frontiers in Microbiology

TECHNOLOGY REPORT
published: 20 April 2016
doi: 10.3389/fmicb.2016.00459

Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics

Juan Jovel[1]*†, Jordan Patterson[1]†, Weiwei Wang[1], Naomi Hotte[1], Sandra O'Keefe[1], Troy Mitchel[1], Troy Perry[1], Dina Kao[1], Andrew L. Mason[1], Karen L. Madsen[1] and Gane K.-S. Wong[1, 2, 3]*

[1] Department of Medicine, University of Alberta, Edmonton, AB, Canada, [2] Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada, [3] BGI-Shenzhen, Shenzhen, China

OPEN ACCESS

Edited by:
Martha E. Trujillo,
Universidad de Salamanca, Spain

Reviewed by:
Ranjit Kumar,
University of Alabama at Birmingham, USA
Henning Seedorf,
Temasek Life Sciences Laboratory, Singapore

Correspondence:
Juan Jovel
jovel@ualberta.ca;
Gane K.-S. Wong
gane@ualberta.ca

† These authors have contributed equally to this work.

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *08 January 2016*
Accepted: *21 March 2016*
Published: *20 April 2016*

Citation:
Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, Perry T, Kao D, Mason AL, Madsen KL and Wong GK-S (2016) Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. Front. Microbiol. 7:459. doi: 10.3389/fmicb.2016.00459

The advent of next generation sequencing (NGS) has enabled investigations of the gut microbiome with unprecedented resolution and throughput. This has stimulated the development of sophisticated bioinformatics tools to analyze the massive amounts of data generated. Researchers therefore need a clear understanding of the key concepts required for the design, execution and interpretation of NGS experiments on microbiomes. We conducted a literature review and used our own data to determine which approaches work best. The two main approaches for analyzing the microbiome, 16S ribosomal RNA (rRNA) gene amplicons and shotgun metagenomics, are illustrated with analyses of libraries designed to highlight their strengths and weaknesses. Several methods for taxonomic classification of bacterial sequences are discussed. We present simulations to assess the number of sequences that are required to perform reliable appraisals of bacterial community structure. To the extent that fluctuations in the diversity of gut bacterial populations correlate with health and disease, we emphasize various techniques for the analysis of bacterial communities within samples (α-diversity) and between samples (β-diversity). Finally, we demonstrate techniques to infer the metabolic capabilities of a bacteria community from these 16S and shotgun data.

Keywords: gut microbiome, 16S rRNA gene sequencing, shotgun metagenomics, bioinformatics, taxonomic classification, diversity analysis, functional profiling

INTRODUCTION

High-throughput comparative metagenomics enabled by development of next-generation sequencing (NGS) platforms (Mardis, 2008; Novais and Thorstenson, 2011) has led to an outburst of research endeavors that have rapidly advanced our understanding of the composition and function of bacterial populations in very diverse environments (Ley et al., 2006; Garrett et al., 2010; Caporaso et al., 2011; Bolhuis et al., 2014; Huttenhower et al., 2014a; Norman et al., 2014; Yoon et al., 2015). In the clinical context, the human gut microbiome has been the subject of intense investigation, which has revealed a sophisticated interplay between the microbiome and the host immune system and metabolism (Garrett et al., 2010; Brown et al., 2013; Huttenhower et al., 2014a; Martín et al., 2014; Broderick, 2015). For instance, it is well known that bacteria aid in many important metabolic pathways, including synthesis of essential compounds like secondary bile acids and short-chain fatty acids (Flint et al., 2012; Nicholson et al., 2012). Moreover, reduced diversity and/or imbalances in the gut microbiome have been associated with a variety of phenotypes, including obesity (Turnbaugh et al., 2009; Turnbaugh and Gordon, 2009), inflammatory bowel diseases (IBD) (Knights et al., 2013; Huttenhower et al., 2014b; Kostic et al., 2014; Norman et al., 2015), type II diabetes (T2D) (Qin et al., 2012; Hartstra et al., 2015), fatty liver disease (Arslan, 2014), and numerous additional disorders (Bhattacharjee and Lukiw, 2013; Dinan et al., 2014; Bajaj et al., 2015; Dash et al., 2015). The mechanisms whereby bacteria affect the host physiology are also well appreciated from a gene content/functional perspective. For example, both IBD and obesity are associated with enrichment of enzymes in the nitrate reductase pathway, the metabolism of choline and p-cresol, as well as the phosphotransferase system, required for assimilation of dietary carbohydrates (Greenblum et al., 2012; Levy and Borenstein, 2014). Bacteria able to synthesize short chain fatty acids, including acetate, butyrate, and propionate, have been found to be critical for colonocyte homeostasis, and their imbalance has been documented in diseases such like IBD and T2D (Qin et al., 2012; Brestoff and Artis, 2013; Kostic et al., 2014; Vital et al., 2014). For the most part, microbiome studies have focussed primarily on the structure and function of bacterial communities, fungi and viruses have received less attention thus far, but are starting to gain momentum (Reyes et al., 2010; Norman et al., 2014, 2015; Wang et al., 2015). There is also renewed interest in better understanding gaseous products from the gut microbiome, including carbon dioxide, hydrogen, methane and hydrogen sulfide (Pimentel et al., 2013). Importantly, methanogenesis from Archaea, mainly *Metanobrevibacter smithii*, is an important source of energy. It therefore influences metabolism and is associated with obesity, diabetes mellitus and other metabolic disorders (Pimentel et al., 2013; Barlow et al., 2015).

Most of the studies to understand bacterial population dynamics have been conducted with metagenomic approaches that are simple and cost-effective, although metatranscriptomic, proteomic, and metabolomic approaches are becoming popular too (Franzosa et al., 2014, 2015; Morgan and Huttenhower, 2014; Heinken and Thiele, 2015; Schaubeck et al., 2015; Yen et al., 2015). Together, these studies promise to provide a high-resolution picture of bacteria-host interactions that may lead to disease (Franzosa et al., 2015). Whole-metagenome shotgun analyses are accomplished by unrestricted sequencing of the genome of all microorganisms present in a sample (hereafter referred to as shotgun libraries); alternatively, inferences can be made by sequencing PCR amplicons from the ribosomal 16S RNA gene (hereafter referred to as 16S libraries), whose domain is restricted to bacteria and archaea (Janda and Abbott, 2007). Data generated by each of these approaches requires sophisticated computational methods and extensive hardware resources for their analysis (Gevers et al., 2012). This poses a significant challenge for microbiologists and clinical researchers interested in diverse aspects of the microbiota. Fortunately, the open-source software community has been diligent in developing user-friendly bioinformatics tools required for the analyses of bacterial NGS datasets. This article provides a compendium of good practices for the analysis of NGS microbiome libraries sequenced with the MiSeq platform but, for the most part, our suggestions are applicable to data generated with other NGS platforms. Using gut microbiome datasets specially designed to illustrate the strengths and weaknesses of 16S or shotgun libraries, we describe several methods for performing taxonomical classification of bacterial sequences, assessment of bacterial diversity within and between samples, and inference of the metabolic capabilities associated with the bacterial microbiome.

PRE-PROCESSING TO ELIMINATE UNINFORMATIVE DATA

Removal of adapters, PCR primers and low quality bases is essential for effective analyses of NGS libraries, and a variety of user-friendly tools have been developed for this purpose. The current Illumina platforms output quality scores "Q" that fit into a 0–41 scale (Q10 corresponds to 1 expected error for every 10 sequenced bases; Q20 = 1 error for every 100 bases, and so on). Setting a quality threshold remains at the researcher's discretion; however, it is good practice to use only those sequences with the highest possible quality. In our experience, sacrificing sequences with low quality scores often improves the accuracy of the analyses by a significant margin. The gain in precision by trimming data is more significant for 16S data than it is for shotgun data, as clustering algorithms have been designed to detect minor differences along the sequence of the 16S rRNA gene. Most sequencing platforms are capable of performing paired-end sequencing. This means that both ends (end1 and end2) of the library insert are sequenced separately. End1 and end2 may or may not overlap and together are referred to as a "read." With Illumina chemistry, bases at the front (5′ end) of each sequence generally exhibit higher quality than those at the back (3′ end) (**Supplemental Figure 1**); however, in the case of 16S libraries, the primers used for amplification can also generate regions of low quality at the front of each sequence. For shotgun data it is recommended to use trimming software that remove low-quality bases from both termini of each sequence, like cutadapt (Martin), sickle (Joshi and Fass, 2011), or fastqMcf (Aronesty, 2011). For 16S rRNA gene sequences, it is advisable to trim sequences along the entire length, starting from the 5′ end and using a quality threshold as high as possible, while leaving sufficient sequences to perform the analyses. Assembly of overlapping paired end sequences is advisable as long as the quality of overlapping regions is high enough to generate a consensus sequence with high quality scores.

TAXONOMICAL CLASSIFICATION OF BACTERIAL SEQUENCES

Precise taxonomy assignments based on sequence alignments remain a computational challenge for both 16S and shotgun libraries, because of the short NGS read lengths. Prior to taxonomic classification, gene marker amplicon sequences, like regions of the bacterial 16S rRNA gene, are clustered by two main approaches (Sun et al., 2012; Chen et al., 2013). First, sequences can be clustered into phylotypes according to their similarity to previously annotated sequences in a reference database (Liu et al., 2008). Second, operational taxonomic units (OTUs) can be constructed by clustering sequences *de novo*, purely based on their similarity (Schloss and Westcott, 2011; Sun et al., 2012), which is computationally much more intensive. A hybrid method that combines both approaches is therefore recommended. In all cases, an arbitrary similarity threshold is used to differentiate clusters. The 99% similarity threshold is generally accepted as a good proxy for species (Stackebrandt and Ebers, 2006). However, this threshold is often insufficient to discriminate between closely related species, such as different members of the Enterobacteriaceae, Clostridiaceae, and Peptostreptococcaceae families. Importantly, higher resolution analytical tools have been published that overcome some of the limitations associated with clustering algorithms (Eren et al., 2013, 2014; Tikhonov et al., 2015).

Comprehensive reference databases have been compiled for annotation of sequenced bacteria metagenomes. For 16S rRNA genes, this includes the Greengenes database (DeSantis et al., 2006), the Ribosomal Database Project (RDP) (Cole et al., 2014), and SILVA (Quast et al., 2013). In addition to their extensive catalogs of curated 16S rRNA sequences, available for downloading, each of those portals also offers a series of bioinformatics tools for analysis of NGS sequences. Comprehensive analysis servers like MG-RAST are also publicly available, which already contain updated databases for annotation purposes (Meyer et al., 2008). More specifically, the human microbiome project (HMP) keeps a curated collection of sequences of microorganisms associated with the human body, including eukaryotes, bacteria, archaea and viruses, from both shotgun and 16S sequencing projects (C. Human Microbiome Project, 2012a,b). One of the approaches to increasing the resolution of taxonomical classification of sequences is to compile databases containing only the sequences likely to exist in the environment under study. For example, specialized databases comprising only members of the human intestinal microbiota (Ritari et al., 2015; Forster et al., 2016) have been created.

Robust bioinformatics approaches have also been developed for analysis of shotgun data (Riesenfeld et al., 2004; Schloss and Handelsman, 2008; Wu and Eisen, 2008; Huson et al., 2011; Boisvert et al., 2012; Gevers et al., 2012; Kultima et al., 2012; Namiki et al., 2012; Segata et al., 2012). Unique clade-specific marker genes (Mende et al., 2013) and lowest common ancestor (LCA) positioning approaches are among the most popular. For the former, a gene marker catalog is pre-computed from previously sequenced bacterial genomes and sequences are taxonomically classified by querying the catalog. For the LCA approach, pre-aligned sequences are hierarchically classified on a taxonomy tree using a placement algorithm (Aho et al., 1973; Huson et al., 2011). Sequences that surpass a dissimilarity threshold (bit-score) are progressively placed on higher taxonomy levels.

VALIDATION OF BIOINFORMATICS APPROACHES IN BACTERIAL COMMUNITIES

To demonstrate some of the most common approaches and pipelines used for taxonomy assignments in 16S and shotgun libraries, we created an artificial bacterial population using DNA from *Salmonella enterica*, *Streptococcus pyogenes*, *Escherichia coli*, *Lactobacillus helveticus*, *Lactobacillus delbrueckii*, *Lactobacillus plantarum*, *Clostridium sordelli*, *Bacteroides thetaiotaomicron*, *Bacteroides vulgatus*, *Bifidobacterium breve,* and *Bifidobacterium animalis*. We then constructed 16S and shotgun libraries in parallel using the NEXTflex 16S V4 Amplicon-Seq (BioO Scientific) and the Nextera XT (Illumina) kits, respectively. The raw data of all libraries generated during this study is publicly available at the Sequence Read Archive (SRA) portal of NCBI under accession number SRP059928.

For the analysis of 16S amplicon libraries, we evaluated QIIME (Caporaso et al., 2010; Navas-Molina et al., 2013) and mothur (Schloss et al., 2009), the most widely adopted pipelines, and the MiSeq Reporter v2.5 (MRS; the software developed by Illumina and accompanying the MiSeq instrument) pipeline, all with default parameters. At the genus level, all pipelines produced similar results, but the Pearson correlation coefficient between the expected (input) and obtained relative abundance was somewhat higher for QIIME (**Figure 1A**). We therefore selected QIIME for our subsequent analyses; however, we do not discourage the use of mothur, which is also a reliable pipeline. None of the 16S pipelines performed satisfactorily at the species level.

We conducted taxonomy assignments using end1, end2, or both paired ends. When using Illumina chemistry, end1 typically exhibits higher quality than end2 (**Supplemental Figure 1**); accordingly, end1 provided a somewhat more accurate classification than end2 or paired ends (**Supplemental Figure 2**). However, the V4 variable region of the 16S rRNA gene is relatively short and in most cases will be covered by any one of the two ends (250 nt in this case); as such, these results may only reflect the higher quality of end1. For single ends, the best results were obtained with the *pick_open_otus.py* script from QIIME to cluster sequences (**Supplemental Figure 2**). Chimeric sequences can be artifactually generated when PCR amplification of the 16S region of interest is incomplete and the resultant partial sequences serve as primers that recombine with heterologous molecules containing a similar 3′ moiety. Several bioinformatics approaches have been developed for detection and removal of chimeric sequences. We used the USEARCH tool (Edgar, 2010) to remove chimeras. However, a desirable approach is to prevent formation of such chimeras *in vitro*, using high fidelity amplification protocols like LEA-Seq (Faith et al., 2013). Sequences were initially clustered into phylotypes using the Greengenes database of 16S rRNA sequences (DeSantis et al., 2006) as reference, while more dissimilar sequences were clustered *de novo* into OTUs. Taxonomy was then assigned using the RDP classifier, using the UCLUST method (Wang et al., 2007). RTAX (Soergel et al., 2012), a method embedded

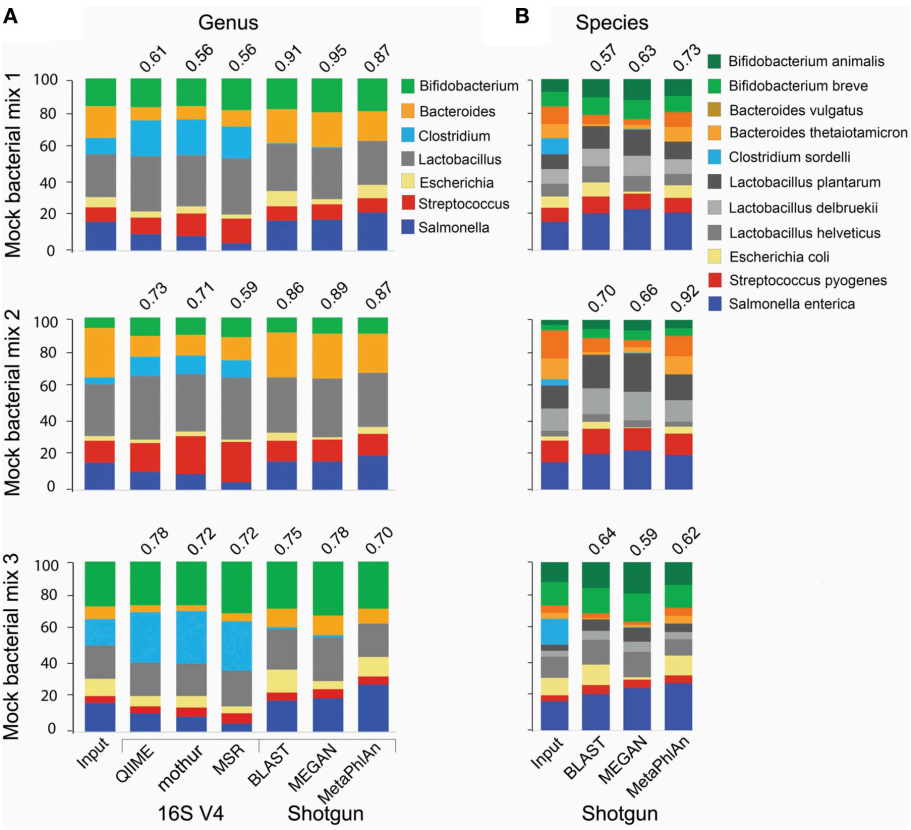

FIGURE 1 | Comparison of taxonomic analyses of a low complexity artificial microbial population using 16S amplicon or shotgun metagenomic approaches. Eleven bacterial species (representing 7 genera) were cultured under standard laboratory conditions. DNA was extracted using the FastDNA spin kit for feces (MPBio). 16S amplicon and shotgun metagenomics libraries were constructed using the NEXTflex 16S V4 Amplicon-Seq (BioO Scientific) and the Nextera XT (Illumina) kits, respectively. Libraries were paired-end sequenced on a MiSeq sequencer using a 500-cycle kit. For 16S libraries, sequences were trimmed with the "split_fastq_libraries.py" script from QIIME. Default parameters were used, with the exception that the quality threshold for trimming was raised to 30. PCR primer sequences were trimmed with in-house Perl scripts. Shotgun metagenomics libraries were trimmed with the fastqMcf tool, and a quality threshold of 15. The relative abundance of each species was determined with the software indicated at the bottom of the bar graph, using default parameters, at the genus **(A)** or species **(B)** levels. The Pearson correlation coefficient between the expected (Input) relative abundance and the classification performed by each program is indicated on top of the bar graph.

in QIIME, and UPARSE (Edgar, 2013) are algorithms especially designed to take advantage of mate pairs information. For paired-end analysis, the UPARSE pipeline (Edgar, 2013) produced more satisfactory results than the RTAX method (Soergel et al., 2012; **Supplemental Figure 2**). Irrespective of the method used for clustering, we found a consistent over-representation of sequences in the *Clostridium* and *Lactobacillus* genera. These two genera contain sequences that are perfectly complementary to the primers used for amplification, while at least one mismatch is found in the rest of genera included in our experimental (mock) bacterial population. This demonstrates how subtle differences in primer binding sites within the 16S rRNA gene sequences lead to biased estimates of relative abundance. Other primers have been reported to present biases, for instance the primer pair 27F/338R results in underrepresentation of *Bifidobacterium* (Martínez et al., 2009; Kuczynski et al., 2010). In our study, the detection of some *Clostridium*, *Escherichia* and *Salmonella* sequences was only possible after computational extraction of representative sequences of OTUs and blasting them against both the nr/nt and the 16S ribosomal RNA databases from NCBI. In general, sequences in the *Enterobacteriaceae* family and the *Clostridiales* order were poorly resolved using the 16S V4 or V3-V4 regions (**Figure 2A**), and this seems to be the case with *Enterobacteriaceae* for other 16S variable regions as well (Chakravorty et al., 2007).

For shotgun libraries, we compared BLAST top hits, the MEtaGenome ANalyzer MEGAN5, and Metagenomic Phylogenetic Analysis (MetaPhlAn) approaches; however, we do acknowledge that many other excellent tools have also been developed, including PhymmBL (Brady and Salzberg, 2009, 2011), PhyloSift (Darling et al., 2014), MOCAT (Kultima et al., 2012), Kraken (Wood and Salzberg, 2014), CLARK (Ounit et al., 2015), and kallisto (Schaeffer et al., 2015). BLAST top hits corresponded to the correct genus in all instances (**Figure 1A**), but there were inaccuracies at finer resolutions. For example, some *C. sordelli* sequences were

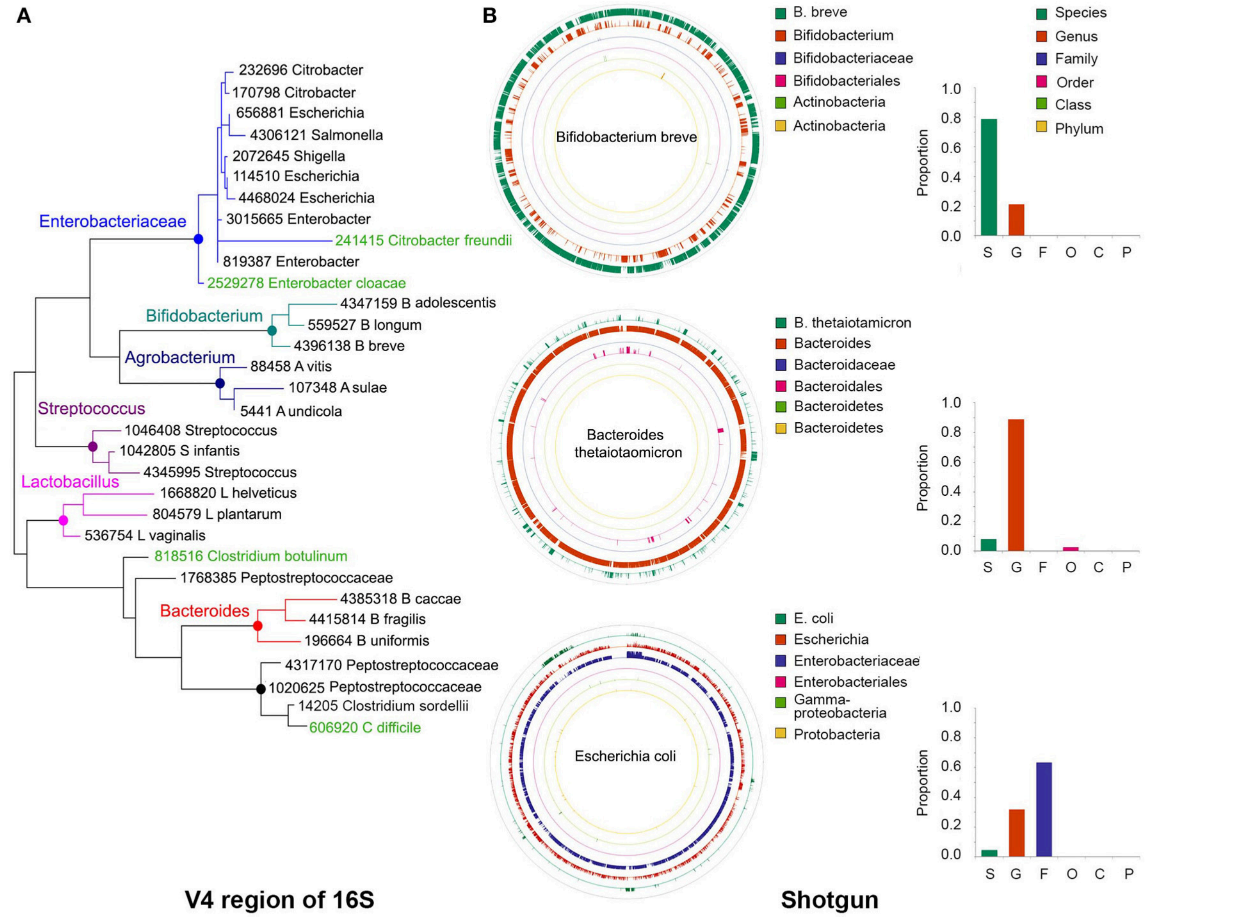

FIGURE 2 | Precision of taxonomy assignments is affected by highly similar sequences in different taxa. (A) For the 16S libraries described in **Figure 1**, sequences were clustered into operational taxonomic units (OTUs) using a 97% similarity threshold and taxonomy assignments were performed with the RDP classifier. Sequences from OTUs classified as Bifidobacterium ($n = 3$), Agrobacterium ($n = 3$), Streptococcus ($n = 3$), Lactobacillus ($n = 3$), Bacteroides ($n = 3$), Peptostreptococcaceae ($n = 4$), or Enterobacteriaceae ($n = 9$) were randomly extracted and aligned to the Greengenes database to extract the closest relative (best hit). In addition, we included Greengenes 16S rRNA gene sequences (in green) from *Clostridium difficile* and *C. botulinum* as reference for Peptostreptococcaceae and *Citrobacter freundii* and *Enterobacter cloacae* as reference for Enterobacteriaceae. The V4 region of the 16S rRNA gene was cropped from the Greengenes sequences to construct a phylogenetic tree with MEGA-6, using UPGMA hierarchical clustering and 10,000 bootstraps. **(B)** Sequences from our bacterial populations in **Figure 1** were aligned against the NCBI nt and human microbiome project (HMP) databases to identify the most similar reference genome. For each bacterium, a simulated library was created by segmenting the reference genome sequence into 500 nt stretches (250 nt paired ends in a head-to-tail orientation), iterating the process to generate ~1.5 million sequences. This simulated library was aligned back to the reference genome and the taxonomy resolved with MEGAN5. As examples, we show the reads classification of *Bifidobacterium breve*, *Bacteroides thetaiotamicron*, and *Escherichia coli*, which accumulated a large proportion of reads that could be resolved at the species, genus or family levels, respectively. Color-matched bars on the right show the proportion of reads accumulated at each level for these particular examples. S, species; G, genus; F, family; O, order; C, class; P, phylum.

erroneously assigned to *C. difficile* or *C. botulinum* because no reference genome was available at the time we conducted the alignments. MEGAN5 (Huson et al., 2011) hierarchically classifies pre-aligned sequences on a taxonomy tree using an LCA algorithm. As BLAST can be prohibitively slow, the LAST aligner was used in comparison for the same analysis (Kielbasa et al., 2011). LAST alignments were several orders of magnitude faster than BLAST, with comparable sensitivity (**Supplemental Figure 3**). The LAST-aligned sequences were fed to MEGAN5 for taxonomic assignments (Huson et al., 2011). Classification with LAST/MEGAN5 was as accurate as BLAST top hits at the genus level (**Figure 1A**). Lastly, we used MetaPhlAn, which infers taxonomy based on unique clade-specific marker genes. MetaPhlAn classification at the genus level was as accurate as the one performed by the other two tools (**Figure 1A**). The three tools correctly classified all species included in our mock populations and also provided a good approximation to their expected relative abundance (**Figure 1B**), but MetaPhlAn outperformed the other two tools in terms of precision and speed. Furthermore, utilization and installation of MetaPhlAn is much simpler than BLAST or MEGAN5 and it requires less computational processing.

SEQUENCES WITH LOW RESOLUTION CANNOT BE CLASSIFIED AT THE SPECIES LEVEL

Resolving the taxonomy of 16S rRNA gene sequences can be problematic based on a limited segment of the 16S rRNA gene, such as the V4 region. In many cases, the sequence to be classified is nearly identical to several other bacterial sequences in the reference database. Similarly, for shotgun metagenomic analyses, when only parts of the bacterial genome are recovered, the classification at a taxonomic level will depend on the degree of conservation of the available sequences. Thus, the taxonomy of species that contain highly similar sequences will be more difficult to resolve, and the analyses will accrue a larger proportion of reads at the higher levels of the taxonomy tree.

For instance, the phylogenetic tree depicted in **Figure 2A** was built using the V4 region of few representative sequences in the Greengenes database (DeSantis et al., 2006; see **Figure 2** caption for details). It can be seen that sequences in some genera form discrete branches on the tree, such as *Lactobacillus*, *Streptococcus*, *Bifidobacterium*, and *Bacteroides*. Other more closely related bacteria intertwine and cannot be delineated solely on the basis of their differences along the V4 region, such as those within the *Enterobacteriaceae* and *Peptostreptococcaceae* family. It has been reported that for ∼42% of bacterial genera there will be pairs of sequences within genus that cannot be easily separated because their 16S rRNA gene sequences are more than 97% similar (Vetrovsky and Baldrian, 2013).

In a taxonomy tree, the lowest common ancestor of two taxa, *a* and *b*, is the immediate upper node that includes *a* and *b* as descendants. When a sequence aligns equally well to nodes *a* and *b*, that sequence will be annotated with the taxonomy corresponding to the lowest common ancestor, which is less accurate but more certain. Using the LCA approach, the lack of resolution of bacterial sequences in certain parts of the genome will also affect the taxonomic classification of shotgun libraries. For example, in bacteria with highly divergent genomes like *Bifidobacterium breve*, a large proportion of the genome can be resolved at the species level (**Figure 2B**, green outer circle), whereas in other genomes like those of *Bacteroides thetaiotamicron* and *Escherichia coli*, the majority of their sequences can only be resolved at the genus (**Figure 2B**, orange ring) and family (**Figure 2B**, purple ring) levels, respectively. MetaPhlAn does not suffer from this problem, as marker genes are chosen based on their uniqueness, with the caveat that sufficient sequences are needed to warrant their representation in shotgun libraries.

In general, classification of whole metagenome sequences improves when more dissimilar regions of the genomes, with greater discriminatory power, are included in the sequenced pool. When relatively large amounts of sequences are available, it is convenient to assemble individual reads into larger fragments, technically known as contigs, which are more amenable for taxonomic classifications. A series of software to assemble metagenomics data have been developed including Ray Meta (Boisvert et al., 2012), MetaVelvet (Namiki et al., 2012), MetaQUAST (Mikheenko et al., 2015), and MetAMOS (Treangen et al., 2013), among others. To increase efficiency, it is also possible to combine different samples in a single assembly procedure, while maintaining the ability to trace the origin of each assembled read.

ASSESSMENT OF REQUIRED SEQUENCING DEPTH

For illustration purposes, we prepared a series of samples of progressively greater complexity. At the low-end, we sequenced the metagenome associated with grains of Kefir, a form of fermented milk with probiotic properties (Nielsen et al., 2014; **Supplemental Figure 4**). The higher complexity libraries included stool samples from subjects affected by Crohn's disease, *C. difficile* infection, and a healthy individual. For comparison, we cultured three experimental (mock) bacteria communities containing 19 species from 12 genera (Mix7-9). All libraries were sequenced at an average depth of $\sim 8.5 \times 10^5$ paired-end reads (minimum 1.57×10^5; maximum 1.67×10^6).

To investigate the minimal sequencing depth sufficient for accurately profiling bacterial community composition, we randomly sampled our libraries at depths of 500, 1000, 5000, 10,000, 50,000, and 100,000 reads. At each depth, sampling and analyses were repeated 20 times. As an example, we show that the taxonomic classification for each type of library at sequencing depths of 1000 and 50,000 was surprisingly consistent (**Figures 3A,B**). It is expected that taxonomic classification performed with each method will be to some extent divergent, as the resolution of the sequences used for taxonomic assignments is distinct and variable depending on which region of the genome is captured in shotgun surveys, which variable region of the 16S rRNA gene is used, and which composition of species is present in the community under analysis. However, the general pattern of relative abundance of taxa was often observed to be similar although the concordance of 16s vs. shotgun methods was higher for simpler bacterial communities, as seen with the Kefir's community (**Figures 3A,B**). In the sample from the CD patient, the most abundant genus (*Lactobacillus*) was detected by both methods (gray bar), but the second was identified as *Klebsiella* in 16S and *Citrobacter* in the shotgun libraries (**Figures 3A,B**). This ambiguity likely occurs because the 16S rRNA gene sequences of these two genera share > 96% similarity. Many other taxa, like *Bifidobacterium* (**Figures 3A,B**) were consistently identified because they are phylogenetically more distant from the other taxa present. For the mock populations, all genera ($n = 12$) were found in shotgun libraries at both depths, but 16S libraries did not allow detection of the *Akkermansia* or *Clostridium* genera, even though they were ∼5% of Mix-9. As expected, increasing sampling depth led to increased detection of taxa; with 1000 sequences 48 and 58 taxa were detected in 16S or shotgun libraries, respectively, and with 50,000 sequences this increased to 72 and 128. Based on our experimental bacterial mock populations, it is clear that some of the assignments are spurious and increasing sequencing depth augments the artifact. Of note, *Propionibacterium* was not included in our experimental mixes but was found in both

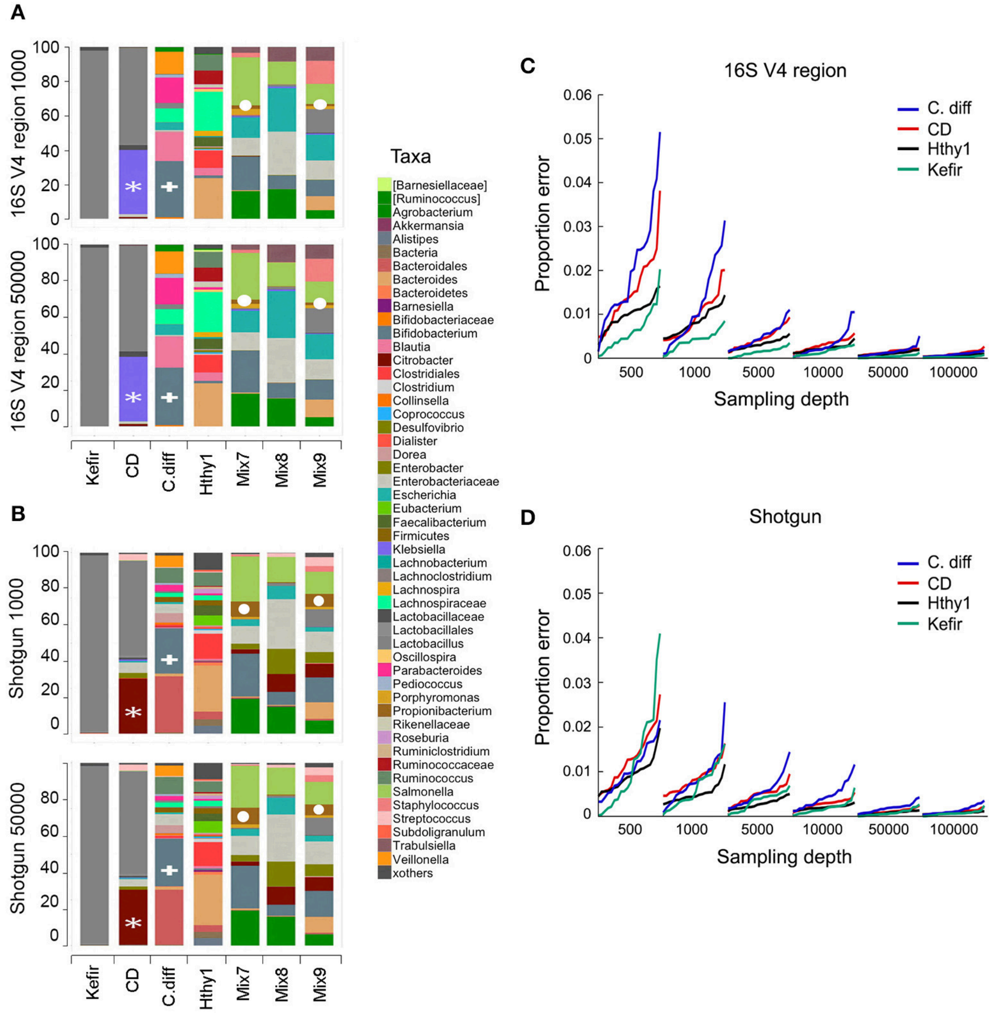

FIGURE 3 | Number of sequences required for taxonomic classification of samples with varying diversity. A series of samples were chosen to assess the effect of library complexity on the accuracy of taxonomy assignments and estimation of diversity of bacterial populations. Kefir represents the lowest point in the bacterial diversity spectrum, followed by a patient affected by Crohn's disease (CD), another one recovered from *C. difficile* infection (C. diff), a healthy individual (Hthy1) and three artificial mixes of bacteria (Mix7-9). **(A,B)** Libraries were randomly sampled at depths of 500, 1000, 5000, 10,000, 50,000 and 100,000 reads. End1 16S rRNA gene sequences were classified with QIIME using the closed reference method to cluster OTUs and a similarity threshold of 97%. Paired-end shotgun metagenomics sequences were aligned with LAST and taxonomically classified with MEGAN5. Each random sampling was repeated 20 times. As an example, the relative abundance of taxa for one of these samplings at a depth of 1000 or 50,000 sequences is presented for 16S and shotgun metagenomics libraries. A white asterisk indicates a group of bacterial sequences identified as *Citrobacter* in the shotgun panel and *Klebsiella* in the 16S panel. *Bifidobacterium* is indicated with a white plus sign. *Propionibacterium* is indicated with a white circle. **(C,D)** For each taxa detected and for each random sample, the proportion error was calculated as the difference between the proportion that each taxon represented in the whole library (i.e., with the maximum number of reads) and in the random sample. This difference was weighted by the proportion that each taxon represented in the whole library. We present the arithmetic mean of all weighted differences for each of the 20 random samples.

types of libraries, indicative of contamination (**Figures 3A,B**). Indeed, environmental contamination poses a serious challenge for construction of NGS libraries (Laurence et al., 2014; Salter et al., 2014; Strong et al., 2014; Weiss et al., 2014).

Increasing the number of sequences results in more consistent estimations of bacteria relative abundance. To illustrate this point, we sampled reads from each library at various depths (500–100,000) and compared the proportion of each taxon to the full library for the Kefir, CD, C. diff., and healthy samples (**Figures 3C,D**). For each depth, we repeated sampling 20 times. We report the weighted arithmetic mean of the differences in proportion between the sampling and the full library. In general, the proportion error and its variance decrease with increasing sampling depth (**Figures 3C,D**). The number of sequences required per library will ultimately depend on the goals of the study and the type of analysis to be conducted (Ni et al., 2013).

In bacterial ecology, alpha (α) diversity refers to the species composition in sampling units, usually at a local scale (Whittaker, 1972; Lozupone and Knight, 2008). While the local scale concept is somewhat diffuse in population ecology, the compartmentalized nature of the human (or mouse) body creates well defined microbial communities (i.e., GI tract, mouth, etc.) on which α-diversity can be estimated for comparison purposes. We used the Shannon diversity and equitability indices (Shannon, 1948) as estimators of α-diversity for each of the random samples extracted from our libraries (**Supplemental Figure 5**). The Shannon diversity index is a sum of the proportion of each species relative to the total number of species in the community under analysis and therefore accounts for both abundance and evenness (Shannon, 1948). It was nearly identical at 1000 and 50000 reads with only a small variance over multiple repetitions for both 16S and shotgun libraries. The trend was the same for both methods: increasing Shannon diversity values were found from the Kefir sample, followed by the CD, *C. difficile*, and the sample from the healthy subject. As noted above, the Kefir microbiota only includes few species of bacteria and yeast (**Supplemental Figure 4**), and both CD and *C. difficile* infection have been reported associated with reduction of faecal bacterial diversity in the patients' stools (Chang et al., 2008; Antharam et al., 2013; Vincent et al., 2013; Kostic et al., 2014; Norman et al., 2015). This was well recapitulated by the Shannon diversity index (**Supplemental Figure 5**). The equitability index compares the actual diversity of a sample with the maximal possible diversity: the situation where all species are equally represented (Monte and Ghelardi, 1964). We found that the equitability decreased slightly with increasing sampling depth from 1000 to 50,000 reflecting the fact that previously unnoticed taxa were identified with increases in sampling depth.

COMPARING MICROBIOMES BY BETA DIVERSITY

Beta (β) diversity considers the difference in bacterial community composition for different environments (Whittaker, 1972; Tuomisto, 2010). To illustrate some ideas and techniques related to beta diversity, we sequenced a set of 16S libraries that constitute three well-defined clusters of samples: three stool samples from mice fed with Chow, high fat or low fat diet; the three mock libraries described in **Figure 1**; and six ileum samples from two patients affected by Crohn's disease (CD). Users should be aware that clustering of samples that are highly disimilar would be more challenging than the illustrative set of data presented here, and will likely form less well-defined clusters. The analyses shown here are equally applicable to shotgun metagenomics data. Before any comparison can be made, the read counts (reads mapped to each taxon) must be normalized (Dillies et al., 2013; Paulson et al., 2013). In **Figure 4A**, we illustrate two popular normalization procedures: the total sum and upper quartile normalization. Respectively, for each sample, the normalization factor is the sum of counts of all bacterial taxa detected or the upper quartile value for each sample. In general, normalization procedures attempt to minimize the technical variability between samples, but also accounts for sample-specific dispersion (Dillies et al., 2013). Despite numerous research endeavors in this area, normalization remains a topic under intense debate, without a consensus on which normalization procedure is the most robust one (Paulson et al., 2013).

One commonly used method to detect discrete patterns of bacterial abundance in a group of samples is hierarchical clustering (Rokach and Maimon, 2005). Samples with similar bacterial profiles are recursively grouped together in branches of a dendrogram. **Figure 4A** presents the results of a hierarchical clustering using the complete linkage method (Rokach and Maimon, 2005). As expected, mice, experimental bacterial populations (mock), and human samples formed three discrete clusters (communities). Within the human samples, using total sum normalization, samples were clustered according to patient, and inside each patient ileal resections were separated from biopsies taken 6 months after surgery, when both patients presented with recurrent disease. With upper quartile normalization, biopsies were separated from resected tissues. Hierarchical clustering is a useful tool for visualizing co-abundance patterns, but in the absence of additional statistical tests, caution should be exercised as visual patterns can be misleading (Caporaso et al., 2010).

There are two main approaches for quantifying β-diversity: those that take into account the evolutionary differences between communities, formally known as phylogenetic β-diversity (Lozupone and Knight, 2005, 2008; Leprieur et al., 2012; Lozupone et al., 2013; Wang et al., 2013), and those that do not, formally known as taxon-based or non-phylogenetic methods (Kuczynski et al., 2010). With phylogenetic methods, differences in abundances that involve closely related species are given lower weights, on the assumption that closely related species have similar genetic capabilities. One example is UniFrac (unique fraction), which has been reported to correlate well with the biological properties of samples (Navas-Molina et al., 2013) and measures the amount of "unique evolution" of a community in comparison to others (Lozupone and Knight, 2005; Lozupone et al., 2006). Phylogenetic metrics are reliant on the quality of the constructed tree for the bacterial communities within the samples, which can be problematic in some cases, contingent on the taxa and the 16S rRNA gene variable region used. One of the most popular non-phylogenetic approaches to quantify β-diversity is the Bray-Curtis dissimilarity (Bray and Curtis, 1957; Beals, 1984). It is robust to the presence of zeroes in a count table, as often is the case for microbiome data (i.e., some bacterial taxa will be present in some but not all samples). QIIME and mothur offer the possibility to readily calculate many β-diversity metrics (Schloss et al., 2009; Navas-Molina et al., 2013) and so does the R package *vegan* (Oksanen et al., 2015).

Once distances/dissimilarities between samples (i.e., differences in bacteria abundance) have been computed, they can be positioned (ordinated) in a low-dimensional space (two or three orthogonal axes) to better appreciate how closely related they are to each other. The main assumption in all ordination methods is that there are a limited number of factors that greatly influence distribution and relative abundance of species. The two most commonly used ordination techniques in bacterial ecology are non-metric multidimensional scaling (NMDS) and principal coordinate analyses (PCoA), also known as metric multidimensional scaling (Quinn and Keough, 2002;

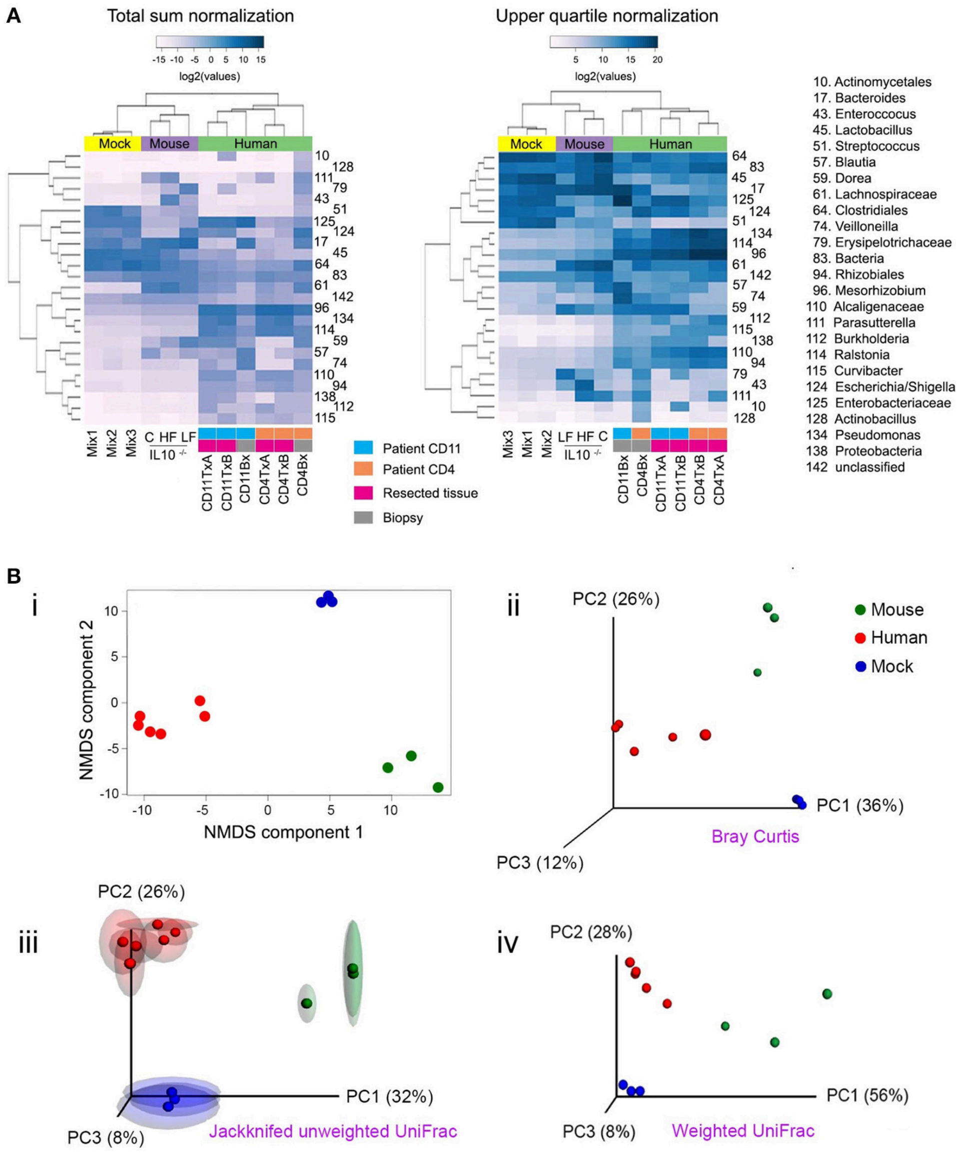

FIGURE 4 | Popular techniques for inspection and quantification of beta diversity. (A) Heatmap of normalized counts for the 50 most abundant taxa. On top of the heatmap, group of samples are color-coded. Lilac (Mouse): mutant IL-$10^{-/-}$ mice that were fed with either high fat (HF), conventional chow (C) or low fat (LF) diet. Yellow (Mock): the three mock bacteria populations described in **Figure 1**. Light green (Human): samples from two patients suffering Crohn's disease (CD4 and CD11), including resections samples from the terminal ileum at the time of surgery (run in duplicate [**A,B**]) and biopsies taken 6 months after surgery. **(B)** Non-metrical multidimensional scaling (NMDS) and Principal Coordinates Analysis (PCoA). Upper panel: Bray-Curtis dissimilarities were ordinated and plotted by either NMDS **(i)** or PCoA **(ii)**. Lower panel: Unweighted **(iii)** or weighted **(iv)** UniFrac distances were analyzed and plotted by PCoA. For unweighted distances, jackknife resampling was performed and the spheres represent the average of such process while semitransparent ellipsoids represent the variance between repeats. Mix1-3 are described in the legend for **Figure 1**; IL$10^{-/-}$C: IL10 deficient mice fed with conventional chow diet; IL$10^{-/-}$HF: as previous one, but fed with high fat diet; IL$10^{-/-}$LF: as previous one but fed with low fat diet; CD11TxA: Patient 11 affected with Crohn's disease, tissue sample from ileocolic resection, repeat **(A)**; CD11TxB: as previous one, repeat **(B)**. CD11Bx: Biopsy from patient 11 colon, 6 months after resection. CD4TxA: Patient 4 affected with Crohn's disease, tissue sample from ileocolic resection, repeat **(A)**; CD4TxB: as previous one, repeat **(B)**. CD4Bx: Biopsy from patient 4 colon, 6 months after resection.

Navas-Molina et al., 2013). The position of samples in the NMDS ordination represents the rank order of inter-sample distances, while in PCoA the ordination attempts to faithfully match their original inter-sample distances, providing results that are more readily interpretable (Ramette, 2007). In most cases, both techniques will lead to similar conclusions and it is a matter of debate which method is more appropriate (Ramette, 2007; Zur et al., 2007). For a more detailed discussion on multidimensional scaling see (Ramette, 2007; Zur et al., 2007; Buttigieg and Ramette, 2014). In **Figure 4B**, we illustrate both NMDS and PCoA analyses. In the upper panels, Bray-Curtis dissimilarities were calculated and are presented by (i) NMDS or (ii) PCoA. In the lower panels, we present UniFrac distances and PCoA ordination, either (iii) unweighted or (iv) weighted. Unweighted UniFrac considers presence/absence of OTUs and therefore emphasizes rare species, while weighted also considers the abundance of OTUs. The selection of each metric will depend on the hypothesis being evaluated as some phenotypes are more strongly influenced by relative abundance of taxa rather than presence or absence of specific taxa (Navas-Molina et al., 2013). As shown in **Figure 4B(iii)**, it is possible to evaluate the stability of the PCoA plot using a resampling procedure known as jackknifing. For this procedure, calculations are reiterated after omitting one observation (taxa, OTU, etc.) and then the average is represented in a PCoA plot while the variance is depicted as confidence ellipsoids (Efron and Stein, 1981; Navas-Molina et al., 2013).

PROFILING THE METABOLIC CAPACITY OF THE MICROBIOME

Determining the functional attributes of the microbiome is essential for understanding their role on host metabolism and disease (Joice et al., 2014). The metabolic capacity of the microbiome can be inferred or cataloged from 16S and shotgun metagenomics libraries, respectively. Gene marker approaches like 16S rely on the correlation between phylogenetic trees and clusters of genes shared between taxa (Langille et al., 2013). Shotgun metagenomics, on the other hand, provides a direct assessment of the functional attributes of the microbiome (Riesenfeld et al., 2004; Knight et al., 2012), although the results are dependent on sequencing depth.

The software PICRUSt (Langille et al., 2013) can be used to infer metabolic capacity of the microbiome contained in 16S libraries. PICRUSt functional inference is implemented in two steps. First, a reference phylogenetic tree is constructed from the Greengenes database (DeSantis et al., 2006) and gene contents are assigned to nodes in such tree if sequenced genomes are available, or otherwise predicted using ancestral state reconstruction algorithms (Langille et al., 2013). Representative sequences from OTUs derived from experimental data and associated with Greengenes identifiers are normalized by 16S rRNA gene copy number and then mapped to the corresponding Greengenes identifiers in the reference tree. The final result is an annotated table of gene counts per sample that can be linked to the Kyoto encyclopedia of genes and genomes (KEGG) orthology (KO) accession numbers (Kanehisa et al., 2004) or to any other orthologous protein family catalog. Similarly, several robust approaches have been developed to determine the functional attributes in shotgun metagenomics data, including MG-RAST (Meyer et al., 2008), MEGAN (Mitra et al., 2011), IMG/M (Markowitz et al., 2008), HUMAnN (Abubucker et al., 2012), and the R package ShotgunFunctionalizeR (Kristiansson et al., 2009). Using software like MEGAN5, each sequence can be directly mapped to KO representative sequences and the sum of KO counts that belongs to the same pathway can be computed. Alternatively, the SEED hierarchy (Overbeek et al., 2005) can be used to map reads to functional roles which can be organized into subsystems (Mitra et al., 2011). Thus, when normalized, results from PICRUSt and MEGAN5 are comparable. Recently, a new approach dubbed ShortBRED (Kaminski et al., 2015) was developed, which is both highly accurate and computer efficient. Essentially, it compiles a *de novo* database of marker peptides derived from reference databases and sequenced data, and then quantifies peptides abundance against such newly generated database.

We derived functional profiles from 16S or shotgun libraries with PICRUSt or MEGAN5, respectively. For this analysis, we used stool samples from three healthy individuals, the CD and the *C. difficile* samples described in **Figure 3**, and the three mice samples described in **Figure 4**. Twenty-three KEGG reference pathways were used to compare relative abundance determined from both type of libraries (**Figure 5A**). The level of concordance between results derived from 16S or from shotgun metagenomics was variable depending on the pathway under consideration. In general both methods recapitulated general patterns of abundance. For example, the metabolic profile of the CD stool sample was clearly distinct from the rest and exhibited the highest gene content related to membrane transport, signal transduction and carbohydrate metabolism and the lowest content related to amino acid metabolism, metabolism of cofactors and vitamins and translation factors, as previously reported for IBD patients (Greenblum et al., 2012; Knights et al., 2013; Kostic et al., 2014). In addition, we show two KEGG reference pathways (at the KO level), which relative abundance was similarly (glycolysis; $r = 0.88$) or distinctly (fatty acid biosynthesis; $r = 0.52$) assessed by both programs (**Figure 5B**). The Pearson correlation coefficient of abundance of KOs detected by at least one of the methods was 0.66.

Although 16S and shotgun metagenomics both allow functional profiling of the microbiome, shotgun metagenomics offers a more reliable assessment, provided that enough sequences are available and, ideally, it should be complemented with metatranscriptomics analyses (Franzosa et al., 2015).

CONCLUDING REMARKS AND PERSPECTIVE

The choice of shotgun or 16S approaches for microbiome analyses is usually dictated by the nature of the studies being conducted. For instance, 16S is well suited for analysis of large number of samples, i.e., multiple patients, longitudinal studies, etc. but offers limited taxonomical and functional resolution. Moreover, it should be pointed out that using primers for different regions of the 16S rRNA gene may lead to discordant results due not only to the distinct binding affinities for the corresponding flanking conserved regions, but also due to the resolution of each variable region across taxa (Soergel et al., 2012). Shotgun metagenomics on the other hand is usually more expensive but offers increased resolution, enabling a more specific taxonomic and functional classification of sequences as well as the discovery of new bacterial genes and genomes (Franzosa et al., 2015), which usually requires assembly of individual reads into contigs. Importantly, shotgun metagenomics allows the simultaneous study of archaea, viruses, virophages, and eukaryotes (Norman et al., 2014,

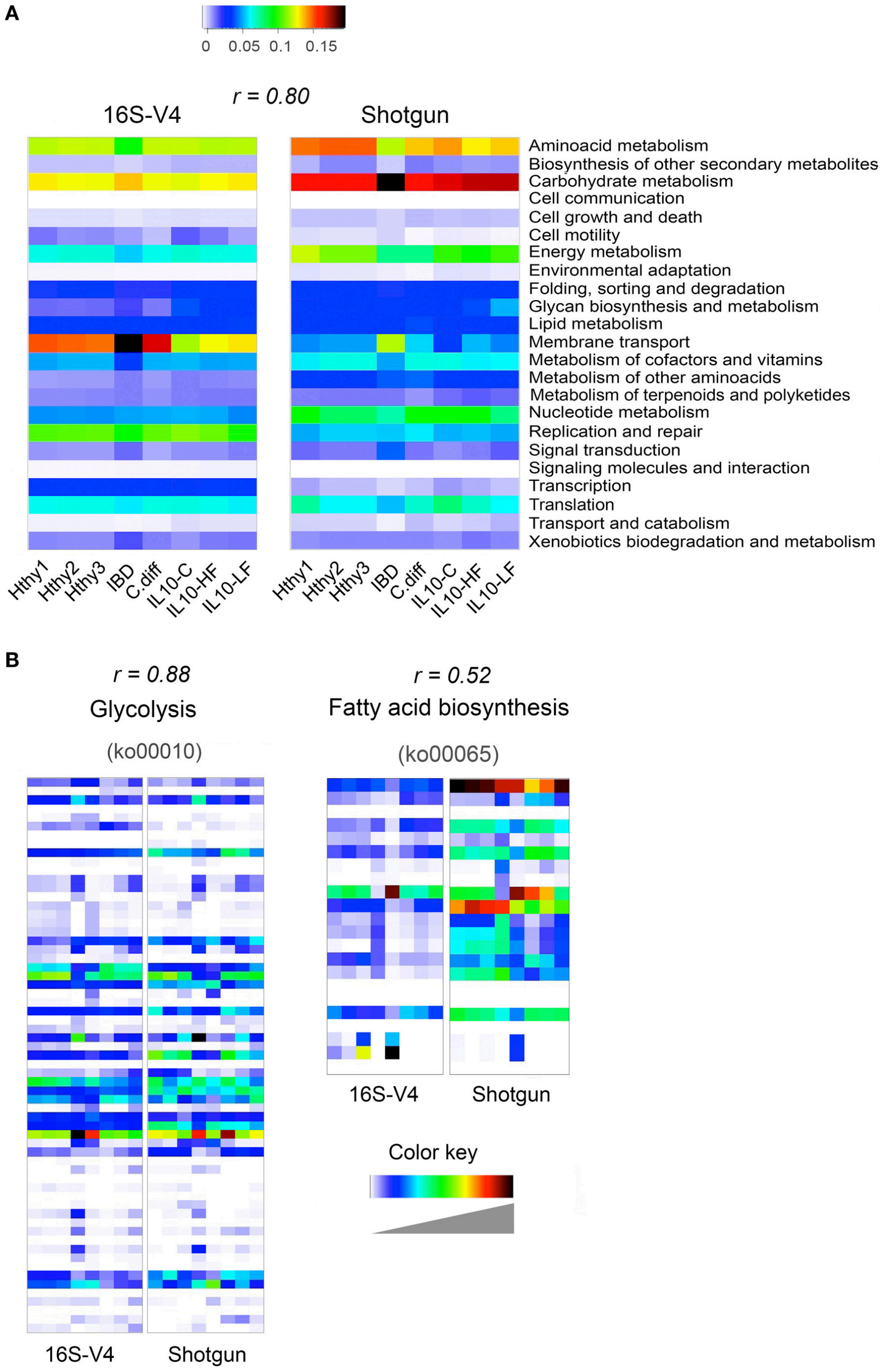

FIGURE 5 | Inference of gut bacterial microbiome functional content from 16S or shotgun metagenomics libraries. Samples from three healthy individuals (Hthy1-3), the CD and the *C. diff* samples described in **Figure 3**, and the three mice samples described in **Figure 4** were used here to illustrate metabolic inference of the gut bacteria microbiome from 16S or shotgun metagenomic libraries. High quality sequences were procured as described in **Figure 1**. **(A)** Twenty-three KEGG reference pathways known to be present in bacteria are depicted for both types of libraries. **(B)** Two KEGG pathways are illustrated at the gene (KEGG orthology, KO, groups) level. On top of each heatmap pair, the Pearson correlation coefficient for relative abundance of KOs derived with each method is presented. Inference of the functional content of the 16S metagenome was performed with PICRUSt, while gene content of shotgun metagenomic libraries was determined with MEGAN5. PICRUSt outputs results in number of bacteria cells that encode a gene (KO) while MEGAN5 outputs counts of sequences that mapped to a KO representative sequence. To make results from both methods comparable, counts were normalized by total sum. In both cases, the results represent the abundance of each KO as a fraction of the abundance of all detected KOs in each library. In order to achieve full representation of all values included in each normalized count table, colors in each heatmap were stretched between the minimum and maximum values. Therefore, the intensity (value) of each cell is not comparable between methods (16S of shotgun). Instead the Pearson correlation coefficient is shown as an estimator of the concordance of results provided by both approaches.

2015). Although several significant efforts to unravel bacterial strains have already been published (Qin et al., 2010; Qichao et al., 2014; Zhu et al., 2015), bacterial strains identification is an issue that remains unsatisfactory with current approaches. This is not only important from an aetiological perspective but also for the study of bacterial populations dynamics in general (Franzosa et al., 2015). Shotgun metagenomics offers a greater potential for identification of strains. Reportedly, the software MetaPhlAn2 has the ability to resolve different strains from the same species when reference genomes are available (https://bitbucket.org/biobakery/metaphlan2), and other software for shotgun data will likely perform well as more comprehensive databases are generated. Shotgun single-cell sequencing efforts also hold promise for bacterial strains deconvolution (Rinke et al., 2013).

In the view of experts in the field, metagenomics should be complemented with metatranscriptomics, proteomics, metabolomics and metadata, like clinical and dietary information, to derive mechanistic models that explain the structure and function of the microbiome (Brown et al., 2013; Morgan and Huttenhower, 2014; Franzosa et al., 2015; Waldor et al., 2015). Data integration will require sophisticated statistical techniques like ordination methods, hierarchical regression analyses, network analysis, and machine-learning approaches, among others (Abubucker et al., 2012; Segata et al., 2013; Franzosa et al., 2014; Joice et al., 2014; Morgan and Huttenhower, 2014). It is hoped that this primer will provide clinicians and researchers with a basic understanding of the main bioinformatics approaches for microbiome analyses with a view of advancing future investigations.

AUTHOR CONTRIBUTIONS

JJ, JP, NH, AM, KM, GW designed the study. NH, SO, TM performed experiments. JJ, JP, WW performed bioinformatics analyses. TP, DK contributed clinical samples. JJ, AM, KM, and GW wrote the manuscript. All authors read and approved the manuscript.

FUNDING

This work was supported by funding from Alberta Innovates Technology Futures-Innovates Centers of Research Excellence (AITF-iCORE) to GW and Canadian Institutes for Health Research and Alberta Health Services to KM.

ACKNOWLEDGMENTS

We would like to thank Matt Emberg and Tracy Jordan for technical assistance. We also thank Drs. Jens Walter and Inés Martínez (University of Alberta) for critically reading a previous version of this manuscript. We apologize to the authors whose work could not be included in this review due to space limitations.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2016.00459

Supplemental Figure 1 | Average quality scores of mock libraries presented in Figure 1 show that end1 (blue line) is higher than the corresponding quality in end2 (red line) especially at the 3′ end of each end.

Supplemental Figure 2 | Comparison of taxonomic classifications using different combinations of end1 and end2 for 16S libraries. The name of the method is at the bottom of each bar and the Pearson correlation coefficient between the expected (Input) and obtained relative abundance of taxa is on top of each bar.

Supplemental Figure 3 | Taxonomical classification of reads from library Mock1 described in Figure 1. Sequences were aligned either with BLAST (upper panel) or LAST (lower panel) and taxonomically classified using MEGAN5. The Krona plot depicts different bacteria taxonomic levels in concentric circles, from subspecies in the outermost circle to the bacteria kingdom in the innermost circle.

Supplemental Figure 4 | Taxonomical classification of sequences derived from the Kefir shotgun metagenomics library described in Figure 3. Taxonomical classification was done by alignment of sequences to the NCBI nt/nr and human microbiome project bacteria databases and then classified using MEGAN5.

Supplemental Figure 5 | Shannon's diversity index was used to describe species diversity in each bacterial community (the so-called α-diversity). It takes into account the number of species and their evenness, and is calculated as a weighted sum of the proportion (*p*) that each species (*i*) constitutes of the total number of species (*S*) in the bacterial community ($H = -\sum_{i=1}^{S} pilnpi$). The higher the number of species and number of individuals inside each species, the higher the Shannon's diversity index will be. Shannon's equitability or evenness (E_H) index compares the actual diversity with the maximal possible diversity (the situation when all species are equally abundant), and is calculated as ($E_H = H/H_{max} = H/lnS$). A bacterial community in which all species are equally represented will have an equitability of 1. The average value of both indices and the corresponding standard deviation were calculated from 20 simulations at depths of 1000 (blue boxes) and 50,000 (magenta boxes) reads. Artificial bacterial mixes were excluded. Kefir, grains of Kefir; CD, sample from a patient affected by Crohn's disease; Cdiff, sample from a patient affected by *Clostridium difficile*; Hthy1, sample from a healthy individual.

REFERENCES

Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., et al. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* 8:e1002358. doi: 10.1371/journal.pcbi.1002358

Aho, A., Hopcroft, J., and Ullman, J. (1973). "On finding lowest common ancestors in trees," in *Proc. 5th ACM Symp. Theory of Computing (STOC)*, (New York, NY: ACM), 253–265.

Antharam, V. C., Li, E. C., Ishmael, A., Sharma, A., Mai, V., Rand, K. H., et al. (2013). Intestinal dysbiosis and depletion of butyrogenic bacteria in Clostridium difficile infection and nosocomial diarrhea. *J. Clin. Microbiol.* 51, 2884–2892. doi: 10.1128/JCM.00845-13

Aronesty, E. (2011). *Command-Line Tools for Processing Biological Sequencing Data ea-utils*. Expression Analysis. Durham, NC. Available online at: http://code.google.com/p/ea-utils

Arslan, N. (2014). Obesity, fatty liver disease and intestinal microbiota. *World J. Gastroenterol.* 20, 16452–16463. doi: 10.3748/wjg.v20.i44.16452

Bajaj, J. S., Betrapally, N. S., Hylemon, P. B., Heuman, D. M., Daita, K., White, M. B., et al. (2015). Salivary microbiota reflects changes in gut microbiota in cirrhosis with hepatic encephalopathy. *Hepatology* 62, 1260–1271. doi: 10.1002/hep.27819

Barlow, G. M., Yu, A., and Mathur, R. (2015). Role of the gut microbiome in obesity and diabetes mellitus. *Nutr. Clin. Pract.* 30, 787–797. doi: 10.1177/0884533615609896

Beals, E. (1984). Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data. *Adv. Ecol. Res.* 14, 1–55.

Bhattacharjee, S., and Lukiw, W. J. (2013). Alzheimer's disease and the microbiome. *Front. Cell. Neurosci.* 7:153. doi: 10.3389/fncel.2013.00153

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* 13:R122. doi: 10.1186/gb-2012-13-12-r122

Bolhuis, H., Cretoiu, M. S., and Stal, L. J. (2014). Molecular ecology of microbial mats. *FEMS Microbiol. Ecol.* 90, 335–350. doi: 10.1111/1574-6941.12408

Brady, A., and Salzberg, S. (2011). PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nat. Methods* 8:367. doi: 10.1038/nmeth0511-367

Brady, A., and Salzberg, S. L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods* 6, 673–676. doi: 10.1038/nmeth.1358

Bray, J. R., and Curtis, J. T. (1957). An ordination of upland forest communities of southern Wisconsin. *Ecol. Monogr.* 27, 325–349.

Brestoff, J. R., and Artis, D. (2013). Commensal bacteria at the interface of host metabolism and the immune system. *Nat. Immunol.* 4, 676–684. doi: 10.1038/ni.2640

Broderick, N. A. (2015). A common origin for immunity and digestion. *Front. Immunol.* 6:72. doi: 10.3389/fimmu.2015.00072

Brown, J., de Vos, W. M., DiStefano, P. S., Doré, J., Huttenhower, C., Knight, R., et al. (2013). Translating the human microbiome. *Nat. Biotechnol.* 31, 304–308. doi: 10.1038/nbt.2543

Buttigieg, P. L., and Ramette, A. (2014). A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. *FEMS Microbiol. Ecol.* 90, 543–550. doi: 10.1111/1574-6941.12437

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4516–4522. doi: 10.1073/pnas.1000080107

Chakravorty, S., Helb, D., Burday, M., Connell, N., and Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J. Microbiol. Methods* 69, 330–339. doi: 10.1016/j.mimet.2007.02.005

Chang, J. Y., Antonopoulos, D. A., Kalra, A., Tonelli, A., Khalife, W. T., Schmidt, T. M., et al. (2008). Decreased diversity of the fecal Microbiome in recurrent Clostridium difficile-associated diarrhea. *J. Infect. Dis.* 197, 435–438. doi: 10.1086/525047

Chen, W., Zhang, C. K., Cheng, Y., Zhang, S., and Zhao, H. (2013). A comparison of methods for clustering 16S rRNA sequences into OTUs. *PLoS ONE* 8:e70837. doi: 10.1371/journal.pone.0070837

C. Human Microbiome Project, R. (2012a). A framework for human microbiome research. *Nature* 486, 215–221. doi: 10.1038/nature11209

C. Human Microbiome Project, R. (2012b). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., et al. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, D633–D642. doi: 10.1093/nar/gkt1244

Darling, A. E., Jospin, G., Lowe, E., Matsen, F. A. IV, and Eisen, J. A. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *Peer J.* 2:e243. doi: 10.7717/peerj.243

Dash, S., Clarke, G., Berk, M., and Jacka, F. N. (2015). The gut microbiome and diet in psychiatry: focus on depression. *Curr. Opin. Psychiatry* 28, 1–6. doi: 10.1097/YCO.0000000000000117

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARApplied, B., and environmental. *Microbiology* 72, 5069–5072. doi: 10.1128/AEM.03006-05

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinformat.* 14, 671–683. doi: 10.1093/bib/bbs046

Dinan, T. G., Borre, Y. E., and Cryan, J. F. (2014). Genomics of schizophrenia: time to consider the gut microbiome? *Mol. Psychiatry* 19, 1252–1257. doi: 10.1038/mp.2014.93

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Efron, B., and Stein, C. (1981). The jackknife estimate of variance. *Ann. Statist.* 9, 586–596.

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., et al. (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119. doi: 10.1111/2041-210X.12114

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., and Sogin, M. L. (2014). Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* 9, 968–979. doi: 10.1038/ismej.2014.195

Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L., et al. (2013). The long-term stability of the human gut microbiota. *Science* 341:1237439. doi: 10.1126/science.1237439

Flint, H. J., Scott, K. P., Louis, P., and Duncan, S. H. (2012). The role of the gut microbiota in nutrition and health. *Nat. Rev. Gastroenterol. Hepatol.* 9, 577–589. doi: 10.1038/nrgastro.2012.156

Forster, S. C., Browne, H. P., Kumar, N., Hunt, M., Denise, H., Mitchell, A., et al. (2016). HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. *Nucleic Acids Res.* 44, D604–D609. doi: 10.1093/nar/gkv1216

Franzosa, E. A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X. C., et al. (2015). Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nat. Rev. Microbiol.* 13, 360–372. doi: 10.1038/nrmicro3451

Franzosa, E. A., Morgan, X. C., Segata, N., Waldron, L., Reyes, J., Earl, A. M., et al. (2014). Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U.S.A.* 111, E2329–E2338. doi: 10.1073/pnas.1319284111

Garrett, W. S., Gordon, J. I., and Glimcher, L. H. (2010). Homeostasis and inflammation in the intestine. *Cell* 140, 859–870. doi: 10.1016/j.cell.2010.01.023

Gevers, D., Pop, M., Schloss, P. D., and Huttenhower, C. (2012). Bioinformatics for the Human Microbiome Project. *PLoS Comput. Biol.* 8:e1002779. doi: 10.1371/journal.pcbi.1002779

Greenblum, S., Turnbaugh, P. J., and Borenstein, E. (2012). Metagenomics systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc. Natl. Acad. Sci. U.S.A.* 109, 594–599. doi: 10.1073/pnas.1116053109

Hartstra, A. V., Bouter, K. E., Bäckhed, F., and Nieuwdorp, M. (2015). Insights into the role of the microbiome in obesity and type 2 diabetes. *Diabetes Care* 38, 159–165. doi: 10.2337/dc14-0769

Heinken, A., and Thiele, I. (2015). Systems biology of host-microbe metabolomics. Wiley interdisciplinary reviews. *Syst. Biol. Med.* 7, 195–219. doi: 10.1002/wsbm.1301

Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N., and Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* 21, 1552–1560. doi: 10.1101/gr.120618.111

Huttenhower, C., Knight, R., Brown, C. T., Caporaso, J. G., Clemente, J. C., Gevers, D., et al. (2014a). Advancing the microbiome research community. *Cell* 159, 227–230. doi: 10.1016/j.cell.2014.09.022

Huttenhower, C., Kostic, A. D., and Xavier, R. J. (2014b). Inflammatory bowel disease as a model for translating the microbiome. *Immunity* 40, 843–854. doi: 10.1016/j.immuni.2014.05.013

Janda, J. M., and Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol.* 45, 2761–2764. doi: 10.1128/JCM.01228-07

Joice, R., Yasuda, K., Shafquat, A., Morgan, X. C., and Huttenhower, C. (2014). Determining microbial products and identifying molecular targets in the human microbiome. *Cell Metab.* 20, 731–741. doi: 10.1016/j.cmet.2014.10.003

Joshi, N. A., and Fass, J. N. (2011). *Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ files. [Software] Version 1.33*. Available online at: https://github.com/najoshi/sickle

Kaminski, J., Gibson, M. K., Franzosa, E. A., Segata, N., Dantas, G., and Huttenhower, C. (2015). High-specificity targeted functional profiling in microbial communities with ShortBRED. *PLoS Comput. Biol.* 11:e1004557. doi: 10.1371/journal.pcbi.1004557

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2004). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–D205. doi: 10.1093/nar/gkt1076

Kielbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21, 487–493. doi: 10.1101/gr.113985.110

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J. A., et al. (2012). Unlocking the potential of metagenomics through replicated experimental design. *Nat. Biotechnol.* 30, 513–520. doi: 10.1038/nbt.2235

Knights, D., Lassen, K. G., and Xavier, R. J. (2013). Advances in inflammatory bowel disease pathogenesis: linking host genetics and the microbiome. *Gut* 62, 1505–1510. doi: 10.1136/gutjnl-2012-303954

Kostic, A. D., Xavier, R. J., and Gevers, D. (2014). The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology* 146, 1489–1499. doi: 10.1053/j.gastro.2014.02.009

Kristiansson, E., Hugenholtz, P., and Dalevi, D. (2009). ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics* 25, 2737–2738. doi: 10.1093/bioinformatics/btp508

Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N., and Knight, R. (2010). Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Methods* 7, 813–819. doi: 10.1038/nmeth.1499

Kultima, J. R., Sunagawa, S., Li, J., Chen, W., Chen, H., Mende, D. R., et al. (2012). MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS ONE* 7:e47656. doi: 10.1371/journal.pone.0047656

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

Laurence, M., Hatzis, C., and Brash, D. E. (2014). Common contaminants in Next-Generation Sequencing that hinder discovery of low-abundance microbes. *PLoS ONE* 9:e97876. doi: 10.1371/journal.pone.0097876

Leprieur, F., Albouy, C., De Bortoli, J., Cowman, P. F., Bellwood, D. R., and Mouillot, D. (2012). Quantifying phylogenetic beta diversity: distinguishing between 'true' turnover of lineages and phylogenetic diversity gradients. *PLoS ONE* 7:e42760. doi: 10.1371/journal.pone.0042760

Levy, R., and Borenstein, E. (2014). Metagenomic systems biology and metabolic modeling of the human microbiome: from species composition to community assembly rules. *Gut Microbes* 5, 265–270. doi: 10.4161/gmic.28261

Ley, R. E., Peterson, D. A., and Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* 124, 837–848. doi: 10.1016/j.cell.2006.02.017

Liu, Z., DeSantis, T. Z., Andersen, G. L., and Knight, R. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.* 36, e120. doi: 10.1093/nar/gkn491

Lozupone, C. A., and Knight, R. (2008). Species divergence and the measurement of microbial diversity. *FEMS Microbiol. Rev.* 32, 557–578. doi: 10.1111/j.1574-6976.2008.00111.x

Lozupone, C. A., Stombaugh, J., Gonzalez, A., Ackermann, G., Wendel, D., Vázquez-Baeza, Y., et al. (2013). Meta-analyses of studies of the human microbiota. *Genome Res.* 23, 1704–1714. doi: 10.1101/gr.151803.112

Lozupone, C., Hamady, M., and Knight, R. (2006). UniFrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* 7:371. doi: 10.1186/1471-2105-7-371

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71, 8228–8235. doi: 10.1128/AEM.71.12.8228-8235.2005

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387–402. doi: 10.1146/annurev.genom.9.081307.164359

Markowitz, V. M., Ivanova, N. N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., et al. (2008). IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* 36, D534–D538. doi: 10.1093/nar/gkm869

Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17. doi: 10.14806/ej.17.1.200. Available online at: http://journal.embnet.org/index.php/embnetjournal/article/view/200

Martín, R., Miquel, S., Langella, P., and Bermudez-Humaran, L. G. (2014). The role of metagenomics in understanding the human microbiome in health and disease. *Virulence* 5, 413–423. doi: 10.4161/viru.27864

Martínez, I., Wallace, G., Zhang, C., Legge, R., Benson, A. K., Carr, T. P., et al. (2009). Diet-induced metabolic improvements in a hamster model of hypercholesterolemia are strongly linked to alterations of the gut microbiota. *Appl. Environ. Microbiol.* 75, 4175–4184. doi: 10.1128/AEM.00380-09

Mende, D. R., Sunagawa, S., Zeller, G., and Bork, P. (2013). Accurate and universal delineation of prokaryotic species. *Nat. Methods* 10, 881–884. doi: 10.1038/nmeth.2575

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386. doi: 10.1186/1471-2105-9-386

Mikheenko, A., Saveliev, V., and Gurevich, A. (2015). MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 32, 1088–1090. doi: 10.1093/bioinformatics/btv697

Mitra, S., Rupek, P., Richter, D., Urich, T., Gilbert, J. A., Meyer, F., et al. (2011). Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics* 12:S21. doi: 10.1186/1471-2105-12-S1-S21

Monte, L., and Ghelardi, R. J. (1964). A table for calculating the equitability component of species diversity. *J. Anim. Ecol.* 33, 217–225.

Morgan, X. C., and Huttenhower, C. (2014). Meta'omic analytic techniques for studying the intestinal microbiome. *Gastroenterology* 146, 1437–1448 e1. doi: 10.1053/j.gastro.2014.01.049

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 40, e155. doi: 10.1093/nar/gks678

Navas-Molina, J. A., Peralta-Sánchez, J. M., González, A., McMurdie, P. J., Vázquez-Baeza, Y., Xu, Z., et al. (2013). Advancing our understanding of the human microbiome using QIIME. *Meth. Enzymol.* 531, 371–444. doi: 10.1016/B978-0-12-407863-5.00019-8

Ni, J., Yan, Q., and Yu, Y. (2013). How much metagenomic sequencing is enough to achieve a given goal? *Sci. Rep.* 3, 1–7. doi: 10.1038/srep01968

Nicholson, J. K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., et al. (2012). Host-gut microbiota metabolic interactions. *Science* 336, 1262–1267. doi: 10.1126/science.1223813

Nielsen, B., Gürakan, G. C., and Unlu, G. (2014). Kefir: a multifaceted fermented dairy product. *Probiot. Antimicrob. Proteins* 6, 123–135. doi: 10.1007/s12602-014-9168-0

Norman, J. M., Handley, S. A., Baldridge, M. T., Droit, L., Liu, C. Y., Keller, B. C., et al. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 160, 447–460. doi: 10.1016/j.cell.2015.01.002

Norman, J. M., Handley, S. A., and Virgin, H. W. (2014). Kingdom-agnostic metagenomics and the importance of complete characterization of enteric microbial communities. *Gastroenterology* 146, 1459–1469. doi: 10.1053/j.gastro.2014.02.001

Novais, R. C., and Thorstenson, Y. R. (2011). The evolution of Pyrosequencing(R) for microbiology: from genes to genomes. *J. Microbiol. Methods* 86, 1–7. doi: 10.1016/j.mimet.2011.04.006

Oksanen, J., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R., et al., (2015). *Vegan Community Ecology Package*. R package version 2.2-1. Available online at: https://cran.r-project.org/web/packages/vegan/index.html

Ounit, R., Wanamaker, S., Close, T. J., and Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* 16:236. doi: 10.1186/s12864-015-1419-2

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 33, 5691–5702. doi: 10.1093/nar/gki866

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10, 1200–1202. doi: 10.1038/nmeth.2658

Pimentel, M., Mathur, R., and Chang, C. (2013). Gas and the microbiome. *Curr. Gastroenterol. Rep.* 15, 356. doi: 10.1007/s11894-013-0356-y

Qichao, T., Zhili, H., and Jizhong, Z. (2014). Strain/species identification in metagenomes using genome-specific markers. *Nucleic Acids Res.* 42, e67. doi: 10.1093/nar/gku138

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490, 55–60. doi: 10.1038/nature11450

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219

Quinn, G. P., and Keough, M. J. (2002). *Experimental Design and Data Analysis for Biologists.* Cambridge: Cambridge University Press.

Ramette, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiol. Ecol.* 62 142–160. doi: 10.1111/j.1574-6941.2007.00375.x

Reyes, A., Haynes, M., Hanson, N., Angly, F. E., Heath, A. C., Rohwer, F., et al. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* 466, 334–338. doi: 10.1038/nature09199

Riesenfeld, C. S., Schloss, P. D., and Handelsman, J. (2004). Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* 38, 525–552. doi: 10.1146/annurev.genet.38.072902.091216

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N., Anderson, I., Cheng, J., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431–437. doi: 10.1038/nature12352

Ritari, J., Salojärvi, J., Lahti, L., and de Vos, W. M. (2015). Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genomics* 16:1056. doi: 10.1186/s12864-015-2265-y

Rokach, L., and Maimon, O. (2005). *Clustering Methods. Data Mining and Knowledge Discovery Handbook.* New York, NY: Springer.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12:87. doi: 10.1186/s12915-014-0087-z

Schaeffer, L., Pimentel, H., Bray, N., Melsted, P., and Pachter, L. (2015). Pseudoalignment for metagenomic read assignment. arXiv 1510.07371.

Schaubeck, M., Clavel, T., Calasan, J., Lagkouvardos, I., Haange, S. B., Jehmlich, N., et al. (2015). Dysbiotic gut microbiota causes transmissible Crohn's disease-like ileitis independent of failure in antimicrobial defence. *Gut* 65, 225–237. doi: 10.1136/gutjnl-2015-309333

Schloss, P. D., and Handelsman, J. (2008). A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics* 9:34. doi: 10.1186/1471-2105-9-34

Schloss, P. D., and Westcott, S. L. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl. Environ. Microbiol.* 10, 3219–3226. doi: 10.1128/AEM.02810-10

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Segata, N., Boernigen, D., Tickle, T. L., Morgan, X. C., Garrett, W. S., and Huttenhower, C. (2013). Computational meta'omics for microbial community studies. *Mol. Syst. Biol.* 9, 666. doi: 10.1038/msb.2013.22

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 9, 811–814. doi: 10.1038/nmeth.2066

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Techn. J.* 27, 379–423.

Soergel, D. A., Dey, N., Knight, R., and Brenner, S. E. (2012). Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME J.* 6, 1440–1444. doi: 10.1038/ismej.2011.208

Stackebrandt, E., and Ebers, J. (2006). Molecular taxonomic parameters: tarnished gold standards. *Microbiol. Today* 33, 152–155. doi: 10.1038/msb.2013.22

Strong, M. J., Xu, G., Morici, L., Splinter Bon-Durant, S., Baddoo, M., Lin, Z., et al. (2014). Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. *PLoS Pathog.* 10:e1004437. doi: 10.1371/journal.ppat.1004437

Sun, Y., Cai, Y., Huse, S. M., Knight, R., Farmerie, W. G., Wang, X., et al. (2012). A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief. Bioinformatics* 13, 107–121. doi: 10.1093/bib/bbr009

Tikhonov, M., Leach, R. W., and Wingreen, N. (2015). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J.* 9, 68–80. doi: 10.1038/ismej.2014.117

Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaya, I., Ondov, B., et al. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.* 14, R2. doi: 10.1186/gb-2013-14-1-r2

Tuomisto, H. (2010). A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography* 33, 2–22. doi: 10.1111/j.1600-0587.2009.05880.x

Turnbaugh, P. J., and Gordon, J. I. (2009). The core gut microbiome, energy balance and obesity. *J. Physiol.* 587, 4153–4158. doi: 10.1113/jphysiol.2009.174136

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457, 480–484. doi: 10.1038/nature07540

Vetrovský, T., and Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS ONE* 8:e57923. doi: 10.1371/journal.pone.0057923

Vincent, C., Stephens, D. A., Loo, V. G., Edens, T. J., Behr, M. A., Dewar, K., et al. (2013). Reductions in intestinal Clostridiales precede the development of nosocomial Clostridium difficile infection. *Microbiome* 1, 18. doi: 10.1186/2049-2618-1-18

Vital, M., Howe, A. C., and Tiedje, J. M. (2014). Revealing the bacterial butyrate synthesis pathways by analyzing (meta)genomic data. *MBio* 5, e00889. doi: 10.1128/mBio.00889-14

Waldor, M. K., Tyson, G., Borenstein, E., Ochman, H., Moeller, A., Finlay, B. B., et al. (2015). Where next for microbiome research? *PLoS Biol.* 13:e1002050. doi: 10.1371/journal.pbio.1002050

Wang, J., Shen, J., Wu, Y., Tu, C., Soininen, J., Stegen, J. C., et al. (2013). Phylogenetic beta diversity in bacterial assemblages across ecosystems: deterministic versus stochastic processes. *ISME J.* 7, 1310–1321. doi: 10.1038/ismej.2013.30

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

Wang, W., Jovel, J., Halloran, B., Wine, E., Patterson, J., Ford, G., et al. (2015). Metagenomic analysis of microbiome in colon tissue from subjects with inflammatory bowel diseases reveals interplay of viruses and bacteria. *Inflamm. Bowel Dis.* 21, 1419–1427. doi: 10.1097/mib.0000000000000344

Weiss, S., Amir, A., Hyde, E. R., Metcalf, J. L., Song, S. J., and Knight, R. (2014). Tracking down the sources of experimental contamination in microbiome studies. *Genome Biol.* 15, 564. doi: 10.1186/s13059-014-0564-2

Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon* 21, 213–251.

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. doi: 10.1186/gb-2014-15-3-r46

Wu, M., and Eisen, J. A. (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* 9:R151. doi: 10.1186/gb-2008-9-10-r151

Yen, S., McDonald, J. A., Schroeter, K., Oliphant, K., Sokolenko, S., Blondeel, E. J., et al. (2015). Metabolomic analysis of human fecal microbiota: a comparison of feces-derived communities and defined mixed communities. *J. Proteome Res.* 14, 1472–1482. doi: 10.1021/pr5011247

Yoon, S. S., Kim, E. K., and Lee, W. J. (2015). Functional genomic and metagenomic approaches to understanding gut microbiota-animal mutualism. *Curr. Opin. Microbiol.* 24C, 38–46. doi: 10.1016/j.mib.2015.01.007

Zhu, A., Sunagawa, S., Mende, D. R., and Bork, P. (2015). Inter-individual differences in the gene content of human gut bacterial species. *Genome Biol.* 16, 82. doi: 10.1186/s13059-015-0646-9

Zur, E. F., Ieno, E. N., and Smith, G. M. (2007). *Analyzing Ecological Data*. New York, NY: Springer.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

GLOSSARY

α-**diversity**: number of species (richness) at one site. Some formulations also consider the proportion that each species represents (evenness), in which case a high diversity implies a large number of species with similar abundances.

β-**diversity**: differences in species composition between sites. Some formulations incorporate phylogenetic information, assigning lower weights to differences in abundances that involve closely related species, on the assumption that closely related species have similar genetic capabilities.

Bray Curtis dissimilarity: a non-phylogenetic measurement of β-diversity based only on the species present in both sites.

Greengenes: a database of 16S rRNA gene sequences at the Lawrence Berkeley National Laboratory.

Hierarchical clustering: a method to detect patterns of bacterial abundance by recursively grouping samples with similar bacterial profiles into branches of a dendrogram.

Lowest common ancestor (LCA): refers to the common node of two descendants in a phylogenetic tree. With respect to taxonomic classifications, assignments are made at the lowest non-ambiguous level.

Non-metric multidimensional scaling (NMDS): an ordination method where the positions on the low-dimensional plot represent the rank orders of the inter-sample distances.

Operational taxonomy unit (OTU): a group of sequences clustered together based purely on similarity and an arbitrary threshold. OTUs may or may not be equivalent to taxonomical entities (species, genera, etc.).

Ordination: statistical techniques to transform multi-dimensional datasets into easier-to-visualize two or three-dimensional representations, such that similar datasets are placed close to each other and dissimilar datasets are placed far from each other.

Principal coordinate analyses (PCoA): an ordination method where the relative positions on the low-dimensional plot attempt to faithfully match the original inter-sample distances.

Read counts normalization: a linear scale factor correction that facilitates dataset comparisons by minimizing technical sources of variability and sample-specific data dispersion patterns.

UniFrac: a measurement of β-diversity that incorporates the phylogenetic distances between species. Both weighted (quantitative) and unweighted (qualitative) variants are used, where the former accounts for abundance, while the latter only considers presence vs absence.

Unique clad-specific marker genes: genes that are universally found in their taxonomic clade and yet are absent outside it, as scored by BLAST. Typically about 5% of bacterial genes will qualify