misclassifications for obligate anaerobes (due to lack of reference sequences) and Enterobacteriaceae (due to highly similar reference sequences from different genera), but lose much of the specificity and accuracy at species level. More dedicated species-classifiers such as SPINGO[95] or UTAX[96] are better options for higher taxonomic resolution.

Sequencing errors can greatly affect OTU clustering performance, and methods removing erroneous reads or alternatives to clustering are desirable[97]. One published alternative to standard OTU clustering is the Divisive Amplicon De-noising Algorithm (DADA2)[98], which corrects for noise and errors in Illumina MiSeq generated sequences. DADA2 also offers a full amplicon workflow including quality filtering, de-replication (the process of consolidating identical sequences for efficiency) and chimaera identification.

Gene calling
Identifying coding regions in a sequence of DNA.

Orthologues
Genes in different species derived from a common ancestral gene following speciation, which usually retain the same function.

Pipelines
A series of tools or scripts optimized for the analysis of a dataset in which the outputs of one step are the inputs for next step.

Barcode sequence
A short series of DNA bases attached to sequence reads, each unique to a sample to enable differentiation after sequencing.

Operational taxonomic units
A collection (cluster) of sequences that are often at least 97% similar to each other and used to classify closely related individuals.

Reference database
A collection of known information (for example, gene sequences or functions) constructed in a format for querying or similarity-based searches.

Chimeric sequences
Artefacts from the PCR process in which an amplified sequence is composed of DNA from two or more parents.

For statistical analysis and visualization, QIIME and MEGAN[99] require a count table (OTUs as rows and their abundances for each sample in columns). However, many other packages and software exist for microbiome-related statistics. A standard count table is used when carrying out analysis in R, the most commonly used statistical programming language. Useful R libraries for such analysis include Vegan, Phyloseq[100] and ggplot2, which all provide methods, ample documentation and vast flexibility for statistical and visual analysis of both metagenomic and marker gene studies.

Finally, as an attempt to bridge the gap between metagenomics and marker gene analysis, PICRUSt[101] and Tax4Fun[102] can assign metabolic functions to 16S rRNA genes based on their read mappings to previously annotated genomes. PICRUSt applies a two-step process involving gene content inference followed by metagenome inference, resulting in estimated abundances of gene families. Although acknowledging the imperfect link between the marker gene and its host genome, PICRUSt and Tax4Fun can be used as exploratory tools to investigate the potential need for more in-depth metagenomic shotgun sequencing.

Actionable recommendations

This section is aimed at new entrants to the microbiome field and distils the information in the preceding sections, along with our own personal experiences over the past decade, into a series of recommendations for starting a new project. These suggestions are necessarily subjective and we concede that alternative choices might lead to equally satisfactory results. For a list of microbiome analysis tools and software, grouped by different categories, see TABLE 2.

Nucleic acid extraction

If there is a chance that studying the host or microbial transcriptome will be required in the future, begin by extracting all nucleic acids (combined RNA and DNA) and store in aliquots. Removal of RNA by ribonuclease treatment of a purposely stored extra aliquot can be used for both 16S rRNA gene amplicon and shotgun metagenomic analysis. For latest optimized methods, consult the International Human Microbiome Standards website.

Microbiota analysis method

Use a parsimonious approach. If the clinical question and the accompanying literature have already identified a microorganism or short-list of microorganisms whose abundance is critical, then a quantitative PCR assay or assays is sufficient and will cost very little to set-up. These assays use established PCR primer pairs (that work well, even though this method is no longer widely used) for groups of microorganisms or high-profile single organisms. In some situations, you might have to design a species-specific or even strain specific primer pair. An NGS approach will usually be required when there is no candidate microorganism whose abundance levels will inform the clinical question being addressed.

Sequencing platform

If the clinical question requires compositional data, as opposed to functional data, the Illumina Nextera chemistry and the 16S rRNA V3–V4 region primers are very widely used, with 2×250 paired-end reads. We multiplex up to 280 samples per 384 well plate or sequencing run, which provides more than ample sequence depth.

Bioinformatics pipeline

For 16S rRNA gene-amplicon-based analysis, start with QIIME or Mothur, which will give you a rudimentary analysis. You or your postgraduate can visit a lab routinely using a QIIME or Mothur-based pipeline and become reasonably proficient in a couple of days. For more advanced statistical analysis involving comprehensive metadata, you might have to pair up with more experienced scientists or research groups. Analysis pipelines that accompany sequencing instruments are very limited in what they can accomplish.

Metagenomics and biostatistics

Unless you are unusually computer literate and can invest a lot of time, shotgun metagenomic analysis is not for beginners working unaided. For shotgun sequence data analysis, involve a collaborator or hire a postdoctoral fellow with published competence in this area. The data is much richer than 16S rRNA gene compositional data, and it is becoming increasingly clear that some questions can only be answered by shotgun sequence analysis, but the costs and data challenges are still formidable. Without a biostatistician who understands multivariate analysis, adjusting for confounders and false discovery rates, it will be very challenging to convert expensive primary datasets into robust high-impact conclusions.

Future perspectives

In a field characterized by rapid progress, making future predictions is difficult, but nevertheless there are trends and technical advances that point towards some likely clinical applications and exploitations of microbiota research.

Microbiota surveillance

A microbiota depleted by multiple rounds of antibiotics enables out-growth of *Clostridium difficile* spores leading to *Clostridium difficile*-associated diarrhoea (CDAD) that might be successfully treated by faecal microbiota transplantation[103]. An altered microbiome is also a feature of diseases such as obesity[104], type 2 diabetes mellitus[4] and frailty in the elderly[105]. We predict that microbiota profiling will become an adjunct to routine health management and disease prevention, and that microbiota-based biomarkers will become clinically informative and widely screened.

Microbiota manipulation

If altered microbiota configurations ultimately prove to be causative of disease states other than CDAD, then these states will become therapy targets. Prime

candidates, those for which the strongest rationale or possible mechanism exist, include type 2 diabetes mellitus (metabolic effects), cardiovascular disease (atherosclerogenic effects), IBD (loss of anti-inflammatory effects), and IBS (outgrowth of fermentative Firmicutes). The next generation of microbiota-based therapeutics are already in development[106] and their regulation is being addressed[107].

Shotgun sequencing

Metagenomics involving shotgun sequencing of clinical samples will probably become routine in some regions, marking the translation of research into routine application[108]. One example is the identification of West Nile virus by shotgun sequencing of cerebrospinal fluid[109]. This approach also has the potential to identify new pathogens.

Systems biology and the microbiome

A striking example in applied microbiome research is the improvement in algorithms for predicting post-prandial glucose response in patients with impaired glucose control[110]. The metabolism of individual microbial genomes can be predicted[111], and combined metabolism from entire microbiota communities can also be integrated into sophisticated models that can be interrogated based on altered conditions[112]. We speculate that it will become possible to predict the interaction of, for example, drugs with the host and the microbiota, treated as a single (computationally combined) genetic and metabolomic entity. Such a scenario will also facilitate developing truly personalized nutrition programmes, combining nutrigenomics data with metagenomics data for each individual and minimizing, for example, the risk of chronic diet-related diseases that have a microbial component.

Table 2 | **Categorization and description of software tools and packages used for published microbiome analysis**

Category	Description	Software
Quality tools	Tools for run assessment, read filtering and trimming to ensure high quality data	FASTQC[113], Fastx-Toolkit[114], Trimmomatic[49], PRINTSEQ[115], NGS QC Toolkit[116], Meta-QC-Chain[117]
Shotgun read assemblers	Software for aligning and merging fragments of DNA to form longer contigs and scaffolds in an attempt to reconstruct an original sequence	SOAPdenovo2 (REF. 52), IDBA-UD[55], Meta-IDBA[118], MetaVelvet[51], MetaVelvet-SL[56], Velvet[119], Ray-Meta[120], MEGAHIT[121], MetaBAT[122], Omega[123], METACAA[124], metaSPAdes[125], MetaORFA[126], MetaAMOS[50]
Shotgun binners or classifiers	• Grouping contigs or reads based on specific features to assign taxonomy • Methods include composition and alignment or a combination of both	MetaPhlAn2 (REF. 60), Kraken[59], GSTaxClassifier[127], Ray Meta[120], mOTU[128], CLARK[66], CLARK-S[129], Amphora2 (REF. 63), TACOA[130], NBC[131], MLTreeMap[132], PhymmBL[61], GOTTCHA[133], CARMA3 (REF. 134), LMAT[135], PhyloPythiaS[58], Taxator-tk[136], MetaClusterTA[62], RITA[137], SORT-Items[138], SPHINX[139], Ralphy[140], WGSQuiker[141], S-GSOM[142], Treephyler[143], TaxSOM[144], ClaMS[145], Genometa[146], Woods[147], DiScRIBinATE[148], MetaCV[149], INDUS[150], MetaBin[151], MetaPhyler[152], TAC-ELM[153], metaBEETL[154], SPANNER[155], DUDes[156], Kaiju[157], MaxBin[158], MetAnnotate[159], MyTaxa[160], Tamer[161], TaxyPro[162], TWARIT[163], WSVDD[164]
Shotgun gene and functional analysis tools	Tools for predicting genes in contigs and assigning functions	MetaGeneMark[68], FragGeneScan[69], MetaProdigal[70], ShotgunFunctionalize R[165], RAMMCAP[166], Glimmer-MG[167], Orphelia[168], MetaGUN[169], Genometa[146], Metaphor[170], MetaPath[171]
Shotgun gene and functional databases	Databases containing functional information for aligning predicted genes	PFAM[75], COG/KOG[73], SEED[172], eggNOG v4 (REF. 173), TIGRFAM[76], KEGG[74]
Shotgun statistical analysis tools	Tools for visualization and analysis of metagenomic shotgun data	HUMAnN2 (REF. 174), LEfSe[175], PPANINI (manuscript under preparation), StrainPhlAn[60], Metastats
Shotgun pipelines	• Software containing all or subsets of steps involved in shotgun metagenomic analysis • Steps include, quality-filtering, assembly, classification and functional assignment along with visualization	MG-RAST[78], MEGAN[99], IMG/MER[176], WEBMGA[177], MetaAMOS[50], EBI Metagenomics[80], METAREP[178], Parallel-META[179], MOCAT[180], GALAXY Portal[181], BIOMaS[182], PHACCS[183], Smashcommunity[184]
Operational taxonomic unit (OTU) picking methods	Methods for clustering reads against each other or aligning reads to a database, or both, to obtain a set of OTUs	Closed Reference, Open Reference, De Novo, Mothur[82], uclust[96], UPARSE[83], DOTUR[185], CD-HIT[186], CROP[187], ESPIRTI-Tree[188], DNA Clust[189], GramCluster[190], M-pick[191], Swarm v2 (REF. 192), oclust[193], MtHc[194]
Chimaera removal	Tools for removing chimeric sequences (reads erroneously consisting of two transcripts)	UCHIME[88], UCHIME2 (preprint), ChimeraSlayer[89], Perseus, CATCh[195]
16S rRNA gene amplicon classifiers	Software to assign taxonomy to 16S marker gene OTUs or reads at each taxonomic level including species for some classifiers	Mothur[82], RDP-Classifier[90], UTAX[71], rTax[196], 16S Classifier[197], SPINGO[95]
Amplicon databases	Databases containing sequences used for assigning taxonomy to OTUs or reads through alignment	SILVA[91], Greengenes[92], RDP[93], rrnDB[198], PhylOPDb[199], HITdb[200], Unite[94]
16S rRNA gene statistical analysis tools	Tools and software packages for the analysis and statistical comparisons of 16S marker gene datasets	Mothur[82], QIIME[81], UniFrac[201], PICRUSt[101], Tax4Fun, Phyloseq[100], LEfSe[175], MaAslin[202], MetagenomeSeq[203], CopyRighter[204], OTUbase[205], Metastats
16S rRNA gene amplicon pipelines	A package of tools and commands contained in a pipeline to analyse 16S marker gene data from raw data to visualization	QIIME[81], Mothur[82], SILVA[91], Megan[99], FASTGroup2 (REF. 206), PANGEA[207], CLOTU[208], Jaguc[209], DADA2 (REF. 98), MICCA[210], FunFrame[211]

Conclusions

Technological advancements in culture-independent methods, NGS and bioinformatics tools have led to the unlocking of the human microbiome. This impressive progress should, however, not mislead stakeholders into believing that microbiome analysis is now a 'done deal' and ready for routine clinical deployment. As we have outlined, there are a multitude of options, considerations and confounders for clinicians and scientists to consider before drawing any conclusions on classification, prediction and causality. If appropriately managed before, during and after studies, microbiome analysis will develop even more clinical utility and will become integrated into health-care management and clinical research, in turn leading to better diagnostics and therapeutics for the benefit of everyone.

1. Manichanh, C., Borruel, N., Casellas, F. & Guarner, F. The gut microbiota in IBD. *Nat. Rev. Gastroenterol. Hepatol.* **9**, 599–608 (2012).
2. Salonen, A., de Vos, W. M. & Palva, A. Gastrointestinal microbiota in irritable bowel syndrome: present state and perspectives. *Microbiology* **156**, 3205–3215 (2010).
3. Tang, W. H. *et al.* Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N. Engl. J. Med.* **368**, 1575–1584 (2013).
4. Pedersen, H. K. *et al.* Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* **535**, 376–381 (2016).
5. Sears, C. L. & Garrett, W. S. Microbes, microbiota, and colon cancer. *Cell Host Microbe* **15**, 317–328 (2014).
6. Marchesi, J. R. & Ravel, J. The vocabulary of microbiome research: a proposal. *Microbiome* **3**, 31 (2015).
7. Costello, E. K., Stagaman, K., Dethlefsen, L., Bohannan, B. J. & Relman, D. A. The application of ecological theory toward an understanding of the human microbiome. *Science* **336**, 1255–1262 (2012).
8. Cho, I. *et al.* Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature* **488**, 621–626 (2012).
9. Dethlefsen, L., Huse, S., Sogin, M. L. & Relman, D. A. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol.* **6**, e280 (2008).
10. Petrosino, J. F., Highlander, S., Luna, R. A., Gibbs, R. A. & Versalovic, J. Metagenomic pyrosequencing and microbial identification. *Clin. Chem.* **55**, 856–866 (2009).
11. Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V. & Polz, M. F. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J. Bacteriol.* **186**, 2629–2635 (2004).
12. Neefs, J. M., Van de Peer, Y., De Rijk, P., Chapelle, S. & De Wachter, R. Compilation of small ribosomal subunit RNA structures. *Nucleic Acids Res.* **21**, 3025–3049 (1993).
13. Claesson, M. J. *et al.* Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Res.* **38**, e200 (2010).
14. Clooney, A. G. *et al.* Comparing apples and oranges?: Next generation sequencing and its impact on microbiome analysis. *PLoS ONE* **11**, e0148028 (2016).
15. Findley, K. *et al.* Topographic diversity of fungal and bacterial communities in human skin. *Nature* **498**, 367–370 (2013).
16. Lavelle, A. *et al.* Spatial variation of the colonic microbiota in patients with ulcerative colitis and control volunteers. *Gut* **64**, 1553–1561 (2015).
17. Huse, S. M. *et al.* Comparison of brush and biopsy sampling methods of the ileal pouch for assessment of mucosa-associated microbiota of human subjects. *Microbiome* **2**, 5 (2014).
18. Chiodini, R. J. *et al.* Microbial population differentials between mucosal and submucosal intestinal tissues in advanced crohn's disease of the ileum. *PLoS ONE* **10**, e0134382 (2015).
19. Watt, E. *et al.* Extending colonic mucosal microbiome analysis-assessment of colonic lavage as a proxy for endoscopic colonic biopsies. *Microbiome* **4**, 61 (2016).
20. Budding, A. E. *et al.* Rectal swabs for analysis of the intestinal microbiota. *PLoS ONE* **9**, e101344 (2014).
21. Shobar, R. M. *et al.* The effects of bowel preparation on microbiota-related metrics differ in health and in inflammatory bowel disease and for the mucosal and luminal microbiota compartments. *Clin. Transl Gastroenterol.* **7**, e143 (2016).
22. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).
23. Flemer, B. *et al.* Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut* http://dx.doi.org/10.1136/gutjnl-2015-309595 (2016).
24. Gorzelak, M. A. *et al.* Methods for improving human gut microbiome data by reducing variability through sample processing and storage of stool. *PLoS ONE* **10**, e0134802 (2015).
25. Cardona, S. *et al.* Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiol.* **12**, 158 (2012).
26. Bahl, M. I., Bergstrom, A. & Licht, T. R. Freezing fecal samples prior to DNA extraction affects the Firmicutes to Bacteroidetes ratio determined by downstream quantitative PCR analysis. *FEMS Microbiol. Lett.* **329**, 193–197 (2012).
27. Shaw, A. G. *et al.* Latitude in sample handling and storage for infant faecal microbiota studies: the elephant in the room? *Microbiome* **4**, 40 http://dx.doi.org/10.1186/s40168-016-0186-x (2016).
28. Vogtmann, E. *et al.* Comparison of collection methods for fecal samples in microbiome studies. *Am. J. Epidemiol.* **185**, 115–123 (2017).
29. Hill, C. J. *et al.* Effect of room temperature transport vials on DNA quality and phylogenetic composition of faecal microbiota of elderly adults and infants. *Microbiome* **4**, 19 (2016).
30. Anderson, E. L. *et al.* A robust ambient temperature collection and stabilization strategy: enabling worldwide functional studies of the human microbiome. *Sci. Rep.* **6**, 31731 (2016).
31. Flores, R. *et al.* Collection media and delayed freezing effects on microbial composition of human stool. *Microbiome* **3**, 33 (2015).
32. Choo, J. M., Leong, L. E. & Rogers, G. B. Sample storage conditions significantly influence faecal microbiome profiles. *Scientif. Rep.* **5**, 16350 (2015).
33. Sherker, A. R., Cherepanov, V., Alvandi, Z., Ramos, R. & Feld, J. J. Optimal preservation of liver biopsy samples for downstream translational applications. *Hepatol. Int.* **7**, 758–766 (2013).
34. Persson, S., de Boer, R. F., Kooistra-Smid, A. M. & Olsen, K. E. Five commercial DNA extraction systems tested and compared on a stool sample collection. *Diagnost. Microbiol. Infecti. Dis.* **69**, 240–244 (2011).
35. Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z. & Forney, L. J. Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS ONE* **7**, e33865 (2012).
36. Li, F., Hullar, M. A. & Lampe, J. W. Optimization of terminal restriction fragment polymorphism (TRFLP) analysis of human gut microbiota. *J. Microbiol. Methods* **68**, 303–311 (2007).
37. Ariefdjohan, M. W., Savaiano, D. A. & Nakatsu, C. H. Comparison of DNA extraction kits for PCR-DGGE analysis of human intestinal microbial communities from fecal specimens. *Nutr. J.* **9**, 23 (2010).
38. Becker, L., Steglich, M., Fuchs, S., Werner, G. & Nubel, U. Comparison of six commercial kits to extract bacterial chromosome and plasmid DNA for MiSeq sequencing. *Scientif. Rep.* **6**, 28063 (2016).
39. Mirsepasi, H. *et al.* Microbial diversity in fecal samples depends on DNA extraction method: easyMag DNA extraction compared to QIAamp DNA stool mini kit extraction. *BMC Res. Notes* **7**, 50 (2014).
40. Wesolowska-Andersen, A. *et al.* Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* **2**, 19 (2014).
41. Gerasimidis, K. *et al.* The effect of DNA extraction methodology on gut microbiota research applications. *BMC Res. Notes* **9**, 365 (2016).
42. Hart, M. L., Meyer, A., Johnson, P. J. & Ericsson, A. C. Comparative evaluation of DNA extraction methods from feces of multiple host species for downstream next-generation sequencing. *PLoS ONE* **10**, e0143334 (2015).
43. Kennedy, N. A. *et al.* The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. *PLoS ONE* **9**, e88982 (2014).
44. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 http://dx.doi.org/10.1186/s12915-014-0087-z (2014).
45. Lauder, A. P. *et al.* Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome* **4**, 29 (2016).
46. Perez-Munoz, M. E., Arrieta, M. C., Ramer-Tait, A. E. & Walter, J. A critical assessment of the "sterile womb" and "*in utero* colonization" hypotheses: implications for research on the pioneer infant microbiome. *Microbiome* **5**, 48 (2017).
47. Edri, S. & Tuller, T. Quantifying the effect of ribosomal density on mRNA stability. *PLoS ONE* **9**, e102308 (2014).
48. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. Nature reviews. *Genetics* **17**, 333–351 (2016).
49. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
50. Treangen, T. J. *et al.* MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.* **14**, R2 (2013).
51. Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Res.* **40**, e155 (2012).
52. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* **1**, 18 (2012).
53. Compeau, P. E. C., Pevzner, P. A. & Tesla, G. How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* **29**, 987–991 (2011).
54. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
55. Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
56. Afiahayati, Sato, K. & Sakakibara, Y. MetaVelvet-SL: an extension of the Velvet assembler to a *de novo* metagenomic assembler utilizing supervised learning. *DNA Res.*: *Int. J. Rapid Publ. Rep. Genes Genomes* **22**, 69–77 (2015).
57. Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glockner, F. O. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformat.* **5**, 163 (2004).
58. Patil, K. R., Roune, L. & McHardy, A. C. The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. *PLoS ONE* **7**, e38581 (2012).
59. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
60. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* **12**, 902–903 (2015).
61. Brady, A. & Salzberg, S. L. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods* **6**, 673–676 (2009).
62. Wang, Y., Leung, H., Yiu, S. & Chin, F. MetaCluster-TA: taxonomic annotation for metagenomic data based on assembly-assisted binning. *BMC Genom.* **15** (Suppl. 1), S12 (2014).
63. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28**, 1033–1034 (2012).

64. Lin, H. H. & Liao, Y. C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Scientif. Rep.* **6**, 24175 (2016).
65. Peabody, M. A., Van Rossum, T., Lo, R. & Brinkman, F. S. Evaluation of shotgun metagenomics sequence classification methods using *in silico* and *in vitro* simulated communities. *BMC Bioinformat.* **16**, 363 (2015).
66. Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genom.* **16**, 236 (2015).
67. Lindgreen, S., Adair, K. L. & Gardner, P. P. An evaluation of the accuracy and speed of metagenome analysis tools. *Scientif. Rep.* **6**, 19233 (2016).
68. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38**, e132 (2010).
69. Rho, M., Tang, H. & Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* **38**, e191 (2010).
70. Hyatt, D., LoCascio, P. F., Hauser, L. J. & Uberbacher, E. C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**, 2223–2230 (2012).
71. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
72. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–D293 (2016).
73. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
74. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
75. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–230 (2014).
76. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).
77. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).
78. Meyer, F. *et al.* The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformat.* **9**, 386 (2008).
79. Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P. & Frazier, M. CAMERA: a community resource for metagenomics. *PLoS Biol.* **5**, e75 (2007).
80. Hunter, S. *et al.* EBI metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.* **42**, D600–606 (2014).
81. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336 (2010).
82. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).
83. Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* **10**, 996–998 (2013).
84. Plummer, E., Twin, J., Bulach, D. M., Garland, S. M. & Tabrizi, S. N. A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. *J. Proteomics Bioinform.* **8**, 283–291 (2015).
85. Westcott, S. L. & Schloss, P. D. *De novo* clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* **3**, e1487 (2015).
86. Jervis-Bardy, J. *et al.* Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. *Microbiome* **3**, 19 (2015).
87. Kopylova, E. *et al.* Open-source sequence clustering methods improve the state of the art. *mSystems* **1**, e00003–00015 (2016).
88. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200 (2011).
89. Haas, B. J. *et al.* Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* **21**, 494–504 (2011).
90. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).
91. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
92. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
93. Maidak, B. L. *et al.* The RDP-II (Ribosomal Database Project). *Nucleic Acids Res.* **29**, 173–174 (2001).
94. Koljalg, U. *et al.* UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytol.* **166**, 1063–1068 (2005).
95. Allard, G., Ryan, F. J., Jeffery, I. B. & Claesson, M. J. SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinformat.* **16**, 324 (2015).
96. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
97. Chen, W., Zhang, C. K., Cheng, Y., Zhang, S. & Zhao, H. A comparison of methods for clustering 16S rRNA sequences into OTUs. *PLoS ONE* **8**, e70837 (2013).
98. Callahan, B. J. *et al.* DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**, 581–583 (2016).
99. Huson, D. H. *et al.* MEGAN Community Edition — Interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Computat. Biol.* **12**, e1004957 (2016).
100. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
101. Langille, M. G. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnol.* **31**, 814–821 (2013).
102. Asshauer, K. P., Wemheuer, B., Daniel, R. & Meinicke, P. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* **31**, 2882–2884 (2015).
103. van Nood, E. *et al.* Duodenal infusion of donor feces for recurrent *Clostridium difficile*. *N. Engl. J. Med.* **368**, 407–415 (2013).
104. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
105. Claesson, M. J. *et al.* Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488**, 178–184 (2012).
106. Olle, B. Medicines from microbiota. *Nat. Biotechnol.* **31**, 309–315 (2013).
107. US Food and Drug Administration. *Early Clinical Trials With Live Biotherapeutic Products: Chemistry, Manufacturing, and Control Information; Guidance for Industry* (FDA, 2016).
108. Goldberg, B., Sichtig, H., Geyer, C., Ledeboer, N. & Weinstock, G. M. Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. *mBio* **6**, e01888–e01815 (2015).
109. Wilson, M. R. *et al.* Acute west nile virus meningoencephalitis diagnosed via metagenomic deep sequencing of cerebrospinal fluid in a renal transplant patient. *Am. J. Transplant*. http://dx.doi.org/10.1111/ajt.14058 (2016).
110. Zeevi, D. *et al.* Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094 (2015).
111. Bauer, E., Laczny, C. C., Magnusdottir, S., Wilmes, P. & Thiele, I. Phenotypic differentiation of gastrointestinal microbes is reflected in their encoded metabolic repertoires. *Microbiome* **3**, 55 (2015).
112. Heinken, A. & Thiele, I. Systems biology of host-microbe metabolomics. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **7**, 195–219 (2015).
113. [No authors listed.] *Babraham Bioinformatics* http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
114. [No authors listed.] *Hannonlab* http://hannonlab.cshl.edu/fastx_toolkit/index.html
115. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
116. Patel, R. K. & Jain, M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE* **7**, e30619 (2012).
117. Zhou, Q., Su, X., Jing, G. & Ning, K. Meta-QC-Chain: comprehensive and fast quality control method for metagenomic data. *Genom. Proteom. Bioinformat.* **12**, 52–56 (2014).
118. Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. Meta-IDBA: a *de novo* assembler for metagenomic data. *Bioinformatics* **27**, i94–i101 (2011).
119. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
120. Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F. & Corbeil, J. Ray Meta: scalable *de novo* metagenome assembly and profiling. *Genome Biol.* **13**, R122 (2012).
121. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
122. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
123. Haider, B. *et al.* Omega: an overlap-graph *de novo* assembler for metagenomics. *Bioinformatics* **30**, 2717–2722 (2014).
124. Reddy, R. M., Mohammed, M. H. & Mande, S. S. MetaCAA: a clustering-aided methodology for efficient assembly of metagenomic datasets. *Genomics* **103**, 161–168 (2014).
125. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. metaSPAdes: a new versatile de novo metagenomics assembler. *arXiv* 1604.03071 (2016).
126. Ye, Y. & Tang, H. An ORFome assembly approach to metagenomics sequences analysis. *J. Bioinformat. Computat. Biol.* **7**, 455–471 (2009).
127. Yu, F., Sun, Y., Liu, L. & Farmerie, W. GSTaxClassifier: a genomic signature based taxonomic classifier for metagenomic data analysis. *Bioinformation* **4**, 46–49 (2010).
128. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods* **10**, 1196–1199 (2013).
129. Ounit, R. & Lonardi, S. Higher classification sensitivity of short metagenomic reads with CLARK-S. *Bioinformatics* http://dx.doi.org/10.1093/bioinformatics/btw542 (2016).
130. Diaz, N. N., Krause, L., Goesmann, A., Niehaus, K. & Nattkemper, T. W. TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformat.* **10**, 56 (2009).
131. Rosen, G. L., Reichenberger, E. R. & Rosenfeld, A. M. NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics* **27**, 127–129 (2011).
132. Stark, M., Berger, S. A., Stamatakis, A. & von Mering, C. MLTreeMap—accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genom.* **11**, 461 (2010).
133. Freitas, T. A., Li, P. E., Scholz, M. B. & Chain, P. S. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Res.* **43**, e69 (2015).
134. Gerlach, W. & Stoye, J. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res.* **39**, e91 (2011).
135. Ames, S. K. *et al.* Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics* **29**, 2253–2260 (2013).
136. Droge, J., Gregor, I. & McHardy, A. C. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics* **31**, 817–824 (2015).
137. MacDonald, N. J., Parks, D. H. & Beiko, R. G. Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic Acids Res.* **40**, e111 (2012).
138. Monzoorul Haque, M., Ghosh, T. S., Komanduri, D. & Mande, S. S. SOrt-ITEMS: sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics* **25**, 1722–1730 (2009).
139. Mohammed, M. H., Ghosh, T. S., Singh, N. K. & Mande, S. S. SPHINX—an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics* **27**, 22–30 (2011).
140. Nalbantoglu, O. U., Way, S. F., Hinrichs, S. H. & Sayood, K. RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC Bioinformat.* **12**, 41 (2011).
141. Koslicki, D., Foucart, S. & Rosen, G. WGSQuikr: fast whole-genome shotgun metagenomic classification. *PLoS ONE* **9**, e91784 (2014).

142. Chan, C. K., Hsu, A. L., Halgamuge, S. K. & Tang, S. L. Binning sequences using very sparse labels within a metagenome. *BMC Bioinformat.* **9**, 215 (2008).
143. Schreiber, F., Gumrich, P., Daniel, R. & Meinicke, P. Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics* **26**, 960–961 (2010).
144. Weber, M. *et al.* Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics. *ISME J.* **5**, 918–928 (2011).
145. Pati, A., Heath, L. S., Kyrpides, N. C. & Ivanova, N. ClaMS: a classifier for metagenomic sequences. *Standards Genom. Sci.* **5**, 248–253 (2011).
146. Davenport, C. F. *et al.* Genometa—a fast and accurate classifier for short metagenomic shotgun reads. *PLoS ONE* **7**, e41224 (2012).
147. Sharma, A. K., Gupta, A., Kumar, S., Dhakan, D. B. & Sharma, V. K. Woods: a fast and accurate functional annotator and classifier of genomic and metagenomic sequences. *Genomics* **106**, 1–6 (2015).
148. Ghosh, T. S., Monzoorul Haque, M. & Mande, S. S. DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences. *BMC Bioinformat.* **11** (Suppl. 7), S14 (2010).
149. Liu, J. *et al.* Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. *Nucleic Acids Res.* **41**, e3 (2013).
150. Mohammed, M. H. *et al.* INDUS - a composition-based approach for rapid and accurate taxonomic classification of metagenomic sequences. *BMC Genom.* **12** (Suppl. 3), S4 (2011).
151. Sharma, V. K., Kumar, N., Prakash, T. & Taylor, T. D. Fast and accurate taxonomic assignments of metagenomic sequences using MetaBin. *PLoS ONE* **7**, e34030 (2012).
152. Liu, B., Gibbons, T., Ghodsi, M., Treangen, T. & Pop, M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genom.* **12** (Suppl. 2), S4 (2011).
153. Rasheed, Z. & Rangwala, H. Metagenomic taxonomic classification using extreme learning machines. *J. Bioinformat. Computat. Biol.* **10**, 1250015 (2012).
154. Ander, C., Schulz-Trieglaff, O. B., Stoye, J. & Cox, A. J. metaBEETL: high-throughput analysis of heterogeneous microbial populations from shotgun DNA sequences. *BMC Bioinformat.* **14** (Suppl. 5), S2 (2013).
155. Porter, M. S. & Beiko, R. G. SPANNER: taxonomic assignment of sequences using pyramid matching of similarity profiles. *Bioinformatics* **29**, 1858–1864 (2013).
156. Piro, V. C., Lindner, M. S. & Renard, B. Y. DUDes: a top-down taxonomic profiler for metagenomics. *Bioinformatics* **32**, 2272–2280 (2016).
157. Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Commun.* **7**, 11257 (2016).
158. Wu, Y. W., Tang, Y. H., Tringe, S. G., Simmons, B. A. & Singer, S. W. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2**, 26 (2014).
159. Petrenko, P., Lobb, B., Kurtz, D. A., Neufeld, J. D. & Doxey, A. C. MetAnnotate: function-specific taxonomic profiling and comparison of metagenomes. *BMC Biol.* **13**, 92 (2015).
160. Luo, C., Rodriguez, R. L. & Konstantinidis, K. T. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res.* **42**, e73 (2014).
161. Jiang, H., An, L., Lin, S. M., Feng, G. & Qiu, Y. A statistical framework for accurate taxonomic assignment of metagenomic sequencing reads. *PLoS ONE* **7**, e46450 (2012).
162. Klingenberg, H., Asshauer, K. P., Lingner, T. & Meinicke, P. Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. *Bioinformatics* **29**, 973–980 (2013).
163. Reddy, R. M., Mohammed, M. H. & Mande, S. S. TWARIT: an extremely rapid and efficient approach for phylogenetic classification of metagenomic sequences. *Gene* **505**, 259–265 (2012).
164. Hou, T. *et al.* Classification of metagenomics data at lower taxonomic level using a robust supervised classifier. *Evol. Bioinformat. Online* **11**, 3–10 S20523 (2015).
165. Kristiansson, E., Hugenholtz, P. & Dalevi, D. ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics* **25**, 2737–2738 (2009).
166. Li, W. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformat.* **10**, 359 (2009).
167. Kelley, D. R., Liu, B., Delcher, A. L., Pop, M. & Salzberg, S. L. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res.* **40**, e9 (2012).
168. Hoff, K. J., Lingner, T., Meinicke, P. & Tech, M. Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res.* **37**, W101–W105 (2009).
169. Liu, Y., Guo, J., Hu, G. & Zhu, H. Gene prediction in metagenomic fragments based on the SVM algorithm. *BMC Bioinformat.* **14** (Suppl. 5), S12 (2013).
170. van der Veen, B. E., Harris, H. M., O'Toole, P. W. & Claesson, M. J. Metaphor: finding bi-directional best hit homology relationships in (meta)genomic datasets. *Genomics* **104**, 459–463 (2014).
171. Liu, B. & Pop, M. MetaPath: identifying differentially abundant metabolic pathways in metagenomic datasets. *BMC Proc.* **5** (Suppl. 2), S9 (2011).
172. Overbeek, R. *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42**, D206–D214 (2014).
173. Powell, S. *et al.* eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.* **42**, D231–D239 (2014).
174. Abubucker, S. *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Computat. Biol.* **8**, e1002358 (2012).
175. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).
176. Markowitz, V. M. *et al.* IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res.* **42**, D568–573 (2014).
177. Wu, S., Zhu, Z., Fu, L., Niu, B. & Li, W. WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genom.* **12**, 444 (2011).
178. Goll, J. *et al.* METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics. *Bioinformatics* **26**, 2631–2632 (2010).
179. Su, X., Pan, W., Song, B., Xu, J. & Ning, K. Parallel-META 2.0: enhanced metagenomic data analysis with functional annotation, high performance computing and advanced visualization. PloS one **9**, e89323 (2014).
180. Kultima, J. R. *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS ONE* **7**, e47656 (2012).
181. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44**, W3–W10 (2016).
182. Fosso, B. *et al.* BioMaS: a modular pipeline for Bioinformatic analysis of Metagenomic AmpliconS. *BMC Bioinformat.* **16**, 203 (2015).
183. Angly, F. *et al.* PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformat.* **6**, 41 (2005).
184. Arumugam, M., Harrington, E. D., Foerstner, K. U., Raes, J. & Bork, P. SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics* **26**, 2977–2978 (2010).
185. Schloss, P. D. & Handelsman, J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* **71**, 1501–1506 (2005).
186. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
187. Hao, X., Jiang, R. & Chen, T. Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* **27**, 611–618 (2011).
188. Cai, Y. & Sun, Y. ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Res.* **39**, e95 (2011).
189. Ghodsi, M., Liu, B. & Pop, M. DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformat.* **12**, 271 (2011).
190. Russell, D. J., Way, S. F., Benson, A. K. & Sayood, K. A grammar-based distance metric enables fast and accurate clustering of large sets of 16S sequences. *BMC Bioinformat.* **11**, 601 (2010).
191. Wang, X., Yao, J., Sun, Y. & Mai, V. M-Pick, a modularity-based method for OTU picking of 16S rRNA sequences. *BMC Bioinformat.* **14**, 43 (2013).
192. Mahe, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* **3**, e1420 (2015).
193. Franzen, O. *et al.* Improved OTU-picking using long-read 16S rRNA gene amplicon sequencing and generic hierarchical clustering. *Microbiome* **3**, 43 (2015).
194. Wei, Z. G. & Zhang, S. W. MtHc: a motif-based hierarchical method for clustering massive 16S rRNA sequences into OTUs. *Mol. bioSystems* **11**, 1907–1913 (2015).
195. Mysara, M., Saeys, Y., Leys, N., Raes, J. & Monsieurs, P. CATCh, an ensemble classifier for chimera detection in 16S rRNA sequencing studies. *Appl. Environ. Microbiol.* **81**, 1573–1584 (2015).
196. Soergel, D. A., Dey, N., Knight, R. & Brenner, S. E. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME J.* **6**, 1440–1444 (2012).
197. Chaudhary, N., Sharma, A. K., Agarwal, P., Gupta, A. & Sharma, V. K. 16S classifier: a tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets. *PLoS ONE* **10**, e0116106 (2015).
198. Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. & Schmidt, T. M. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Res.* **43**, D593–D598 (2015).
199. Jaziri, F. *et al.* PhylOPDb: a 16S rRNA oligonucleotide probe database for prokaryotic identification. *Database (Oxford)* http://dx.doi.org/10.1093/database/bau036 (2014).
200. Ritari, J., Salojarvi, J., Lahti, L. & de Vos, W. M. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genom.* **16**, 1056 (2015).
201. Lozupone, C., Hamady, M. & Knight, R. UniFrac—an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformat.* **7**, 371 (2006).
202. Gilmore, R. D., Cieplak, W., Policastro, P. F. & Hackstadt, T. The 120 kilodalton outer membrane (rOmpB) of *Rickettsia rickettsii* is encoded by an unusually long open reading frame: evidence for protein processing from a large precursor. *Mol. Microbiol.* **5**, 2361–2370 (1991).
203. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods* **10**, 1200–1202 (2013).
204. Angly, F. E. *et al.* CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome* **2**, 11 (2014).
205. Beck, D., Settles, M. & Foster, J. A. OTUbase: an R infrastructure package for operational taxonomic unit data. *Bioinformatics* **27**, 1700–1701 (2011).
206. Seguritan, V. & Rohwer, F. FastGroup: a program to dereplicate libraries of 16S rDNA sequences. *BMC Bioinformat.* **2**, 9 (2001).
207. Giongo, A. *et al.* PANGEA: pipeline for analysis of next generation amplicons. *ISME J.* **4**, 852–861 (2010).
208. Kumar, S. *et al.* CLOTU: an online pipeline for processing and clustering of 454 amplicon reads into OTUs followed by taxonomic annotation. *BMC Bioinformat.* **12**, 182 (2011).
209. Nebel, M. E. *et al.* JAGUC—a software package for environmental diversity analyses. *J. Bioinformat. Computat. Biol.* **9**, 749–773 (2011).
210. Albanese, D., Fontana, P., De Filippo, C., Cavalieri, D. & Donati, C. MICCA: a complete and accurate software for taxonomic profiling of metagenomic data. *Scientif. Rep.* **5**, 9743 (2015).
211. Weisman, D., Yasuda, M. & Bowen, J. L. FunFrame: functional gene ecological analysis pipeline. *Bioinformatics* **29**, 1212–1214 (2013).

Acknowledgements
This work was supported by Science Foundation Ireland through a Centre Award to the APC Microbiome Institute (SFI/12/RC/2273).

Author contributions
All authors contributed equally to this work.

Competing interests statement
The authors declare no competing interests.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

FURTHER INFORMATION
International Human Microbiome Standards:
http://www.microbiome-standards.org/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

ERRATUM

A clinician's guide to microbiome analysis

Marcus J. Claesson, Adam G. Clooney & Paul W. O'Toole

Nature Reviews Gastroenterology & Hepatology http://dx.doi.org/10.1038/nrgastro.2017.97 (2017)

In the version of this Review initially published online, the article should have indicated that Marcus J. Claesson and Adam G. Clooney contributed equally to this work. The error has been corrected for the HTML, PDF and print versions of the article.