Methods for Studying Gut Microbiota: A Primer for Physicians

Aditya N. Sarangi [*,†], Amit Goel [†], Rakesh Aggarwal [*,†]

[*]Biomedical Informatics Center, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow 226014, India and [†]Department of Gastroenterology, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow 226014, India

Human gastrointestinal tract contains a large variety of microbes, in particular bacteria. Studies in recent years have strongly suggested a role for these microbes, collectively referred to as gut microbiota, in the maintenance of homeostasis during health. In addition, alterations in gut microbiota have been reported in several diseases, including those related to the gastrointestinal tract and several systemic conditions, and are believed to play a pathogenetic role in at least some of these. Given the close association between the human gut and liver, the association with gut microbiota appears to be particularly strong for a wide variety of liver diseases. This piece, aimed primarily at physicians, reviews in brief the methods used to study gut microbiota, with particular emphasis on those that use sequences of bacterial 16S rRNA gene or its components. (J Clin Exp Hepatol 2019;9:62–73)

Gut Microbiota

The term 'human microbiota' refers to the complete set of microbes that live in and on the human body.[1] It appears to play a major role in health and disease, either directly through the expression of microbial genes that provide the human host some metabolic capabilities which its own genome lacks, or indirectly through interaction with human physiology, particularly with the immune system. The main locations on the human body where the microbiota exists are the gastrointestinal tract, female genital tract, oral cavity and the respiratory tract. Of these, the gastrointestinal tract is the site that is the richest in microbial organisms.

Several methods have been used to study gut microbiota, and these have undergone a major change over time (Table 1). This article describes the various methods used to study microbiota, and the advantages and limitations of each. The gut microbiota include several different groups of organisms, including bacteria, viruses, fungi, archaea, etc. However, of these, bacteria have been the most extensively studied, and much less is known about the viruses (virome), fungi (fungome) and other prokaryotes (e.g. archea) present in the gastrointestinal tract. This article therefore focuses on the study of gut bacteria, and henceforth the term gut microbiota has been used interchangeably with the set of bacteria present in a person's gut. Also, before one proceeds further, it may be useful to understand a few terms that are used in relation to the study of microbiota (e.g. microbiome, metagenome, etc.), which represent concepts quite similar to, but not identical with, the term 'microbiota' (see Box).

CULTURE-BASED METHODS

The initial studies used traditional bacterial culture techniques, followed by phenotyping of the cultured bacteria using morphological and biochemical characteristics. However, a large proportion of bacteria in the gut are obligate anaerobes, which often do not survive the procedures used for obtaining specimens from the gastrointestinal tract, or for transport to the laboratory and storage. Furthermore, various organisms present in the human gut differ in their propensity to grow in culture. Thus, the results of relative abundance of various bacteria in the gut lumen deduced using culture-based techniques are heavily biased in favour of aerobic organisms that grow easily in in vitro culture, while missing the anaerobic bacteria. Also, these techniques markedly underestimate the diversity of bacteria in the intestinal luminal contents, and hence their usefulness in studying changes in the profile of gut microbiota is limited. Hence, these techniques never gained sufficient traction for the study of profile of gut microbiota, and their use was limited to the study of individual culturable bacterial groups (e.g. a particular genus) in particular clinical situations.

To overcome these limitations of culture techniques, and with the development in the late 20th century of techniques for the study of bacterial genomic material, several molecular approaches were developed in which different bacterial species were identified based on the sequences of their 16S ribosomal RNA (16S rRNA) genes.

Keywords: gut microbiota, metagenome, 16S rRNA, next generation sequencing, 16S rRNA data analysis, microbial diversity
Received: 25 April 2018; *Accepted:* 27 April 2018; *Available online:* 4 May 2018
Address for correspondence: Rakesh Aggarwal, Department of Gastroenterology, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow 226014, India.
E-mail: aggarwal.ra@gmail.com
Abbreviations: BDI: Beta Diversity Index; OTUs: Operational Taxonomic Units; PCR: Polymerase Chain Reaction
https://doi.org/10.1016/j.jceh.2018.04.016

Table 1 Techniques Used for Study of Microbiota in the Gut as well as Other Body Sites.

A. Culture-based methods
B. Molecular-based (nucleic-acid based) methods
a. Non-sequencing methods
i. Fluorescence in situ hybridization flow cytometry
ii. Pulsed field gel electrophoresis
iii. Denaturing gradient gel electrophoresis
iv. Temperature gradient gel electrophoresis
v. Single-strand conformation polymorphism
b. Sequence-based methods
i. Sequencing of 16S rRNA genes or their hypervariable regions (targeted gene sequencing)
ii. Whole bacterial genome DNA (metagenome) sequencing
iii. Whole bacterial mRNA (meta-transcriptome) sequencing
C. Methods based on detection and quantification of small metabolites
i. Gas chromatography mass spectrometry
ii. Capillary electrophoresis coupled to mass spectrometry
iii. Fourier-transform infrared spectroscopy
iv. Nuclear and proton magnetic resonance spectroscopy

Box 1 Definitions

Microbiota
The assemblage of microorganisms present in a defined environment.
Metagenome
The collection of genomes and genes from the members of a microbiota. This collection is obtained through shotgun sequencing of nucleic acid extracted from a specimen (metagenomics) followed by assembly or mapping to a reference database and annotation.
Microbiome
This term refers to the entire habitat, including the microorganisms (bacteria, archaea, lower and higher eurkaryotes, and viruses), their genomes (i.e., genes), and the surrounding environmental conditions. However, the term is often also used for what is described as 'metagenome'.
Metatranscriptomics
Analysis of the suite of expressed RNAs (meta-RNAs) by high-throughput sequencing of the corresponding meta-cDNAs. This approach provides information on the regulation and expression profiles of complex microbiomes. The resulting census of all expressed RNAs present in a specimen is called 'metatranscriptome'.
Metaproteomics
Characterization of the entire protein complement of environmental or clinical samples at a given point in time. The resulting census of all proteins present in any given specimen or tissue is called 'proteome'.
Metabolomics
Determination of metabolite profile(s) in any given specimen or tissue. The resulting census of all metabolites present in any given specimen or tissue is called 'metabolome'.
These definitions are adapted from Marchesi et al.[1]

BACTERIAL 16S RIBOSOMAL RNA

Each living cell contains ribosomes, which are composed of two subunits, one large and one small. The small ribosomal subunit contains an RNA molecule which is 16S in size in case of prokaryotic cells (including bacteria) and 18S in case of eukaryotic cells. These small RNA molecules are encoded by the bacterial genome.

The bacterial 16S rRNA is around 1500 nucleotide long,[2] with some variation across species. Several stretches of this gene are highly conserved across all bacterial groups. These conserved or constant sequences are interspersed with regions that show marked variation (the 'hypervariable regions'); nine such regions have been recognized and are referred to as V1 to V9 (Figure 1). The variations in nucleotide sequences in these hypervariable regions reflect evolutionary divergence of bacteria, and hence, these sequences provide a reliable method for identification and phylogenetic classification of bacterial species. Methods for bacterial identification based on nucleotide sequences in these regions have the advantage that these do not need prior bacterial culture, and hence can detect bacteria that are culturable as well as those that do not grow well. Further, when these methods are applied to bacterial mixtures, their results provide a relatively unbiased assessment of the relative abundance of various bacterial groups, irrespective of their capability to grow and growth rate in culture.

The molecular techniques that were initially developed could exploit only differences in the length (e.g. those identified by gel electrophoresis) and major variations in the nucleotide sequences (e.g. using restriction fragment length polymorphism) of these hypervariable regions across various bacterial species. However, in the last 10–15 years, rapid development in nucleic acid sequencing technology has led to high-throughput multi-parallel sequencing becoming widely available and at a reasonable price; this has made these techniques virtually the current gold standard for the study of gut microbiota.

NON-SEQUENCING BASED MOLECULAR METHODS FOR STUDY OF MICROBIOTA

In these techniques, bacterial nucleic acid is extracted from the specimen to be analyzed, followed by amplification either of the entire length of the 16S rRNA gene or a segment of this gene that includes one or more selected hypervariable regions. This can be done using Polymerase Chain Reaction (PCR) with universal primers corresponding to conserved regions in the bacterial genome flanking the entire 16S rRNA gene or its selected hypervariable

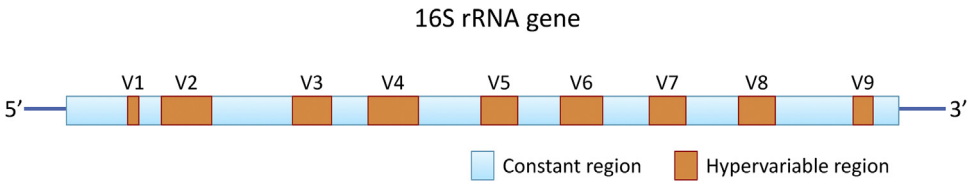


Figure 1 A representation of 16S ribosomal RNA gene showing the nine hypervariable regions (in brown colour) and constant regions (in blue).

region(s). The resultant amplified mixture of 16S rRNA genes or of its hypervariable fragments from all the bacteria contained in the specimen can then be resolved using one of several techniques. These techniques have included electrophoresis-based separation based on fragment length [e.g. denaturing gradient gel electrophoresis or on a temperature-gradient gel electrophoresis], or those based on the presence of specific nucleotide sequences [e.g. Fluorescence In Situ Hybridization flow cytometry (FISH-flow)[3] and bacterial DNA microarrays].

The main drawback of these methods is a limited resolution of bacterial groups. This results from the fact that differences in length as well as sequences of 16S rRNA gene from closely-related bacterial groups (e.g. species, genera, and often larger phylogroups, such as families and orders) are relatively small, precluding their separation. Further, the bacterial groups present in low abundance are missed. Hence, these methods have over time been replaced by newer-generation sequencing techniques.

16S RRNA GENE SEQUENCING METHODS

The traditional Sanger technique for nucleic acid sequencing needs relatively pure DNA as the starting material, and provides only one sequence per experiment. Thus, it was not possible to sequence a specimen containing a mixture of related nucleic acids using this technique, except by cloning each of these nucleic acid molecules into separate vectors and sequencing each clone, a very tedious and costly undertaking. Since microbiota contains a mixture of bacteria with somewhat diverse genomic material, these could not be sequenced using this technique.

Several newer sequencing technologies, developed over the last 15 years, have permitted massively-parallel sequencing, i.e. simultaneous sequencing of each molecule contained in a DNA mixture, such as that isolated from a microbiota specimen. These techniques however pose two major challenges. First, these generate a large amount of data, with the number of sequences from each specimen often reaching several million, posing a nightmare for analysis. Second, these technologies generally provide much shorter read-lengths than were possible from Sanger sequencing. Several computational software tools and a high computational power that have since become available allow matching of a large number of nucleotide sequences to a large database, as also identification and merger of various overlapping and contiguous short sequence reads into longer reads (the so called *contigs*). These tools have however helped overcome these limitations, and permitted the widespread use of such sequencing.

Several different technologies were developed and commercialized for multi-parallel sequencing. However, most of these have fallen by the wayside, and most of the current studies on microbiota use one of the two equipments from one manufacturer (Illumina), namely: MiSeq (250 or 300-base length reads, lower output) and Illumina HiSeq (150-base length, higher output). In view of their limited read-lengths, these techniques allow sequencing of only one or two adjacent hypervariable regions of the 16S rRNA gene. This information permits one to determine the types of bacteria present as also their relative frequencies (abundance) in a mixed specimen. This sequence length, though practically reasonable for most work, may not effectively classify all bacterial species. A newer alternative, which allows for sequencing of the full-length bacterial 16S rRNA gene, is offered by the more-recently developed Single Molecule, Real-Time (SMRT) circular consensus sequencing equipment from Pacific Biosciences.[4] However, given the high cost of this technology, it has not yet become popular for the study of microbiota.

The study of gut microbiota using the newer-generation multi-parallel sequencing techniques involves several sequential steps, which are described in brief below.

Specimen Collection, Preparation and Sequencing

Choice and Collection of Specimens

The accuracy of gut microbiota analysis depends on appropriate selection, collection and pre-processing of specimens. Specimens used for analysis of human gut microbiota have included stool, intestinal tissue biopsy and intestinal mucosal lavage material – the latter two being collected during endoscopic examination.[5] Each of these specimens has certain advantages and disadvantages.

If the aim is to assess the interaction of a certain segment of the host gut with microbiota, tissue biopsy may be the most preferable, permitting assessment of both the host tissue characteristics and the microbiota. However, several parts of the gastrointestinal tract are not

easily amenable to tissue biopsy (e.g. small intestine). Biopsies from other parts of the gut may need specific preparation (e.g. lavage for colonic biopsies), which may itself alter the microbiota. Lavage material too suffers the same limitation.

By contrast, fecal specimens draw from several segments along the length of the gastrointestinal tract, though primarily the distal gut. Thus, these provide a good surrogate for bacteria in the colon, the site where gastrointestinal bacteria are anyway the most numerous in density.[5]

Irrespective of the choice of specimen type, all the specimens used in a particular study (whether from one group of subjects at one or multiple time points, or multiple groups that are to be compared with each other, e.g. patients and controls) should be collected, stored and processed in an identical manner. Ideally, all specimens from one study should also be processed simultaneously, and in the same laboratory by the same personnel, to minimize any batch effect.[6]

DNA Extraction

In the next step, the specimen is subjected to DNA extraction. Several different protocols have been developed for this step. These methods vary by the type of specimen used for analysis.[7–9] Also, the results obtained may vary with the method used. Hence, International Human Microbiome Standards (IHMS) Consortium has provided standard operating procedures to standardize specimen collection and DNA extraction methods for such studies (http://www.microbiome-standards.org), so that data obtained can be compared across studies.

Selection of HVR, Amplification of DNA and Generation of DNA Libraries

Of the nine hypervariable regions in 16S rRNA, V3, V4 and V6, or pairs of adjacent HVRs (e.g. V3–V4 or V4–V5) have been the most widely used. Of these, the V4–V5 region is particularly suited for the study of microbiota, since it provides the most comparable results across platforms[10] and provides a high taxonomic resolution.[11,12] However, the sequencing of these hypervariable regions requires a technique with a longer read length than the methods that use only one hypervariable region.

The choice of region of 16S rRNA gene to be amplified and sequenced is based on its ability to accurately classify as many genera or species as possible (this needs inputs from previous studies in the literature), level of conservation of the flanking region across microbial species (the higher the better) and its length (whether the sequencing platform chosen can sequence this length in a cost-effective manner).

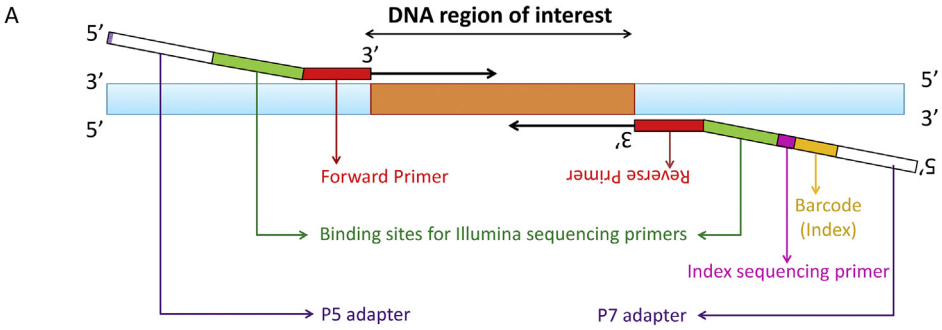
Once the choice of hypervariable region(s) of the 16S rRNA gene to be studied (the region of interest) is made, custom-designed primers which include the priming sequences flanking it as also sequences complementary to Illumina forward and reverse sequencing primers (located on Illumina sequencing flow cell – see below) are used to amplify the region of interest using polymerase chain reaction (Figure 2A).

The existing sequencing methods can generate enormous amount of data in one run (i.e. one experiment). This amount is much larger than the number of sequences (depth of sequencing) that one needs for adequate study of one specimen. Hence, it makes perfect sense to somehow combine multiple specimens in one run, to reduce costs. This is easily done by using slightly different reverse primers, each containing in its sequence a unique six-nucleotide 'index' sequence (Figure 2B). Thus, the amplification products for each specimen will contain different sequences for this 'index'. These products carrying distinct 'index' markers can then be pooled (in roughly equimolar quantities) and run in the same sequencing experiment. This process of pooling of different specimens is referred to as 'multiplexing'. Once the sequence data are obtained, these are computationally segregated (demultiplexed) by reading the 'index' region of each sequence to identify its origin.

Sequencing

The DNA library (or a mixture of libraries – if multiplexing is done) is loaded on to a flow cell, which resembles a glass slide to which several oligonucleotide molecules of two different types (the sequencing primers) are attached. The sequencing primers have sequences that are complementary to those of the adapters included at the ends of the two amplification primers used to generate the DNA library. Thus, each DNA molecule in the DNA library attaches to the flow cell via one of the adapters, and carries the other adapter at its free end. Since the number of attachment sites on the flow cell is much larger than the number of DNA molecules added to it, these molecules are widely separated from each other. Through several steps, as described in detail elsewhere,[13] each DNA molecule is then used to generate a local cluster consisting of its several identical copies around it. This results in formation of several million distinct clusters, each derived from a separate molecule in the DNA library, on the flow cell. In the next steps, DNA in each of these clusters is sequenced first in one (forward) direction and then in the opposite (reverse) direction. This generates several million pairs of data, with one pair representing data for each cluster, and hence for each individual DNA molecule in the original library. If the DNA fragments in the library are short, then the 3′-ends of the two reads (one forward and one reverse) in each pair can be made to overlap and their data fused with each other during analysis (Figure 3).

The two currently-available machines that use the Illumina platform provide reads of 150 (HiSeq) and



B

Primer name	Primer nucleotide sequence
V3F	5′-aatgatacggcgaccaccgagatct acactctttccctacacgacgctcttccgatct NNNNCCTACGGGAGGCAGCAG-3′
V3R	5′-caagcagaagacggcatacgagat XXXXX gtgactggagttcagacgtgtgctcttccgatct ATTACCGCGGCTGCTGG-3′

Figure 2 A schematic representation of various functional elements included in primers used to amplify the region of interest in the 16S rRNA gene for generation of Illumina DNA sequencing libraries (A), and corresponding example of actual primer sequences used to amplify the V3 hypervariable region in a particular study (B). (A) The template DNA (16S rRNA gene) is shown as thick horizontal line, with hypervariable region to be sequenced in brown and the flanking constant regions in blue. Each primer includes (beginning from the 5′-end) three main regions: (i) a P5 or P7 adapter sequence (purple) for binding to the Illumina sequencing flow-cell, (ii) a binding site for forward or reverse Illumina sequencing primer (green), and (iii) an annealing sequence that actually help bind the primer to the 16S rRNA gene (red), in the latter's constant region. In addition, one of the primers contains a short (usually 6 nucleotides in length) 'index sequence' (or 'barcode'; shown in yellow) needed for multiplexing (for running several specimens in one flow cell) and another short region, known as index sequencing primers (pink), that helps sequence the index/barcode. (B) In the primer sequences (forward primer: 341F; reverse primer: 518R), the colours of letters correspond to those of various segments of the primers in 'A'. Lower case purple letters (purple) at 5′ end represent adapter sequences necessary for binding of the library to the Illumina flow-cell, the underlined lowercase letters (green) represent binding site for Illumina sequencing primers, and the uppercase letters represent the actual annealing sequencing (red) for binding to the constant regions flanking the V3 region. The letters XXXXXX (yellow) represents the 6-nucleotide index region for 'multiplexing' (see main text). NNNN represent a few degenerate bases (i.e. these locations can carry any of the four possible nucleotide bases); these are added to help to provide sequence diversity, which is necessary for proper cluster identification by the sequencer.

300 (MiSeq) bases in either direction. Merger of read pairs can thus generate sequences of up to ~250 and ~550 nucleotides, respectively, while providing for an overlap of ~50 bases in the opposing reads. Individual HVRs from V2 to V7 have average lengths of 86 to 207 nucleotides. Hence either of these platforms can be used to sequence one of these HVR regions. In contrast, the average length of V8 HVR is 322 nucleotides.[14] Hence, to sequence this HVR region, or two adjacent HVRs, paired-end sequencing using MiSeq platform is advisable.

Processing and Analysis of 16S rRNA Sequence Data

The raw sequence data obtained contain sequences corresponding to sequencing adaptors and primers used for amplification; as the first step, these latter segments are trimmed away.

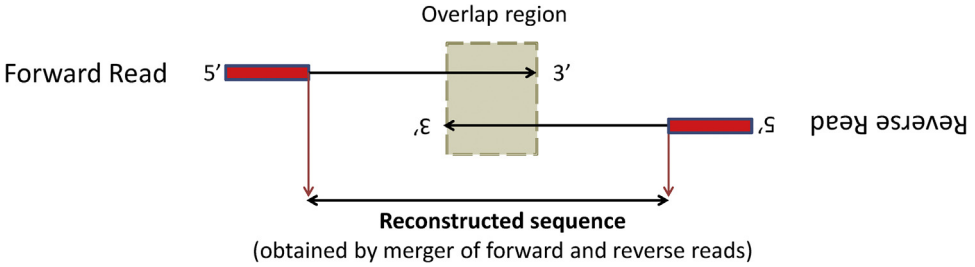


Figure 3 Schematic representation of merger of paired-end reads. The red boxes represent the primer regions which are trimmed out. Depending on the length of the region of interest and lengths of the forward and reverse reads, the two reads may overlap to a variable extent (shown as a grey box). In such cases, during analysis, the overlapping 3′-ends of the forward and reverse reads are compared to verify each other's accuracy. After this verification, sequences in the two directions can be merged to yield a 'reconstructed' or 'merged' sequence which is longer than the either starting sequence.

If the paired-end sequencing technique has been used, in which each DNA molecule is sequenced in both directions and the reads in the two directions partially overlap, then the next step is merge the paired forward and reverse reads into one read. This has two main advantages. First, since reads in the two directions overlap only partially, the merger provides a longer read than is possible by reading in only one direction. Second, the merger helps in excluding any low-quality reads. The quality of raw NGS reads declines as sequences proceeds towards the 3′-end. Thus, when bidirectional sequencing is done, the non-overlapping portions (5′ ends) of the forward and reverse reads represent the best quality data, and the overlapping portions (3′ ends) have the relatively poor-quality data. The merger process verifies that the overlapping data in the two directions are identical, serving to ensure that no errors have crept in, thus helping ensure that the overall data quality is good.

The sequencing equipment also provides an estimate of the data quality (higher quality = less risk of reading error) for each nucleotide that is read. There is always a possibility that certain bases in some sequences are of low-quality and hence more likely to represent sequencing errors. Quality-control filters are used to identify such poor-quality reads and purge these from the data. Generally, only reads with average quality score of 30 or above (which represents an expected error rate of fewer than one base for every 1000 bases) are selected for further analysis.

Widely used open-source tools for primer and adapter trimming, paired-end read merging and quality control analysis are listed in Table 2. Details on the usage, selectable features, strengths and limitations of these tools are usually available on the servers where these are hosted.

Assignment of Reads to Operational Taxonomic Units

The next step is clustering or binning the pre-processed high-quality sequences into operational taxonomic units (OTUs). Each OTU represents a cluster of nucleotide sequences that are highly similar and are likely to represent one (or a few closely-related) organisms.[15] This presumes that sequences with a high degree of nucleotide identity (usually >97%) belong to the same bacterial species. This assumption not only accounts for intra-species sequence variations, but also helps overcome the problem of occasional errors introduced during DNA sequencing; for instance, if two sequences differ by only 1–2 nucleotides, this difference may not be real and be due to sequencing errors, and hence, it makes sense to treat these as one. A lower clustering threshold of 95% is used for genus-level analysis.[16] The clustering also reduces the large data set of several sequences (usually in hundreds of thousands) to representative consensus sequences for a few clusters or OTUs and the count of number of sequences in each cluster – this helps reduce the run time of subsequent steps in data analysis.

Taxonomy Assignment

A representative sequence from each OTU is then mapped to a reference 16S-rRNA sequence database. The OTU is then assigned the taxonomy of the closest match found in the database on such mapping. By doing this for all the OTUs, one can obtain information on the various types of bacteria present and relative abundance of each, in a particular specimen.

Table 2 Popular Bioinformatics Tools Used for 16S rRNA Metagenome Analysis.

Purpose	Tools	URL
Trimming of primers and adapters	Cutadapt	https://github.com/marcelm/cutadapt
	Sickle	https://github.com/najoshi/sickle
	cutPrimers	https://github.com/aakechin/cutPrimers
	AdaperRemoval	https://github.com/MikkelSchubert/adapterremoval
Quality control	NGS-QC ToolKit	http://www.nipgr.res.in/ngsqctoolkit.html
	Trimmomatic	http://www.usadellab.org/cms/?page=trimmomatic
	clinQC	https://sourceforge.net/projects/clinqc/
	AfterQC	https://github.com/OpenGene/AfterQC
Merger of paired-end reads	Pandaseq	https://github.com/neufeld/pandaseq
	PEAR	https://sco.h-its.org/exelixis/web/software/pear/
	FLASH	https://ccb.jhu.edu/software/FLASH/
	MeFiT	https://github.com/nisheth/MeFiT
16S-rRNA metagenome analysis pipelines	QIIME	http://qiime.org/
	MOTHUR	https://www.mothur.org/
	MG-RAST	http://metagenomics.anl.gov/
	MICCA	http://micca.org/

These reference databases contain several thousand 16S-rRNA gene sequences, with information on the bacterium (name and phylogeny) from which each is derived. Several such databases are currently available, namely SILVA,[17] Ribosomal Database Project (RDP),[18] Greengenes[19] and EzTaxon-e.[20] Another alternative is the use of RNACentral,[21] an aggregated RNA resource, which allows the use of 16S rRNA sequences from several or all the above-mentioned databases. Alternatively, a specialized database of 16S rRNA sequences of human intestinal organisms (HITdb)[22] can be used when one is working with data from microbiota of intestinal origin.

Admittedly, no good match can be found for representative sequences from some OTUs, and these OTUs thus remain unclassified. Also, some OTUs can be assigned to a higher-level taxon (e.g. a particular family or order) but not to a specific lower-level taxon (e.g. genus or species).

USING THE MICROBIOTA COMPOSITION DATA FOR DECISION MAKING

Data on microbiota in one specimen are of little use. Almost always, one looks at data from several specimens. Once data on the type and abundance of various bacteria present in each of several specimens have been obtained, further analyses of such data can be broadly categorized into three types:

i) Estimation of diversity within a specimen and between groups of specimens;
ii) Identification of specific taxa that differ significantly between study groups; and
iii) Functional profiling to predict the genes and metabolic pathways.

Various steps involved in each of the above analyses are described below.

As a first step, data noise is reduced by purging data for OTUs (e.g. species) that are observed in only a few specimens (e.g. fewer than 10% specimens – i.e. in only 2 or fewer specimens, if the study has a total of 20 specimens) or account for very few reads (e.g. <0.005% of reads in all the specimens taken together). These bacterial groups that are present in only a few subjects with a particular disease or in a very small concentration are unlikely to be important for disease pathogenesis, and hence can be safely ignored. The specimen-wise observation count of each OTU remaining in the dataset is then tabulated as an OTU table, with each specimen represented as a column and each taxon as a row, and the cells contain information on abundance of a particular taxon in a particular specimen. This table describes the bacterial composition of each specimen, and forms the basis of analyses that follow.

Estimation of Diversity of Microbiota

Estimation of microbial diversity has clinical importance, as alterations in gut flora composition (also sometimes referred to as 'dysbiosis') are often associated with reduced microbiome diversity.[23] Diversity is assessed using two separate types of measures – namely alpha diversity and beta diversity – which represent entirely different constructs.

Alpha Diversity

Alpha-diversity is defined as the number of unique taxa (richness) and their distribution (evenness) in a particular specimen. Thus, a specimen which contains several different types of bacteria is considered to have a greater diversity than another specimen with fewer types of bacteria. Another factor that affects the assessment of diversity is the distribution of various bacterial types. For instance, let us think of two specimens (A and B) having four types of bacteria (a, b, c and d) each. Further, let us assume that specimen A contains the four types of bacteria in equal numbers, accounting for 25% of the bacterial cells each; by contrast, in specimen B, bacterium 'a' accounts for 97% of cells, and 'b', 'c' and 'd' for 1% each. In this case, the former specimen is more diverse and is considered to have a greater alpha diversity than the latter.

Several indices are available for estimation of alpha-diversity. These include Chao1 and Abundance-based Coverage Estimator (ACE), which primarily measure the species richness. By contrast, other commonly-used indices of alpha diversity, namely the Shannon Index and the Simpson index measure both the richness and the evenness of distribution of taxa.[24] These indices act as summary statistics of alpha diversity of individual specimens. A comparison of alpha diversity indices in two groups of specimens (e.g. from patients with a particular disease and healthy controls; using a parametric [unpaired *t* test] or non-parametric [Mann–Whitney *U* or Wilcoxon's rank sum test] statistical test) can inform us whether the disease is associated with a change in the diversity of gut microbiota.

Beta Diversity

Beta diversity provides a measure to assess the difference in species composition of two groups of specimens, e.g. those from patients with a particular disease and healthy subjects (control group). Looked at in another way, this measure calculates the number of species that are different between the two groups.

Let us look at an example of two studies, A and B with a few specimens each. In study A (Figure 4A), let us assume that the total number of species in the specimens included in two groups (let us call them *K* and *L*) is K_{tot} and L_{tot}, respectively. Further, let us assume that M_{tot} is the number of species common to both groups (assumed here as 3 species). In this situation, beta diversity is calculated as: Beta Diversity Index BDI(*K*,*L*) = $1 - [(M_{tot} \times 2)/(K_{tot} + L_{tot})]$, which would be 0.7. Similarly, if we look at study B consisting of two other groups X and Y (Figure 4B) which also have 10 species each, but share a larger proportion of bacterial species (say 5) (Z_{tot}), their beta diversity 'BDI(*X*,*Y*)' would be expected to be 0.5. Their values

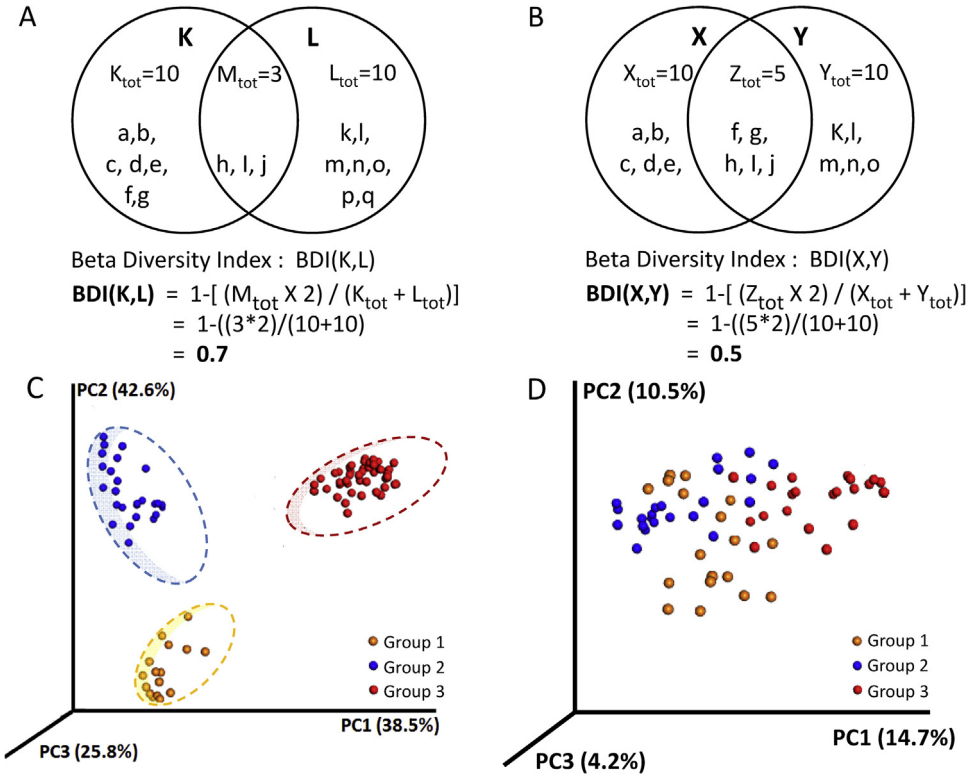


Figure 4 Examples of calculation of beta diversity index (panels A and B). In panel A, two specimens (K and L) have 10 species each (shown by lower case letters). Of these 3 species (h, i and j) are shared and 7 each are distinct being present in only one specimen. The beta diversity index for these two specimens can be calculated as 0.70. By comparison, in panel B, two specimens (X and Y), with similar number of species (10 each) have a greater overlap with 5 shared species and 5 unshared species each. In this case, the beta diversity index is calculated to be 0.50, a lower value indicating lower diversity (and greater sharing) than in panel 'A'. Panels C and D show examples of principal component analysis plots. In panel C, the three groups of specimens (shown using red, yellow and blue dots) have small intra-group beta-diversity values and large inter-group beta-diversity values. Hence, the dots for each group cluster together, but are placed at larger distances from the other groups. This clustering suggests that the microbiota in three groups are quite distinct from each other. By comparison, in panel D, the specimens from 3 groups show substantial overlap, indicating that microbiota in the three groups are not much different. In this situation, the inter-group and intra-group diversities are similar.

range from zero to one, with a high value indicating a lower level of similarity between the two groups, whereas a low value shows a higher level of similarity.

The example shown above (and in Figure 4) is a bit simplistic, since it looked at only the species counts. In real-life, somewhat more complex methods are used. Thus, for estimation of beta diversity, data from different specimens (e.g. from individual patients with a disease) are assembled into a table where each row represents a bacterial group (e.g. species) and each column represents a specimen. In this table, the values in individual cells contain observation counts for the particular bacterial group in a particular specimen (often after normalization, using one of several available normalization methods). Based on the above table, a distance/dissimilarity matrix is generated for each pair of specimens, using either a non-phylogeny-based or a phylogeny-based method. The methods based on non-phylogeny distances, such as Bray–Curtis, Euclidean, Jaccard or Hamming distance matrices, take into account abundances of various taxa in the two specimens being compared. The methods based on phylogeny-based distances also take into consideration the relative phylogenetic distances between various taxa; these are further of two types, i.e. un-weighted UniFrac method (which considers only the presence and absence of OTUs across specimens and the phylogenetic distances of taxa) and weighted UniFrac method (which also considers relative abundance information for each OTU and phylogenetic distances between them).[25]

The clustering patterns of specimens belonging to different groups, as represented in the beta-diversity matrices, can also be visualized in 2D or 3D plots using principal coordinate analysis. In these plots, the specimens with smaller distances between them appear to cluster closer together (Figure 4C) than specimens that have greater distances between them (Figure 4D).

Normalization of Data for Diversity Analyses

The absolute number of reads often varies across specimens included in a study. This poses a problem in alpha

and beta diversity analyses. For instance, if a specimen is sequenced in two different experiments to different depths (i.e. to obtain different numbers of total sequence reads), the experiment with larger depth is likely to pick up a larger number of unique taxa. Hence, during diversity analysis, it is important that data used from each specimen have the same depth of sequencing. This is done using a 'rarefaction technique', whereby, for each specimen, an identical number of sequence reads are randomly selected and used for analysis (referred to as normalization of data by equalizing the sampling depth of all the specimens to the one with the fewest reads). Besides simple data reduction, some advanced normalization techniques such as DESeq2, edgeR or Cumulative Sum Scaling (CSS) normalization have been proposed,[26] each with some advantages and limitations. The choice of normalization method should be such that it has the minimum risk of introducing a bias.

Identification of Specific Taxa that Differ Significantly Between Study Groups

Analysis of information on gut microbiota often involves comparison of two (and sometimes more than two) groups of specimens. These specimens can either belong to two different sets of individuals, as in case-control study design. Alternatively, the two groups of specimens can belong to the same of individuals but at different time points.

Case-Control Analysis (Unpaired Data)

Identification of core bacterial taxa that are significantly enriched in one group of specimens through case-control comparison study is an important aim of microbiota analysis. This analysis is used in situations when microbiota of a group of patients is compared to a group of controls, or when microbiota of two groups of patients, with different disease profiles, are compared. Thus, a specific bacterial taxon (phylum, family, order, genus or species) may be absent or have a low relative abundance in one group of specimens and a higher relative abundance in the other group of specimens.

For such case-control comparison for identification of differential abundance of taxa, unpaired statistical tests are used. The statistical test used may be parametric (e.g. unpaired *t* test), or non-parametric (e.g. Mann–Whitney *U* test or Wilcoxon's rank sum test); the latter is preferred since often the underlying data cannot be assumed to follow a normal distribution.

Comparison of Paired Data

Paired comparisons often refer to analysis of data when specimens are collected from the same set of subjects at two points, e.g. analysis of microbiota using stool specimens of a group of individuals before and after a particular intervention. Such analysis requires the use of a paired parametric (paired *t* test) or non-parametric (Wilcoxon's signed-rank test) statistical test.

Controlling for Multiple Comparisons

Due to the extremely complex nature of gut microbiota, each specimen may contain several hundred taxa. Thus, comparison of abundances of these taxa between two sets of specimens, whether paired or unpaired, implies the use of multiple statistical tests, one for each bacterial taxon with the null hypothesis that the groups do not differ.

Let us assume a situation where abundances of 100 bacteria taxa (often at different taxonomic ranks – phyla, orders, families, genera and species) are compared between a group of patients and controls, using the usual *P* value cut-off of 0.05 for significance. In such analysis, it can be shown that around 5 of these 100 comparisons can show *P* value < 0.05 just by chance, leading to a false conclusion of difference between groups where none actually exists. To avoid this, one of the several available methods (referred to as 'correction for multiple comparisons') is applied – these adjust the calculated *P* values to remove the effect of multiple comparisons, permitting the comparison of resultant 'adjusted *P* value' against the usual cut-off. Two methods that are most commonly used for this purpose are: Bonferroni correction[27] and Benjamini-Hochberg false discovery rate correction.[28]

Specific Measures for Comparison of Microbiota Composition Across Groups

Some specific measures have been proposed for comparison of gut microbiota composition between groups. These are meant to be used in specific situations, based on experience accumulated from studies on gut microbiota. For instance, based on data from patients with liver disease, a measure named as 'Cirrhosis Dysbiosis Ratio' has been proposed.[29] It is computed as the natural log (ln) of the ratio of aggregated abundance of autochthonous (Lachnospiraceae, Ruminococcaceae and Veillonellaceae) and non-autochthonous (Enterobacteriaceae and Bacteroidaceae) taxa. Thus:

$$Cirrhosisdysbiosisratio = loge[(a+b+c)/(d+e),$$

where a, b, c, d and e represent individual abundances of Lachnospiraceae, Ruminococcaceae, Veillonellaceae, Enterobacteriaceae and Bacteroidaceae, respectively.

Similarly, another measure 'Microbial Dysbiosis Index' has been proposed for use in Crohn's disease.[30]

16S rRNA Metagenomic Analysis Pipelines

Installation, configuration and use of individual tools for trimming, QC analysis, taxonomy profiling is somewhat complicated for clinicians. To overcome this difficulty, several bioinformatic pipelines have been developed to automate the various steps of 16S rRNA gene-based metagenome analysis. Some of the available software tools are

listed in Table 2. For comprehensive statistical, visual and comparative analysis of microbiome data, online servers, such as MicrobiomeAnalyst[31] and METAGENassist,[32] can also be used. The websites where these tools are hosted also provide the details of procedures used, and these should be useful to those readers who are interested in knowing more about the analytical algorithms and procedures.

Prediction of Functional Profiles from 16S rRNA Data

One of the purposes of studying microbiota is to know what metabolic functions it is capable of. Knowing the changes in functions of altered microbiota in a disease could help understand how the alteration in composition of the microbiota may relate to the pathogenesis of the particular disease.

It is possible to assess the function of microbiota from its composition using tools such as PICRUSt (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States)[33] and Tax4Fun,[34] which use data annotated by Greengenes and SILVA databases, respectively. These tools, use the data on composition of a particular microbial community (i.e. presence and abundance of various bacterial taxa in it) and the annotated genome sequences of these taxa to estimate the likely gene content of the community. Thus, these tools provide as output a table containing information about the abundance in each specimen of various genes coding for each metabolic pathway. These data can then be further analyzed using a statistical tool, such as STAMP (Statistical Analysis of Taxonomic and Functional Profiles).[35]

SHOTGUN METAGENOME (MICROBIOME) SEQUENCING

Sequencing of 16S rRNA gene or its segments, as discussed above is a powerful technique. However, it has the drawback that the determination of bacteria present in a specimen is based on the association of various sequences of the region of 16S rRNA gene studied with particular bacterial taxa. However, this association is not perfect. Second, this method is limited to the analysis of taxa for which informative sequences are included in the 16S rRNA reference databases. In addition, errors during sequencing may prevent accurate species assignment. More importantly, this method provides information only on taxonomic composition of the specimens studied, but cannot directly assess the biological functions of the microbial communities that these specimens represent. Though phylogenetic reconstruction tools (see "Prediction of functional profiles from 16S rRNA data" section) have been used to estimate the biological functions of a microbial community based on 16S rRNA data, these remain indirect and their accuracy is limited by the non-availability of compete genome annotations for a large number of bacteria.

Another method for study of microbiota involves sequencing of all the genomic material present in a specimen (referred to as 'microbiome' – a term used to denote the collective genetic material of the microorganisms in a particular environment, and 'metagenome' – all the genetic material of microbial or host origin contained in an environment) without the use of any culture method, instead of just the 16S rRNA gene. These methods have the advantage of providing information on the metabolic capabilities of the microbiota present in a particular specimen.

In this technique, DNA is extracted from all the cells in a microbial community. Thereafter, instead of targeting a specific genomic locus (e.g. 16S rRNA gene) for amplification, all the DNA is sheared into tiny fragments that are independently sequenced using a newer-generation sequencing technique to obtain information on the entire 'microbiome' or 'metagenome'. This provides several million sequence reads that belong to various locations on the genomes of the diverse bacteria, as also the host DNA, present in the starting specimen. These reads thus contain sequences not only of the taxonomically-informative 16S rRNA genes for the bacteria contained in the specimen, but also those corresponding to coding regions for enzymes that serve critical biological functions and are contained in the bacterial community. Hence, these metagenomic sequence data provide an opportunity to simultaneously explore two different aspects of the microbial community: which bacteria does it contain and what are these bacteria capable of doing? This capability has led to formation of major consortia (such as the human microbiome project) around the globe that are trying to use metagenomics as a tool for understanding the intestinal microbiota in human health and disease in several populations around the world.

In brief, the metagenomic sequence reads obtained in this procedure are mapped to a large number of bacterial reference genomes. Reads that uniquely map to adjacent locations on a reference genome are then assembled to form contigs, or continuous stretches of DNA sequence to reconstruct partial or complete draft microbial genomes. These contigs are then used to identify the gene families these belong to. By analysing the abundance of contigs for a particular gene family or those for all the genes in a particular pathway, a fairly accurate estimate of overall functional capabilities of all the bacteria present within a community can be obtained.

However, processing and analysis of shotgun metagenome sequence data poses several major challenges. First, the sequence data obtained relate not only to the bacterial DNA but also to the unwanted host DNA. In certain situations, for instance in analysis of human fecal specimens, these 'contaminant' human DNA sequences may overwhelm the bacteria DNA. To deal with this problem,

methods have been developed to selectively enrich microbial DNA sequences in the sequencing dataset by filtering host DNA sequences from the raw metagenome data, as the first step in data processing. Second, to ensure adequate representation of most of the bacteria present in a community, the amount of data generated needs to be very large. These large number of reads then have to be compared to the entire genomes of a large number of bacteria, many of which are quite closely related. This poses a huge computational challenge in terms of computer power. Third, the publically-available databases do not contain full reference genomic sequences for many bacteria. Finally, and possibly most importantly, metagenome data are several-fold more expensive to generate than the 16S rRNA data. Hence, this technique has not become commonplace for the study of gut microbiota.

OTHER NEWER TECHNIQUES (TABLE 1)

Meta-transcriptomics[36] is a tool similar to metagenomics, except that RNA, instead of DNA, is extracted and sequenced. The DNA analysis allows us to assess the functional capability of the genomic material contained in the bacteria present in a particular microbial community; however, one cannot be certain whether these genes are actually being expressed or not. The study of RNA allows us instead to study the expression of various genes in the bacterial genomes, taking us one step closer to the real-life functional characterization of the specimen.

It is theoretically possible to achieve an even better insight into the functional potential of microbiota in a particular specimen by study the profile of proteins contained in it (metaproteomics)[37] or various metabolites resulting from various metabolic pathways (metabolomics).[38] The use of these techniques is currently at an early stage, but with the ongoing development of tools for the measurement of these substances and the analysis of data generated, we should hear more about these in the coming years.

CONFLICTS OF INTEREST

The authors have none to declare.

ACKNOWLEDGEMENTS

The Biomedical Informatics Center at the authors' institution is supported by Indian Council of Medical Research (ICMR), New Delhi. ANS was supported during this work by ICMR.

REFERENCES

1. Marchesi JR, Ravel J. The vocabulary of microbiome research: a proposal. *Microbiome.* 2015;3:31.
2. Bouchet V, Huot H, Goldstein R. Molecular genetic basis of ribotyping. *Clin Microbiol Rev.* 2008;21:262–273. table of contents.
3. Inglis GD, Thomas MC, Thomas DK, Kalmokoff ML, Brooks SP, Selinger LB. Molecular methods to measure intestinal bacteria: a review. *J AOAC Int.* 2012;95:5–23.
4. Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, Parkhill J. Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. *BMC Microbiol.* 2016;16:274.
5. Tong M, Jacobs JP, McHardy IH, Braun J. Sampling of intestinal microbiota and targeted amplification of bacterial 16S rRNA genes for microbial ecologic analysis. *Curr Protoc Immunol.* 2014;107(7). 41.1–11.
6. Sinha R, Chen J, Amir A, et al. Collecting fecal samples for microbiome analyses in epidemiology studies. *Cancer Epidemiol Biomarkers Prev.* 2016;25:407–416.
7. Claesson MJ, Jeffery IB, Conde S, et al. Gut microbiota composition correlates with diet and health in the elderly. *Nature.* 2012;488:178–184.
8. Human Microbiome Project C. Structure, function and diversity of the healthy human microbiome. *Nature.* 2012;486:207–214.
9. Human Microbiome Project C. A framework for human microbiome research. *Nature.* 2012;486:215–221.
10. Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD. 16S rRNA gene sequencing of mock microbial populations – impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiol.* 2016;16:123.
11. Clooney AG, Fouhy F, Sleator RD, et al. Comparing apples and oranges? Next generation sequencing and its impact on microbiome analysis. *PLOS ONE.* 2016;11:e0148028.
12. Claesson MJ, Wang Q, O'Sullivan O, 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.* 2010;38:e200.
13. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet.* 2008;9:387–402.
14. Chaudhary N, Sharma AK, Agarwal P, Gupta A, Sharma VK. 16S classifier: a tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets. *PLOS ONE.* 2015;10:e0116106.
15. Blaxter M, Mann J, Chapman T, et al. Defining operational taxonomic units using DNA barcode data. *Philos Trans R Soc Lond Ser B Biol Sci.* 2005;360:1935–1943.
16. Drancourt M, Bollet C, Carlioz A, Martelin R, Gayral JP, Raoult D. 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. *J Clin Microbiol.* 2000;38:3623–3630.
17. Yilmaz P, Parfrey LW, Yarza P, et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* 2014;42:D643–D648.
18. Cole JR, Wang Q, Fish JA, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 2014;42:D633–D642.
19. DeSantis TZ, Hugenholtz P, Larsen N, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol.* 2006;72:5069–5072.
20. Kim OS, Cho YJ, Lee K, et al. Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *Int J Syst Evol Microbiol.* 2012;62:716–721.
21. The RC. RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Res.* 2017;45:D128–D134.
22. Ritari J, Salojarvi J, Lahti L, de Vos WM. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genomics.* 2015;16:1056.
23. Carding S, Verbeke K, Vipond DT, Corfe BM, Owen LJ. Dysbiosis of the gut microbiota in disease. *Microb Ecol Health Dis.* 2015;26:26191.
24. Gotelli NJ, Chao A. *Measuring and Estimating Species Richness, Species Diversity, and Biotic Similarity from Sampling Data A2 –*

Levin, Simon A. Encyclopedia of Biodiversity. 2nd ed. Waltham: Academic Press; 2013:195–211.
25. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. *ISME J.* 2011;5:169–172.
26. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol.* 2014;10:e1003531.
27. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ.* 1995;310:170.
28. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. *Stat Med.* 1990;9:811–818.
29. Bajaj JS, Heuman DM, Hylemon PB, et al. Altered profile of human gut microbiome is associated with cirrhosis and its complications. *J Hepatol.* 2014;60:940–947.
30. Gevers D, Kugathasan S, Denson LA, et al. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe.* 2014;15:382–392.
31. Dhariwal A, Chong J, Habib S, King IL, Agellon LB, Xia J. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.* 2017.
32. Arndt D, Xia J, Liu Y, et al. METAGENassist: a comprehensive web server for comparative metagenomics. *Nucleic Acids Res.* 2012;40:W88–W95.
33. Langille MG, Zaneveld J, Caporaso JG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol.* 2013;31:814–821.
34. Asshauer KP, Wemheuer B, Daniel R, Meinicke P. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics.* 2015;31:2882–2884.
35. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics.* 2014;30:3123–3124.
36. Bashiardes S, Zilberman-Schapira G, Elinav E. Use of metatranscriptomics in microbiome research. *Bioinform Biol Insights.* 2016;10:19–25.
37. Kolmeder CA, de Vos WM. Metaproteomics of our microbiome – developing insight in function and activity in man and model systems. *J Proteomics.* 2014;97:3–16.
38. Larsen PE, Dai Y. Metabolome of human gut microbiome is predictive of host dysbiosis. *GigaScience.* 2015;4:42.