Published in final edited form as:
Clin Gastroenterol Hepatol. 2019 January ; 17(2): 218–230. doi:10.1016/j.cgh.2018.09.017.

Microbiome 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians

Celeste Allaband[*], **Daniel McDonald**[‡], **Yoshiki Vázquez-Baeza**[‡], **Jeremiah J. Minich**[§], **Anupriya Tripathi**[‖], **David A. Brenner**[¶], **Rohit Loomba**[¶,#], **Larry Smarr**[#,**,‡‡], **William J. Sandborn**[#,§§], **Bernd Schnabl**[¶,#,§§], **Pieter Dorrestein**[‡,#,‖‖], **Amir Zarrinpar**[¶,#,§§], and **Rob Knight**[‡,#,**,¶¶]

[*] Biomedical Sciences Graduate Program, University of California San Diego, La Jolla, California

[‡] Department of Pediatrics, University of California San Diego, La Jolla, California

[§] Scripps Institution of Oceanography, University of California San Diego, La Jolla, California

[‖] Division of Biological Sciences, University of California San Diego, La Jolla, California

[¶] Department of Medicine, University of California San Diego, La Jolla, California

[#] Center for Microbiome Innovation, University of California San Diego, La Jolla, California

[**] Department of Computer Science and Engineering, University of California San Diego, La Jolla, California

[‡‡] California Institute of Telecommunications and Information Technology, University of California San Diego, La Jolla, California

[‖‖] Skaggs School of Pharmacy, University of California San Diego, La Jolla, California

[¶¶] Department of Bioengineering, University of California San Diego, La Jolla, California

[§§] Division of Gastroenterology, Veterans Administration San Diego Health System, La Jolla, California

Abstract

Advances in technical capabilities for reading complex human microbiomes are leading to an explosion of microbiome research, leading in turn to intense interest among clinicians in applying these techniques to their patients. In this review, we discuss the content of the human microbiome, including intersubject and intrasubject variability, considerations of study design including important confounding factors, and different methods in the laboratory and on the computer to read the microbiome and its resulting gene products and metabolites. We highlight several common pitfalls for clinicians, including the expectation that an individual's microbiome will be stable, that diet can induce rapid changes that are large compared with the differences among subjects, that everyone has essentially the same core stool microbiome, and that different laboratory and computational methods will yield essentially the same results. We also highlight the

Reprint requests, Address requests for reprints to: Rob Knight, PhD, Department of Pediatrics, University of California San Diego, 9500 Gilman Drive, MC 0763, La Jolla, California 92093-0763. robknight@ucsd.edu; fax: (858) 246-1184.

Conflicts of interest The authors disclose no conflicts.

current limitations and future promise of these techniques, with the expectation that an understanding of these considerations will help accelerate the path toward routine clinical application of these techniques developed in research settings.

Keywords

Gut Microbiome; Clinician; Study Design; Prognosis; Diagnosis

Interest in the microbiome is at an all-time high, with the microbiome connected to an increasing range of diseases of interest to gastroenterologists and hepatologists. For example, obesity,[1–4] inflammatory bowel disease,[5–7] alcoholic and nonalcoholic fatty liver disease,[8–10] and hepatocellular carcinoma[11–14] all have been linked to the microbiome in human beings, and changes in the microbiome have been shown to induce or modify these diseases in animal models. However, moving this linkage to a clinically relevant diagnostic is still in the research phase.

We have seen enormous progress in the past decade in using genomic sequencing coupled with computational pipelines to decipher the human gut microbiome.[15] These tools are necessary because of the incredible information density of the microbiome. Each teaspoon of stool contains in its bacterial DNA alone the amount of data that it would take 100,000 of today's highest-capacity thumb drives to store. (This number was reached using the following calculations: 1 g stool contains 100 billion microbes [27541692; https://doi.org/10.1371/journal.pbio.1002533]. We then assume 5 million bases per microbe [~1 million bytes], which then yields 10E11 10E6 = 10E17 or 100,000 terabytes. The current highestcapacity thumb drive size is 1 terabyte, so at approximately 12 g each, this data would weigh 1200 kg [1.3 tons], or approximately the weight of a young giraffe.) This information also is dynamic because the microbiome profile changes with diet and medical interventions. These problems create challenges for clinicians in deciding whether it will be medically informative to ask a patient to collect stool or for the physician to obtain colonoscopy biopsy specimens and send them off for sequencing. Interpreting and discussing the results with patients can be challenging, especially with a lack of standard parameters and reference data for comparison.

In this review, we cover what the microbiome is, how it can be collected, what molecular methods can be used to analyze it, how the data can be interpreted, and what some of the limitations are in combining conclusions from different studies. Our goal is to highlight which areas are solid, which areas are emerging, and where the greatest potential is for future work to provide actionable information that benefits patients.

What Is the Microbiome?

The human gut is home to a variety of microbes, including bacteria, archaea (single-celled organisms without nuclei that are related more closely to eukaryotes than to bacteria), fungi (mostly yeasts), microbial eukaryotes (usually *Blastocystis* in the United States, but a variety of pathogenic and nonpathogenic taxa in developing countries), and viruses/phages. This collection of microbes is called the *microbiota*; their genes are called the *microbiome*.[16]

However, the term *microbiome* has come into popular use to refer to the microbes themselves. Whether the microbiome includes the virome (the repertoire of viral genes) is open to debate. Because of the technical ease and widespread utility of approaches that just read the bacteria (see later), many people assume that the microbiome refers only to the bacteria, but this is not correct. Rather, if a difference is shown in the bacterial compartment of the microbiome between cases and controls, it is necessarily true that the microbiome is different; however, if no difference in the bacteria is found, there still might be a difference in other kinds of microbes (eg, yeast or viruses).

Until recently, a frequently repeated slogan was that the human microbiome contained 10 times as many cells as the human body. This figure was based on a rough calculation 40 years ago,[17] and the correct claim was that the true figure was somewhere between 1:10 and 10:1, but could be as much as 10:1. Since then, the errors on the estimates of the number of human cells and microbial cells have narrowed considerably, with the true figure being much closer to 1:1, with the balance slightly in favor of the microbes.[18] Therefore, it is fascinating to consider that one can tip the balance in an individual’s body from having more microbial cells to having more human cells by simply administering the bowel preparation for a colonoscopy. In this article, we focus on the gut microbiome, although microbiomes in other parts of the body (eg, the skin, mouth, and vagina) also are important for health and in numerous diseases.

In most healthy human beings, the gut microbiome is dominated in cellular relative abundance by bacteria, specifically members of the phyla Bacteroidetes and Firmicutes, with only small amounts of nonbacterial microbes. It is important to recognize that among healthy people, their percentage of each of these 2 dominant phyla can vary from 10% to 90%, even though the combined percentage tends to be approximately 95%.[19] However, some individuals, particularly in the disease state, can have large percentages of other bacterial phyla, such as Proteobacteria (which contains *Escherichia coli*), Verrucomicrobia, Actinobacteria, or Fusobacteria.

The earliest culture-independent projects showed that different people can differ greatly from one another in terms of their microbiomes,[1,19–21] and the diversity spanned by human stool is comparable with the diversity spanned by completely different kinds of environments in the Earth Microbiome Project (Figure 1A). In fact, some bacterial species that were as abundant as 5% of the total in 1 individual, turned out to be no more abundant than 0.01% in another individual, even in a small cohort.[21] We have seen the same breadth of composition differences in the American Gut Project data (Figure 1B). Therefore, there is no standard microbiome ecology that all healthy people share. However, because of this high variability among individuals, extreme caution must be taken in interpreting results from fewer than hundreds of people, and the reference range approach that has worked for blood tests will not work for the microbiome[22,23] (Figure 1).[24,25]

What Is the Best Way to Collect a Sample for Microbiome Analysis?

The first topic a clinician faces is the following: what is the optimal protocol for collecting a microbiome sample for analysis? There is still an ongoing debate on the best way to collect and store a sample for analysis of the microbiome. In short, there is no perfect method because the choice will depend on feasibility, cost, patient acceptance, and which methods will be used to read the microbiome downstream.

The first important question is what to sample. Stool is by far the most accessible material, and can be collected as often as your subjects produce stool, enabling longitudinal studies (eg, of daily samples) that would not be feasible with biopsy specimens. For studying gastrointestinal and liver diseases, assessing the gut microbiome using stool provides a unique opportunity to study pathophysiology and disease states in both cross-sectional and longitudinal study designs. However, stool does not capture all the microbes in the gut,[20,26] and in particular mucosally adherent microbes and microbes in the small intestine, particularly the ileum, can be missed. In addition, stool often is quite distant from the gastrointestinal region of the pathology being investigated, and has been stored in the rectum, where there is active dehydration and where fermentation selects for bacteria that are not found commonly in other parts of the lumen. This implies that it is difficult to use the stool microbiome to understand the pathophysiology of a disease because it likely fails to reflect the microbiome of the region of pathology, and it is imperative to choose a sample collection method that is inherently consistent with the scientific or clinical question being asked.

Culturomics approaches,[27] in which large numbers of cells are isolated and cultured, show that metagenomics approaches miss many rare bacteria that are not well represented in the reference databases or that are below the filtering thresholds used to eliminate noise (see later). They also suggest that even the most aggressive homogenizing procedure to break bacterial cell walls still may miss important organisms. On the other hand, approximately 85% of microbes in the human gut are anaerobic and therefore do not culture in an open Petri dish, although they can be grown in research laboratory anaerobic chambers. However, despite advances in culturing methods,[28,29] what can be cultured still is biased, especially because any given culture condition will allow some bacteria to grow much faster than others.

Despite these limitations, the gold standard protocol for stool sampling is to collect the whole stool, homogenize it immediately (eg, with a blender or a tissue homogenizer), then flash freeze the homogenate in liquid nitrogen or in dry ice/ethanol slurry, with an aliquot preserved in 20% glycerol in Lysogeny Broth for culturing. Nucleic acid protectors such as RNAlater (Thermo Fisher Scientific, San Diego, CA), although popular, have had mixed success in different laboratories, and render the sample unsuitable for metabolomics, so should be used on a separate aliquot. However, this protocol is expensive and often impractical, especially given the limitations inherent in subjects’ ability to produce stool on demand. Although stool is not homogeneous, in general the differences between whole stool and a small sample of stool are small compared with the differences between individuals. Although stool consistency is correlated with microbiome changes,[30] stool consistency does not interfere with DNA extraction in people with chronic gastrointestinal conditions such as irritable bowel syndrome, inflammatory bowel disease (IBD), and constipation.

For DNA analyses, several studies have shown that Flinders Technology Associate and fecal occult blood test cards are stable at room temperature for at least days,[31–33] and although they induce small, systematic shifts in the resulting taxon profiles compared with flash-frozen samples, the practical ease of use of these methods is a considerable attraction. Another widely used method is dry swabs of fecal material left behind on bathroom tissue, such as those used in the American Gut Project,[25] which can be used for amplicon analysis (eg, for 16S ribosomal RNA [rRNA] gene profiling by polymerase chain reaction [PCR], see later) with appropriate filtering for overgrowth, but are problematic for shotgun metagenomics, and only cotton-based swabs, not polyester-based, can be used for metabolomics because of issues with polymers. An important practical consideration for using swabs in the mail is that polycarbonate housings are not nearly as robust to the vagaries of mail handling as polyethylene, and require padded envelopes to arrive intact. However, despite these limitations, dry swabs from bathroom tissue have yielded useful results in many studies.[25,34,35]

Going beyond the stool, many studies have shown that the mucosa and lumen differ in their microbiomes from each other at a given site in the gut,[36,37] and that the microbiome varies dramatically along the length of the gut, with the stomach and small intestine being essentially entirely distinct from the large intestine. More subtle intersample variations are therefore found within the small intestine and within the large intestine. This raises the question of where one should look for microbiome associations. However, practically speaking, obtaining biopsy samples from the small intestine is quite challenging clinically, and obtaining them from the large intestine during colonoscopy requires skill and protocols for extracting microbial contents from the biopsy specimens.

Some studies (eg, Gevers et al[6]) have shown that better classifiers for IBD can be developed using samples of luminal content collected directly from the gut rather than stool, but this has been contradicted by other studies that show high classifier accuracy for stool (see later for explanation of these terms). In an ideal world, sampling design would be driven by a hypothesis about mechanism. Most microbial biomass and therefore metabolism occurs in the luminal contents of the large intestine, so microbes that produce and release small-molecule metabolites that enter the bloodstream would be expected to be most important there. In contrast, microbes that interact directly with epithelial cells or dendritic cells would be expected to be concentrated in mucosal biopsy specimens. Microbes that produce metabolites from dietary components that are absorbed in the ileum, duodenum, or jejunum should be sought there. However, we still lack the general understanding about the distribution of microbes and metabolism along the length of the gut to draw general conclusions about where to take samples. The advent of very low biomass protocols, such as KatharoSeq (which uses a series of positive control spike-ins to define what is real and what is contamination at different stages), allows even tiny specimens to be processed.[38]

An important question is how often to sample stool, because the microbiome ecology is intrinsically dynamic. This largely comes down to what question you are trying to answer. Remarkable changes have been observed between one day and the next, especially in the times surrounding colonoscopy and surgery,[25,39] as well as during clinical situations such as IBD flares. These changes would be missed entirely with a less-frequent study design. For episodic diseases, such as IBD, it is known that patients can have large changes in the microbiome composition on time scales of weeks to months.[40] On the other hand, changes induced by diet (eg, those associated with weight loss), take place on a timescale closer to months than days in human beings[41–43] (Figure 2). Having several serial samples provides considerable insight into microbiome dynamics,[40,44] with samples of up to half a dozen providing substantially better classifiers from stool regardless of sampling interval. However, answering this question conclusively will require detailed study of many patients, which is prohibitively expensive at present and impossible to perform with anything beyond a stool sample. However, although it is difficult to obtain serial mucosal/biopsy or luminal samples from individuals because of the cost and invasiveness of the procedures, this may be the best strategy for patients who are receiving multiple, often scheduled, endoscopies as part of their routine care or in the event of exacerbations (eg, variceal screening esophagogastroduodenoscopies for patients with cirrhosis; colonoscopy for patients with IBD) (Figure 2).

On the other hand, adequately collected and optimally stored fecal samples from chronic liver disease patients, such as nonalcoholic fatty liver disease, can provide unique insights into differentiation between those with a milder form of fibrosis vs advanced fibrosis in a cross-sectional setting.[45] Furthermore, integrating the gut microbiome with the metabolome may offer deeper insights into the metabolic perturbations linking the gut microbiome with disease states.[45] Recent studies also have suggested that certain bacterially derived metabolites may be associated with shared gene effects with disease states of interest.[46] Longitudinal studies are needed to assess causality, and are discussed later in this review.

What Sort of Microbiome Data Should I Collect?

There is a bewildering diversity of microbiome-relevant molecular analyses that can be performed on biological specimens today, each with strengths and weaknesses (Figure 3). The correct type of analyses for an experiment is completely dependent on the scientific question and hypothesis. Some of the more traditional methods focus on species identification or toxin presence for pathogens,[47] while newer methods seek to describe and detect whole communities rather than individual organisms (Figure 3).[48]

For known organisms with well-characterized selective culture conditions, culturing is still the most sensitive detection method, and comparisons of colony-forming units per milliliter is the best way to obtain the absolute abundance of viable organisms. This method can be used on a variety of sample types including stool, blood, and skin. Various organisms found in the stool are susceptible to antibiotic resistance including *Clostridium difficile* and *Enterococcus* species, which also are highly infectious pathogens.[49] Culturing enables a phenotypic classification of an isolate including pathogenicity, antibiotic resistance mechanisms, and antibiotic susceptibility.[47] However, this method is best suited to reading a small number of well-known organisms that can survive in the presence of oxygen, not to characterizing the entire complex and largely anaerobic gut microbiome.

A broader view can be obtained by assay panels that target a set of known bacteria, viruses, parasites, or functional genes such as toxins or antibiotic resistance. Stool samples generally are processed through nucleic acid extraction followed by complementary DNA synthesis and subsequent amplification using mixtures of primers specific for a given range of organisms. Either genomic DNA or PCR product then is qualified and quantified across the organism panel, either through a hybridization array using a fluorescence-based measure or a melt curve analysis.[50,51] Both quantitative PCR and reverse-transcription quantitative PCR also are examples of these methods that are used to detect and quantify specific organisms.[52] Various companies (Verigene, Luminex, Riverside, CA; Biofire, Salt Lake City, UT; and Luminex) have developed Food and Drug Administration–approved platforms to detect microbial pathogens from bulk stool samples.[53] The platforms can be microfluidic chips that perform multiple processes including DNA extraction, PCR, and read-outs.[54] Through-put ranges from 1 to 24 samples, while time ranges from 1 to 5 hours.[53] The mentioned technologies target between 14 and 22 analytes, including 7 to 14 bacteria, 2 toxins, 2 to 5 viruses, and 0 to 4 parasites. The advantage to these assays is that they provide absolute abundance of each taxon per gram or milliliter of input material, and have a high dynamic range. The disadvantage is that there are many undiscovered taxa in the gut that may be important, and these will be missed in a targeted panel. However, as we understand more about the specific microbes that make the difference between clinical indications, these targeted panels will be increasingly valuable. However, one important concern is whether panels developed in one population will apply to another (see later).

Amplicon analyses, in which a specific piece of DNA is amplified by orders of magnitude using various methods including PCR, have been the workhorse of the microbiome for the past 15 years.[55] In these analyses, PCR primers that match a specific gene, usually the 16S rRNA for bacteria and archaea and the internal transcribed spacer for fungi, are used to amplify all the variants that occur between the highly conserved regions used to construct the primers. For example, bacterial 16S rRNA genes contain 9 hypervariable regions (V1–V9) that show sequence diversity and therefore often are used as a barcode-like method to differentiate many bacterial taxa, sometimes but not always at the species level. Then next-generation sequencing, typically on the Illumina (San Diego, CA) platform,[56] is used to read all the sequences, which then can be placed into a phylogenetic tree or matched to a database. There are many considerations in choosing which primers to use, and the difference between the microbiome profiles obtained with different PCR primers is much greater than the difference between the stool of different healthy individuals.[19] Consequently, the best option is to use the same PCR primers as other studies with which you would like to compare your results, or if there is no specific study in mind then using widely used primers such as the V1 to V3 or V3 to V5 primers from the Human Microbiome Project or the V4 primers from the Earth Microbiome Project (which have the advantage that they pick up archaea such as *Methanobrevibacter* and *Methanosphaera*, which are both important in the gut) is the best plan. Critically, many primers can target the same variable region, so it is important to know not just which region is being sequenced but the specific primers themselves. In general, the specific region is much more important than the length of the fragment,[57,58] and a long sequence with biased primers can provide a spectacularly incorrect result. Therefore, it is important to beware of claims about the value of longread sequencing that are not backed by extensive validation in the form of peer-reviewed reports. Many species of bacteria are identical along the full length of the 16S rRNA gene, and in principle it therefore is impossible to distinguish all bacterial species using that gene, despite claims of some vendors. In general, genus-level resolution is possible for most bacterial taxa, but species resolution is difficult.[59] Amplicon analyses in general are challenging to apply to viruses, which is mostly because there is no gene common to all viruses like there is in bacteria.

Although 16S rRNA sequencing has enabled a great deal of scientific research on microbiomes, simply knowing the genera of bacteria and its relative abundance is not as useful for clinical analysis. This is because each genus can have a wide range of strains that are genomically distinct. This is true even within a species: *E coli*, for instance, has a genome that can vary from 4 to 6 million DNA bases,[60] which group into several thousand distinct genes, some of which can be quite virulent. As a result, there are thousands of known strains of *E coli* that have been sequenced (only approximately a third of the *E coli* genome is core to all its strains) and found to be genomically distinct, with at least 1 strain considered a probiotic and another that can cause debilitating illness.

In contrast to the use of 1 gene, such as 16S rRNA, shotgun metagenomics is a method that fragments all the DNA from a sample into small pieces, sequences these fragments, then tries to puzzle these fragments together into a view of the microbiome.[61] The advantage to shotgun metagenomics is that it is very easy to explain what it does: you are trying to infer the complete list of microbial strains present in a microbiome, including the fungi and viruses that are missed by 16S rRNA amplicon analysis, and how abundant each of those strains is. However, the technical challenges are considerable: for example, analyses rely on genomes of the organisms in the gut, many of which are unknown (especially outside the bacteria). Shotgun metagenomics was traditionally orders of magnitude more expensive than amplicon analyses, but with rapid decreases in the cost of DNA sequencing and library preparation this technique is becoming much more accessible on a large scale. In addition, the amount of DNA required for shotgun metagenomics recently has decreased from micrograms to less than a nanogram, allowing it to be used on biopsy specimens. An important limitation to shotgun metagenomics is that all the DNA will be sequenced, including human DNA, which is a problem if your subjects are not consented for human DNA analysis or if your biopsy specimen is dominated by host tissue (resulting in very expensive resequencing of the human genome, with only a small trace of microbial reads; this is common in biopsy specimens, which is why 16S rRNA amplicon analysis typically is used for such specimens; “host DNA depletion” techniques, although successful in saliva[62] have not yet worked for biopsy specimens, although this is an active area of methods development). Shotgun metagenomics is rapidly displacing 16S rRNA amplicon analysis because of its expanded taxonomic range and strain-level resolution, but is subject to many of the same reproducibility issues that have not yet been as well characterized because of the increased expense of the assays.

Metatranscriptomics, in which the transcribed RNA is sequenced, and metaproteomics, which uses mass spectrometry to sort out the wide range of proteins in a sample, have tremendous promise because they read gene expression, but are still very challenging. Most bacterial transcripts only last a few minutes,[63] so the interpretation of RNA left in a stool sample is challenging. Moreover, in the few comparisons that have been performed, the correlation between gene expression in the RNA and proteins at the whole-community level has been close to zero, complicating interpretation of the expression profiles. These should be considered emerging technologies rather than ready for routine use, although techniques are rapidly improving. Studies of expression require metagenomic data from the same sample to back them so that changes in the relative expression of particular genes can be distinguished from changes in the representation of these genes in the community.[64]

Metabolomics, the study of the nonprotein small molecules including products of metabolism, is a very exciting emerging area because it relates directly to the function of the community. The most common approaches separate metabolites by gas chromatography or liquid chromatography before analysis by mass spectrometry as charged ions. There are 2 main approaches of metabolomics analysis: targeted metabolomics, in which we have a predetermined list of molecules, typically for which the reference standards are available. It is usually the most sensitive approach for detecting molecules of interest and has better quantification compared with untargeted mass spectrometry but does not allow for discovery.[65,66] Most molecules that are made by the microbiome are not commercially available or still remain to be discovered and therefore cannot be analyzed via targeted methods. On the other hand, untargeted metabolomics aims to detect as many small molecule metabolites as possible. The main challenge for untargeted metabolomics is the annotation of these metabolites. For untargeted metabolomics, tandem mass spectrometry (which weighs the ions, then breaks them into fragments, then weighs the fragments) often is used to provide annotations by matching against a reference library of known molecules. However, this fails to annotate molecules that are modified by the microbiome or host metabolism. However, fragmentation data from related spectra can be found by linking their mass spectra through a technique called molecular networking[67,68] (see later), allowing identification of new molecules that are related to known ones. An important consideration when choosing a metabolomics platform is whether the target molecules will be captured, for example, many standard untargeted liquid chromatography/mass spectrometry/mass spectrometry approaches do not pick up short-chain fatty acids such as butyrate and acetate, which are known to play important physiological roles in the gut, on the other hand gas chromatography–mass spectrometry does not pick up molecules from the host that are modified by microbes. Examples of such molecules include lithocholic acid, the oral bacteria produced fungal biofilm inhibitor mutanobactin A,[69] and the microbial molecule 4-phenyl-ethyl sulfate, which results in autism-like symptoms in rodent models.[70] The current preferred methods for stool are a combination of shotgun metagenomics and metabolomics. It is likely that metabolomics will not only be able to report on microbially modified or microbially biosynthesized molecules, but also provide a direct read of the medications as well as diet that affect the gut microbiome.

How Should I Analyze My Data?

The main question clinicians usually have is either “how do my cases and controls differ?” Or “is this sample from this patient indicative of a particular disease?” These questions can be difficult to answer with the current state of the science, especially given the many options for conducting the molecular analysis.

The wrong approach is to decide to perform a microbiome study, pick a type of sample to collect, decide which molecular assay to run, and then decide to analyze the data yourself or hand it off to a bioinformatics or biostatistics collaborator, core facility, or company. The greatest expense in many studies is data analysis, and if the study was not designed in a way that allows the data to be analyzed easily, this can take years and cost hundreds of thousands of dollars (if accurately accounted). We cover issues of study design extensively elsewhere in other recent reviews.[71–74] Briefly, it is important to consider confounding factors such as age, drugs, diet, and co-housing, issues of causality. Your patients might be sick because their microbiomes are different or their microbiomes may be different as a consequence of their medical condition or treatment. It is important to begin to appreciate that studies designed with equal numbers of samples per group with consistent time points are dramatically easier to analyze. Furthermore, a common tactic is to use the microbiome differences to infer that they underlie a pathophysiological process that was not even part of the initial intent of the study. Not only does this assume a causative relationship between the microbiome and the pathology being investigated, but also our knowledge of the relationship of the gut microbiome on host processes is often not yet sufficient to support such conclusions. Finally, for all next-generation sequencing–based methods of microbiome analysis, it is paramount to include positive and negative controls to help distinguish between signal and noise.[38]

The most important consideration with data analysis is that different methods will provide different results, even using the same raw data from the DNA sequencing instrument. This issue stems from several distinct sources. First, algorithms for assigning DNA sequences to particular genomes or classes of organisms are approximate. For example, the popular RDP classifier[59] has an accuracy of approximately 80% at the genus level using short 16S rRNA fragments. This means that approximately 20% of the assignments are wrong, which is not ideal. In shotgun metagenomics, approaches, such as Kraken[75] or Centrifuge,[76] based on k-mers (short fragments of sequences, often only a few bases long) are much more sensitive (likely to find an organism if it is present, especially at low abundance), but less specific (likely to report an organism even if it is not present) than those based on profile matches to marker genes, such as PhyloPhlAn.[77] Whether it is more dangerous to miss an organism that is present or accidentally report an organism that is absent depends on the clinical application. In any case, none of these techniques is currently suitable for clinical use. The diagnosis of pathogens still should be performed by Food and Drug Administration–approved, culture-based, PCR-based, or antibody-based assays.

In addition, most approaches rely on reference databases that are highly incomplete. Consequently, matches to a given sequence will vary depending on what sequences are actually in the reference database and the name given to the closest sequence, which results in different bacterial names given to the same DNA sequence depending on the database used. Because of this, you can get wildly different results. In the past this was an enormous problem, although cooperation among the rRNA-based taxonomy databases such as SILVA,[78] RDP,[79] and Greengenes[80] have reduced this problem and resulted in more consistency between results in recent years. However, taxonomy based on whole genomes rather than on single-marker genes is likely to prompt large-scale revision of taxonomy as we discover more about the relationships among major groups of organisms.

The major considerations in data analysis are as follows (Figure 4).[81–91] (1) How do I go from my raw DNA sequence data to a table of how many of each species (or gene/strain, for metagenomics) is observed in each sample? (2) How do I link this table to relevant clinical variables for analysis? (3) How do I perform appropriate analyses either at the level of the whole microbiome (typically, α diversity and β diversity analyses) or at the level of individual taxa or genes?

There are several features of microbiome data from a statistical standpoint such as sparsity, compositionality, and zero inflation that make standard statistical tools inappropriate for most microbiome analyses. It is therefore critical to use tools designed for these analyses, such as QIIME/Qiita,[81] the BioBakery,[92] or PhyloSeq[93] that take these considerations into account. Providing details of how to analyze microbiome data is beyond the scope of this article, but we have covered this topic recently in several other reviews that will be of interest to readers who want more details.[74]

What Are the Limits to Combining Data From Different Studies?

One frequently encountered issue is reading an exciting research report that links a particular microbe, pathway, or gene to a condition or treatment, then wanting to see if the same relationship holds true in a new cohort or a new individual patient. This apparently simple question turns out to be surprisingly difficult.

As noted earlier, a very large number of factors can affect the read-out of the microbiome, especially at finer taxonomic levels, but they are by no means limited to these levels. The same samples can yield completely different assessments of which phyla are abundant in a given specimen when using PCR-based methods, including primers that target different hypervariable regions (eg, V1–3 vs V4) or different primers that target the same region but pick up different taxa with different efficiency. It is especially true when trying to make an assessment at the species level, which current sequencing techniques are poorly suited to determine. Consequently, if you are designing a new study and want to compare it with an existing study, the safest approach is to use exactly the same methods in every detail, including sample collection, sample storage, DNA extraction, PCR or library construction, sequencing, and bioinformatics analysis. Standardized reporting such as the Genomic Standards Consortium MIxS standards[94] help immensely with this task by capturing the information in a structured way and, in the context of databases such as Qiita (https://doi.org/10.1038/s41592-018-0141-9), allow automatic retrieval of studies that used comparable methods.

The Human Microbiome Project[19] showed that even when everything else is kept exactly the same, the choice of PCR primers (V1–V3 vs V3–V5) and the choice of whether to perform shotgun metagenomics or 16S rRNA sequencing on the same samples can produce completely different results. Similarly, the Microbiome Quality Control project showed that differences in the computational pipeline, even on the same data, could lead to large differences in the inferred outcomes at levels from the species to the phylum.[95] However, one valuable outcome of the Microbiome Quality Control project was that many different laboratories could independently reproduce similar results on the same samples by following a consistent written protocol.[95] Similarly, in the American Gut Project, we found that dozens of sequencing runs over many years yielded consistent results when consistent protocols were used, and this was highlighted in the Supplementary Video 1 of that report (https://figshare.com/articles/movie_s2_mp4/5936482).[25, 96]

In general, whether and how studies can be combined depends on the subtlety of the effect and the type of analysis being performed. Different parts of the human body differ radically in their microbiomes, and neonatal microbiomes are completely different from adults. Therefore, even studies using different DNA extraction methods and sequencing techniques often will yield the same pattern in combined analysis (eg, through principal coordinates analysis).[97] In contrast, subtle differences such as those yielded by day-to-day variation within a healthy individual are much smaller, and will be obscured by even minor technical variation such as lot numbers of sequencing reagents. A general guideline is that the more technical factors differ between 2 studies, the more obvious the difference will need to be to be visible. Although the American Gut Project and other recent projects have started to construct an effect size scale for factors that affect the microbiome in large or small ways, incorporating technical variation at these scales would be an arduous and expensive undertaking. One approach that often is useful is asking whether particular taxa or gene functions are reliably increased or decreased with a given clinical state (eg, ulcerative colitis, nonalcoholic fatty liver disease) across many studies, although different methods can in principle lead to different conclusions, even with data analysis at this level.

Of particular concern to clinicians is whether data from companies offering testing, or from citizen-science projects such as the American Gut Project,[25] is comparable with studies performed in the scientific literature. The American Gut Project is part of the Earth Microbiome Project, and uses the Earth Microbiome Project protocols[24] that have been applied in literally thousands of microbiome studies, including those that are clinically relevant. Unsurprisingly, testing services that use proprietary protocols produce different results, even on the same biological specimens. In general, to understand these differences, it is necessary to have detailed information about all the protocols being used.

Another important issue is that although many associations between the microbiome and disease or between the microbiome and treatment have been found within the context of individual research studies, there are many reasons why these might not generalize to new individuals or populations. It is well known in the field of human genetics that environmental factors have a major impact on which genes are important for a given trait, and the same likely is true for the microbiome, therefore validation cohorts are essential to prove the generality of microbiome findings just as they are for human genetic findings. Some conditions, such as IBD, have very robust signatures across populations,[6,98,99] with diagnostic models trained in human beings working even on dogs[100]; in contrast, while there are typical signatures that separate lean from obese individuals within one population, these signatures do not apply across other cohorts.[23,98,101,102] This result is surprising given that obesity can be transmitted from human beings into germ-free mice by transmitting the microbiome from obese people, showing the direct effects of the microbiome.[2,103] Understanding which findings will generalize to new subjects, and which will not, remains an important outstanding challenge in the field. It is possible that new ecosystem-level or pathway-level concepts and methods need to be developed to develop such an understanding.

Conclusions

Although there is great interest in the microbiome, there is still a long way to go before microbiome-based diagnostics become a routine part of clinical care. Microbiome studies have been enormously valuable both in understanding mechanisms of disease in animal models and finding associations with disease in human beings. A good analogy is machine translation of natural languages: there has been interest since the 1950s, and poorly functioning systems have been available since the 1980s, but only in the past couple of years has it been possible to have a conversation with someone who speaks no common language using a mobile app on a smartphone, or to translate signs or menus from Chinese into English in real time using that smartphone's camera. In the same way, microbiome testing right now is primarily of interest as a science project. However, there will be rapid progress in the near term to develop better technical capability, including better user interfaces with readouts at the level of bacterial strains, and integration of ecologic dynamic concepts to better understand the transitions from health to illness.

Abbreviations used in this paper:

IBD	inflammatory bowel disease
PCR	polymerase chain reaction
rRNA	ribosomal RNA

References

1. Turnbaugh PJ, Hamady M, Yatsunenko T, et al. A core gut microbiome in obese and lean twins. Nature 2009;457:480–484. [PubMed: 19043404]
2. Ridaura VK, Faith JJ, Rey FE, et al. Gut microbiota from twins discordant for obesity modulate metabolism in mice. Science 2013;341:1241214. [PubMed: 24009397]
3. Le Chatelier E, Nielsen T, Qin J, et al. Richness of human gut microbiome correlates with metabolic markers. Nature 2013; 500:541–546. [PubMed: 23985870]
4. Cotillard A, Kennedy SP, Kong LC, et al. Dietary intervention impact on gut microbial gene richness. Nature 2013; 500:585–588. [PubMed: 23985875]
5. Frank DN, St Amand AL, Feldman RA, et al. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci U S A 2007;104:13780–13785. [PubMed: 17699621]
6. 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. [PubMed: 24629344]
7. Schirmer M, Smeekens SP, Vlamakis H, et al. Linking the human gut microbiome to inflammatory cytokine production capacity. Cell 2016;167:1125–1136 e8. [PubMed: 27814509]
8. Henao-Mejia J, Elinav E, Jin C, et al. Inflammasome-mediated dysbiosis regulates progression of NAFLD and obesity. Nature 2012;482:179–185. [PubMed: 22297845]
9. Hoyles L, Fernandez-Real JM, Federici M, et al. Molecular phenomics and metagenomics of hepatic steatosis in nondiabetic obese women. Nat Med 2018;24:1070–1080. [PubMed: 29942096]
10. Tripathi A, Debelius J, Brenner DA, et al. The gut-liver axis and the intersection with the microbiome. Nat Rev Gastroenterol Hepatol 2018;15:397–411. [PubMed: 29748586]

11. Xie G, Wang X, Liu P, et al. Distinctly altered gut microbiota in the progression of liver disease. Oncotarget 2016;7: 19355–19366. [PubMed: 27036035]

12. Grat M, Wronka KM, Krasnodebski M, et al. Profile of gut microbiota associated with the presence of hepatocellular cancer in patients with liver cirrhosis. Transplant Proc 2016; 48:1687–1691. [PubMed: 27496472]

13. Xie G, Wang X, Zhao A, et al. Sex-dependent effects on gut microbiota regulate hepatic carcinogenic outcomes. Sci Rep 2017;7:45232. [PubMed: 28345673]

14. Shalapour S, Lin XJ, Bastian IN, et al. Inflammation-induced IgAþ cells dismantle anti-liver cancer immunity. Nature 2017; 551:340–345. [PubMed: 29144460]

15. Knight R, Callewaert C, Marotz C, et al. The microbiome and human biology. Annu Rev Genomics Hum Genet 2017; 18:65–86. [PubMed: 28375652]

16. Turnbaugh PJ, Ley RE, Hamady M, et al. The human microbiome project. Nature 2007;449:804–810. [PubMed: 17943116]

17. Savage DC. Microbial ecology of the gastrointestinal tract. Annu Rev Microbiol 1977;31:107–133. [PubMed: 334036]

18. Sender R, Fuchs S, Milo R. Revised estimates for the number of human and bacteria cells in the body. PLoS Biol 2016; 14:e1002533. [PubMed: 27541692]

19. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 2012; 486:207–214. [PubMed: 22699609]

20. Eckburg PB, Bik EM, Bernstein CN, et al. Diversity of the human intestinal microbial flora. Science 2005;308:1635–1638. [PubMed: 15831718]

21. Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010; 464:59–65. [PubMed: 20203603]

22. Arumugam M, Raes J, Pelletier E, et al. Enterotypes of the human gut microbiome. Nature 2011;473:174–180. [PubMed: 21508958]

23. He Y, Wu W, Zheng HM, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat Med 2018;24:1532–1535. [PubMed: 30150716]

24. Thompson LR, Sanders JG, McDonald D, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 2017;551:457–463. [PubMed: 29088705]

25. McDonald D, Hyde E, Debelius JW, et al. American gut: an open platform for citizen science microbiome research. mSystems 2018;3.

26. Aguirre de Carcer D, Cuiv PO, Wang T, et al. Numerical ecology validates a biogeographical distribution and gender-based effect on mucosa-associated bacteria along the human colon. ISME J 2011;5:801–809. [PubMed: 21124491]

27. Lagier JC, Armougom F, Million M, et al. Microbial culturomics: paradigm shift in the human gut microbiome study. Clin Microbiol Infect 2012;18:1185–1193. [PubMed: 23033984]

28. Zengler K, Toledo G, Rappe M, et al. Cultivating the uncultured. Proc Natl Acad Sci U S A 2002;99:15681–15686. [PubMed: 12438682]

29. Browne HP, Forster SC, Anonye BO, et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. Nature 2016;533:543–546. [PubMed: 27144353]

30. Vandeputte D, Falony G, Vieira-Silva S, et al. Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. Gut 2016; 65:57–62. [PubMed: 26069274]

31. 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. [PubMed: 26604270]

32. Loftfield E, Vogtmann E, Sampson JN, et al. Comparison of collection methods for fecal samples for discovery metabolomics in epidemiologic studies. Cancer Epidemiol Biomarkers Prev 2016;25:1483–1490. [PubMed: 27543620]

33. Vogtmann E, Chen J, Amir A, et al. Comparison of collection methods for fecal samples in microbiome studies. Am J Epidemiol 2017;185:115–123. [PubMed: 27986704]

34. Costello EK, Lauber CL, Hamady M, et al. Bacterial community variation in human body habitats across space and time. Science 2009;326:1694–1697. [PubMed: 19892944]

35. Caporaso JG, Lauber CL, Costello EK, et al. Moving pictures of the human microbiome. Genome Biol 2011;12:R50. [PubMed: 21624126]

36. Zhang Z, Geng J, Tang X, et al. Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. ISME J 2014;8:881–893. [PubMed: 24132077]

37. Hillman ET, Lu H, Yao T, et al. Microbial ecology along the gastrointestinal tract. Microbes Environ 2017;32:300–313. [PubMed: 29129876]

38. Minich JJ, Zhu Q, Janssen S, et al. KatharoSeq enables highthroughput microbiome analysis from low-biomass samples. mSystems 2018;3.

39. Smarr L, Hyde ER, McDonald D, et al. Tracking human gut microbiome changes resulting from a colonoscopy. Methods Inf Med 2017;56:442–447. [PubMed: 29582916]

40. Halfvarson J, Brislawn CJ, Lamendella R, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol 2017;2:17004. [PubMed: 28191884]

41. Ley RE, Turnbaugh PJ, Klein S, et al. Microbial ecology: human gut microbes associated with obesity. Nature 2006; 444:1022–1023. [PubMed: 17183309]

42. Wu GD, Chen J, Hoffmann C, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science 2011; 334:105–108. [PubMed: 21885731]

43. David LA, Maurice CF, Carmody RN, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature 2014; 505:559–563. [PubMed: 24336217]

44. Vazquez-Baeza Y, Gonzalez A, Xu ZZ, et al. Guiding longitudinal sampling in IBD cohorts. Gut 2018;67:1743–1745. [PubMed: 29055911]

45. Loomba R, Seguritan V, Li W, et al. Gut microbiome-based metagenomic signature for non-invasive detection of advanced fibrosis in human nonalcoholic fatty liver disease. Cell Metab 2017;25:1054–1062 e5. [PubMed: 28467925]

46. Caussy C, Hsu C, Lo MT, et al. Link between gut-microbiome derived metabolite and shared gene-effects with hepatic steatosis and fibrosis in NAFLD. Hepatology 2018.

47. Varadi L, Luo JL, Hibbs DE, et al. Methods for the detection and identification of pathogenic bacteria: past, present, and future. Chem Soc Rev 2017;46:4818–4832. [PubMed: 28644499]

48. Forslund K, Hildebrand F, Nielsen T, et al. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Nature 2015;528:262–266. [PubMed: 26633628]

49. Bassetti M, Merelli M, Temperoni C, et al. New antibiotics for bad bugs: where are we? Ann Clin Microbiol Antimicrob 2013; 12:22. [PubMed: 23984642]

50. Kempf VA, Trebesius K, Autenrieth IB. Fluorescent In situ hybridization allows rapid identification of microorganisms in blood cultures. J Clin Microbiol 2000;38:830–838. [PubMed: 10655393]

51. Harris DM, Hata DJ. Rapid identification of bacteria and Candida using PNA-FISH from blood and peritoneal fluid cultures: a retrospective clinical study. Ann Clin Microbiol Antimicrob 2013; 12:2. [PubMed: 23295014]

52. Zautner AE, Gross U, Emele MF, et al. More pathogenicity or just more pathogens?-On the interpretation problem of multiple pathogen detections with diagnostic multiplex assays. Front Microbiol 2017;8:1210. [PubMed: 28706515]

53. Huang RS, Johnson CL, Pritchard L, et al. Performance of the Verigene(R) enteric pathogens test, Biofire FilmArray gastrointestinal panel and Luminex xTAG(R) gastrointestinal pathogen panel for detection of common enteric pathogens. Diagn Microbiol Infect Dis 2016;86:336–339. [PubMed: 27720206]

54. Liu RH, Yang J, Lenigk R, et al. Self-contained, fully integrated biochip for sample preparation, polymerase chain reaction amplification, and DNA microarray detection. Anal Chem 2004; 76:1824–1831. [PubMed: 15053639]

55. Tringe SG, Hugenholtz P. A renaissance for the pioneering 16S rRNA gene. Curr Opin Microbiol 2008;11:442–446. [PubMed: 18817891]

56. Caporaso JG, Lauber CL, Walters WA, et al. Ultra-highthroughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 2012;6:1621–1624. [PubMed: 22402401]

57. Liu Z, DeSantis TZ, Andersen GL, et al. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res 2008;36:e120. [PubMed: 18723574]
58. Soergel DA, Dey N, Knight R, et al. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. ISME J 2012;6:1440–1444. [PubMed: 22237546]
59. Wang Q, Garrity GM, Tiedje JM, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 2007;73:5261–5267. [PubMed: 17586664]
60. Ogura Y, Ooka T, Asadulghani, et al. Extensive genomic diversity and selective conservation of virulence-determinants in enterohemorrhagic Escherichia coli strains of O157 and nonO157 serotypes. Genome Biol 2007;8:R138. [PubMed: 17711596]
61. Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. Annu Rev Genet 2004;38:525–552. [PubMed: 15568985]
62. Marotz CA, Sanders JG, Zuniga C, et al. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome 2018;6:42. [PubMed: 29482639]
63. Har-El R, Silberstein A, Kuhn J, et al. Synthesis and degradation of lac mRNA in E. coli depleted of 30S ribosomal subunits. Mol Gen Genet 1979;173:135–144. [PubMed: 386032]
64. Segata N, Boernigen D, Tickle TL, et al. Computational meta’omics for microbial community studies. Mol Syst Biol 2013; 9:666. [PubMed: 23670539]
65. Melnik AV, da Silva RR, Hyde ER, et al. Coupling targeted and untargeted mass spectrometry for metabolome-microbiome-wide association studies of human fecal samples. Anal Chem 2017;89:7549–7559. [PubMed: 28628333]
66. Aksenov AA, da Silva R, Knight R, et al. Global chemical analysis of biology by mass spectrometry. Nat Rev Chem 2017;1:0054.
67. Watrous J, Roach P, Alexandrov T, et al. Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci U S A 2012;109:E1743–E152. [PubMed: 22586093]
68. Wang M, Carver JJ, Phelan VV, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 2016; 34:828–837. [PubMed: 27504778]
69. Wang X, Du L, You J, et al. Fungal biofilm inhibitors from a human oral microbiome-derived bacterium. Org Biomol Chem 2012;10:2044–2050. [PubMed: 22281750]
70. Hsiao EY, McBride SW, Hsien S, et al. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell 2013;155:1451–1463. [PubMed: 24315484]
71. Goodrich JK, Di Rienzi SC, Poole AC, et al. Conducting a microbiome study. Cell 2014;158:250–262. [PubMed: 25036628]
72. Gilbert JA, Quinn RA, Debelius J, et al. Microbiome-wide association studies link dynamic microbial consortia to disease. Nature 2016;535:94–103. [PubMed: 27383984]
73. Debelius J, Song SJ, Vazquez-Baeza Y, et al. Tiny microbes, enormous impacts: what matters in gut microbiome studies? Genome Biol 2016;17:217. [PubMed: 27760558]
74. Knight R, Vrbanac A, Taylor BC, et al. Best practices for analysing microbiomes. Nat Rev Microbiol 2018;16:410–422. [PubMed: 29795328]
75. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 2014;15:R46. [PubMed: 24580807]
76. Kim D, Song L, Breitwieser FP, et al. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res 2016;26:1721–1729. [PubMed: 27852649]
77. Segata N, Bornigen D, Morgan XC, et al. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun 2013;4:2304. [PubMed: 23942190]
78. Pruesse E, Quast C, Knittel K, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 2007; 35:7188–7196. [PubMed: 17947321]
79. 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. [PubMed: 24288368]

80. McDonald D, Price MN, Goodrich J, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 2012; 6:610–618. [PubMed: 22134646]

81. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010;7:335–336. [PubMed: 20383131]

82. Amir A, McDonald D, Navas-Molina JA, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2017;2.

83. 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. [PubMed: 16820507]

84. Koljalg U, Larsson KH, Abarenkov K, et al. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. New Phytol 2005;166:1063–1068. [PubMed: 15869663]

85. Huson DH, Beier S, Flade I, et al. MEGAN Community Edition interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comput Biol 2016;12:e1004957. [PubMed: 27327495]

86. Abubucker S, Segata N, Goll J, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol 2012;8:e1002358. [PubMed: 22719234]

87. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455–477. [PubMed: 22506599]

88. Li D, Liu CM, Luo R, et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 2015;31:1674–1676. [PubMed: 25609793]

89. Eren AM, Esen OC, Quince C, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 2015; 3:e1319. [PubMed: 26500826]

90. Knights D, Kuczynski J, Charlson ES, et al. Bayesian community-wide culture-independent microbial source tracking. Nat Methods 2011;8:761–763. [PubMed: 21765408]

91. Lax S, Smith DP, Hampton-Marcell J, et al. Longitudinal analysis of microbial interaction between humans and the indoor environment. Science 2014;345:1048–1052. [PubMed: 25170151]

92. McIver LJ, Abu-Ali G, Franzosa EA, et al. bioBakery: a meta'omic analysis environment. Bioinformatics 2018;34:1235–1237. [PubMed: 29194469]

93. McMurdie PJ, Holmes S. Phyloseq: a bioconductor package for handling and analysis of high-throughput phylogenetic sequence data. Pac Symp Biocomput 2012;235–246. [PubMed: 22174279]

94. Yilmaz P, Kottmann R, Field D, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol 2011;29:415–420. [PubMed: 21552244]

95. Sinha R, Abu-Ali G, Vogtmann E, et al. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. Nat Biotechnol 2017;35:1077–1086. [PubMed: 28967885]

96. Pribram HF, Hass AC, Nishioka H. Radiographic localization of a spontaneous cerebrospinal fluid fistula. Case report. J Neurosurg 1966;24:1031–1033. [PubMed: 5936482]

97. Lozupone CA, Knight R. Global patterns in bacterial diversity. Proc Natl Acad Sci U S A 2007;104:11436–11440. [PubMed: 17592124]

98. Walters WA, Xu Z, Knight R. Meta-analyses of human gut microbes associated with obesity and IBD. FEBS Lett 2014; 588:4223–4233. [PubMed: 25307765]

99. Zhou Y, Xu ZZ, He Y, et al. Gut microbiota offers universal biomarkers across ethnicity in inflammatory bowel disease diagnosis and infliximab response prediction. mSystems 2018; 3.

100. Vazquez-Baeza Y, Hyde ER, Suchodolski JS, et al. Dog and human inflammatory bowel disease rely on overlapping yet distinct dysbiosis networks. Nat Microbiol 2016;1:16177. [PubMed: 27694806]

101. Sze MA, Schloss PD. Looking for a signal in the noise: revisiting obesity and the microbiome. MBio 2016;7.

102. Finucane MM, Sharpton TJ, Laurent TJ, et al. A taxonomic signature of obesity in the microbiome? Getting to the guts of the matter. PLoS One 2014;9:e84689. [PubMed: 24416266]

103. Goodrich JK, Waters JL, Poole AC, et al. Human genetics shape the gut microbiome. Cell 2014;159:789–799. [PubMed: 25417156]

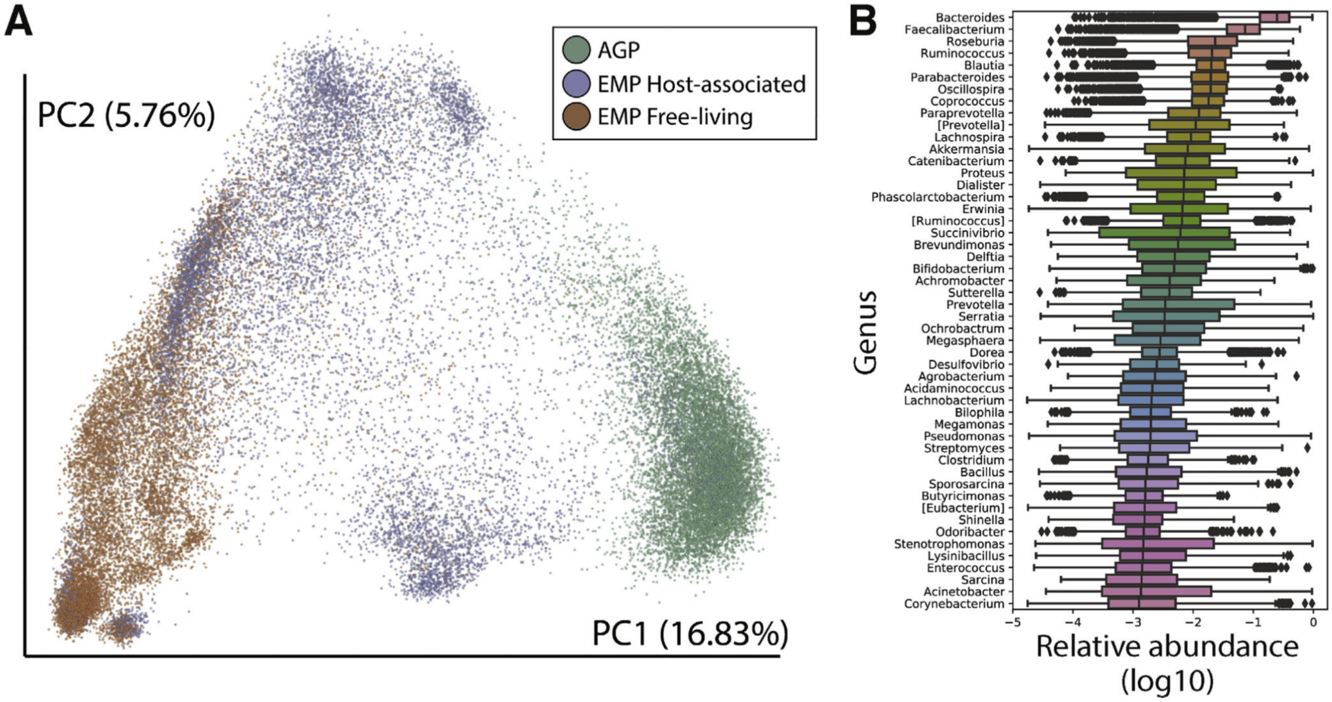


Figure 1.

Intersubject variability of the gut microbiome. (*A*) A principal coordinates plot of unweighted UniFrac distances computed using the Earth Microbiome Project (EMP) data set[24] and the fecal samples from the American Gut Project (AGP) data set.[25] Even though the EMP data include samples from many of the environments on the planet, including hydrothermal vents, soils, marine sediment, and many others, the extent of diversity associated with just the large intestine of a single mammal is one of the dominating clusters of microbial diversity. (*B*) Dynamic ranges of the 50 most abundant genera in the human fecal microbiome from 9316 individuals. These data are based off of a single sample per person, and only consider organisms observed in at least 100 people. Even though Bacteroides are ranked the highest, there are individuals with up to 3 orders of magnitude lower relative abundance of those genera, and that genera was not detected in approximately 1% of the individuals. PC, principal coordinates

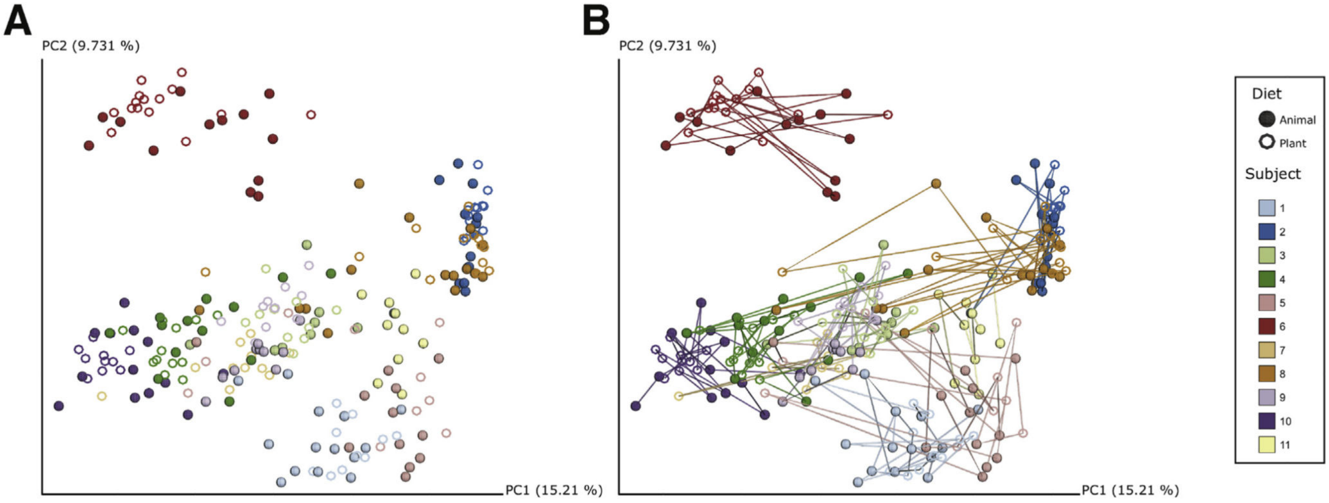


Figure 2.

Interindividual variability is a stronger discriminatory factor than diet, even under extreme dietary changes. (*A*) Principal coordinates analysis plot of unweighted UniFrac distances of the subjects (color) and their diets (shape). (*B*) Principal coordinates analysis plot with traces to show the individual variation over time, each edge is connected according to the collection time point.[43] PC, principal coordinates.

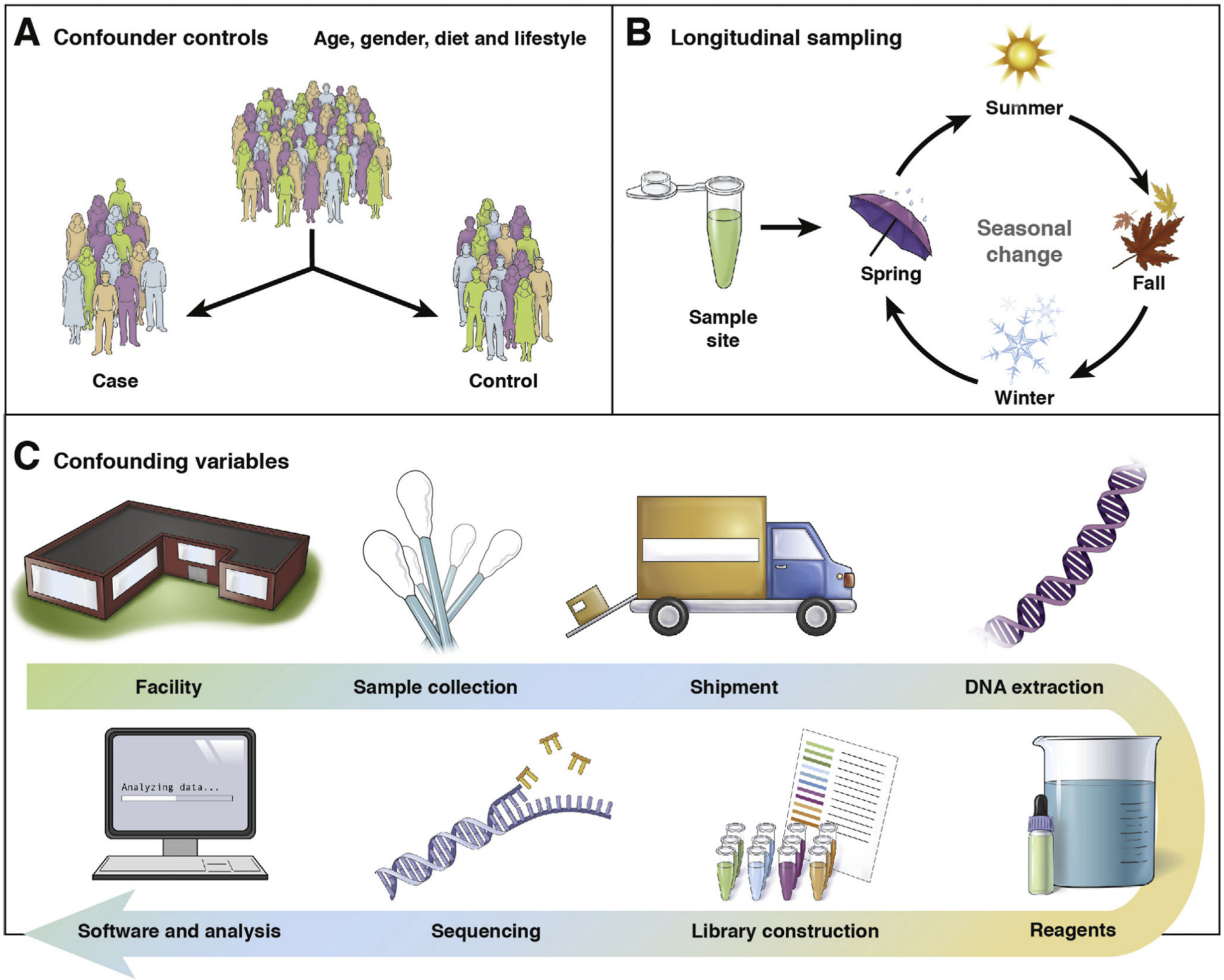
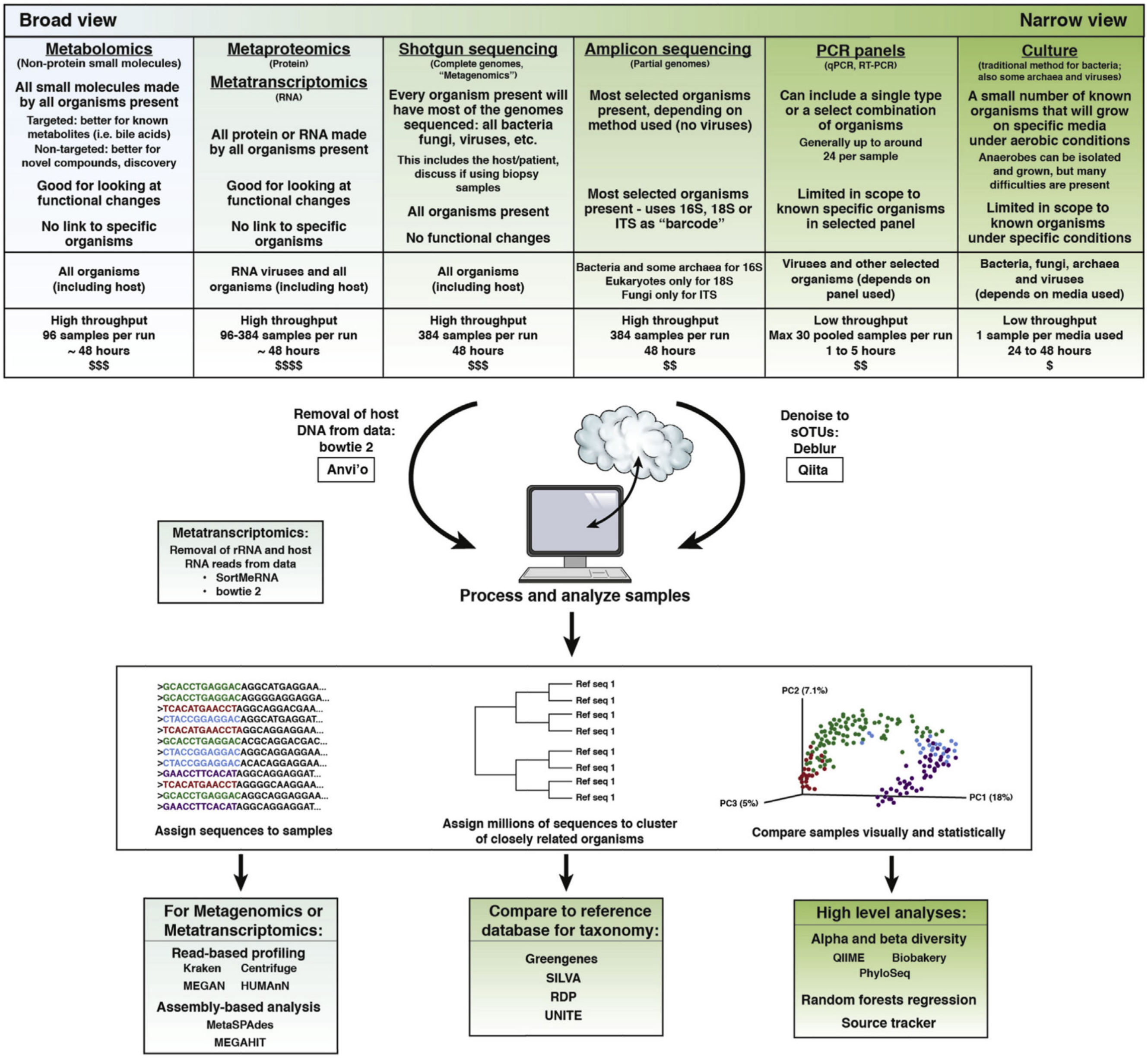


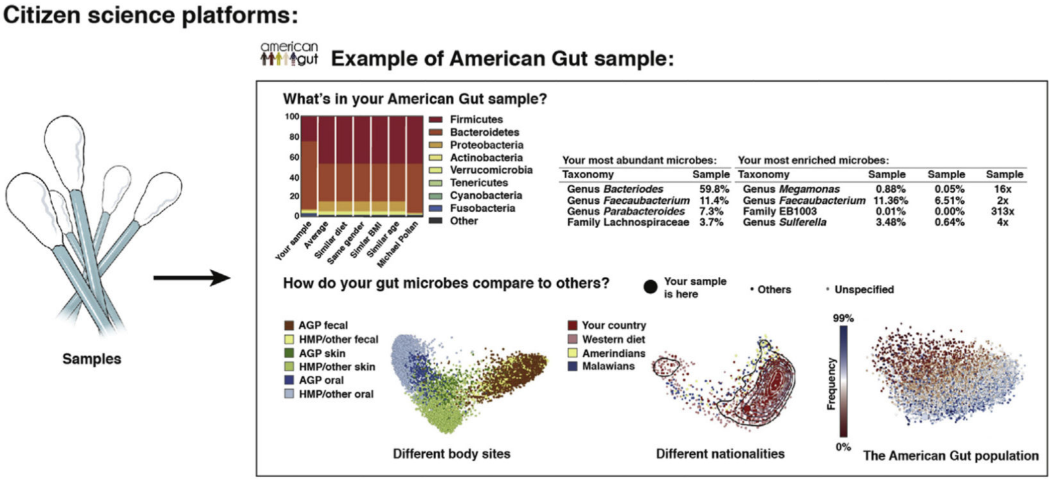
Figure 3.

Conducting a clinical microbiome experiment warrants careful attention to numerous factors. (*A*) Stratification by potential confounders (eg, age, sex, diet, lifestyle factors, and medications) can help resolve differences in microbiota between groups of interest that might otherwise be masked by a confounder effect.[48] (*B*) Longitudinal studies are especially powerful because they both control for confounding factors and allow for the assessment of community stability.[40] (*C*) For all studies, standardizing technical factors and sample processing are essential to control for variation introduced by every step of the process: kit reagents, primers, sample storage, and other factors. The collection and curation of metadata about all aspects of each sample, from clinical variables to sample processing, are crucial for data interpretation; without metadata, it is difficult to draw meaningful conclusions from sequencing data



Broad view					Narrow view
Metabolomics (Non-protein small molecules)	**Metaproteomics** (Protein) **Metatranscriptomics** (RNA)	**Shotgun sequencing** (Complete genomes, "Metagenomics")	**Amplicon sequencing** (Partial genomes)	**PCR panels** (qPCR, RT-PCR)	**Culture** (traditional method for bacteria; also some archaea and viruses)
All small molecules made by all organisms present Targeted: better for known metabolites (i.e. bile acids) Non-targeted: better for novel compounds, discovery Good for looking at functional changes No link to specific organisms	All protein or RNA made by all organisms present Good for looking at functional changes No link to specific organisms	Every organism present will have most of the genomes sequenced: all bacteria fungi, viruses, etc. This includes the host/patient, discuss if using biopsy samples All organisms present No functional changes	Most selected organisms present, depending on method used (no viruses) Most selected organisms present - uses 16S, 18S or ITS as "barcode"	Can include a single type or a select combination of organisms Generally up to around 24 per sample Limited in scope to known specific organisms in selected panel	A small number of known organisms that will grow on specific media under aerobic conditions Anaerobes can be isolated and grown, but many difficulties are present Limited in scope to known organisms under specific conditions
All organisms (including host)	RNA viruses and all organisms (including host)	All organisms (including host)	Bacteria and some archaea for 16S Eukaryotes only for 18S Fungi only for ITS	Viruses and other selected organisms (depends on panel used)	Bacteria, fungi, archaea and viruses (depends on media used)
High throughput 96 samples per run ~ 48 hours $$$	High throughput 96-384 samples per run ~ 48 hours $$$$	High throughput 384 samples per run 48 hours $$$	High throughput 384 samples per run 48 hours $$	Low throughput Max 30 pooled samples per run 1 to 5 hours $$	Low throughput 1 sample per media used 24 to 48 hours $

Removal of host DNA from data: bowtie 2
Anvi'o
Denoise to sOTUs: Deblur
Qiita
Metatranscriptomics:
Removal of rRNA and host RNA reads from data
• SortMeRNA
• bowtie 2
Process and analyze samples
>GCACCTGAGGACAGGCATGAGGAA...
>GCACCTGAGGACAGGGGAGGAGGA...
>TCACATGAACCTAGGCAGGACGAA...
>CTACCGGAGGACAGGCATGAGGAT...
>TCACATGAACCTAGGCAGGAGGAA...
>GCACCTGAGGACACGCAGGACGAC...
>CTACCGGAGGACAGGCAGGAGGAA...
>CTACCGGAGGACACACAGGAGGAA...
>GAACCTTCACATAGGCAGGAGGAT...
>TCACATGAACCTAGGGGCAAGGAA...
>GCACCTGAGGACAGGCAGGAGGAA...
>GAACCTTCACATAGGCAGGAGGAT...
Assign sequences to samples
Ref seq 1
Ref seq 1
Ref seq 1
Ref seq 1
Ref seq 1
Ref seq 1
Ref seq 1
Ref seq 1
Assign millions of sequences to cluster of closely related organisms
PC2 (7.1%)
PC3 (5%)
PC1 (18%)
Compare samples visually and statistically
For Metagenomics or Metatranscriptomics:
Read-based profiling
Kraken Centrifuge
MEGAN HUMAnN
Assembly-based analysis
MetaSPAdes
MEGAHIT
Compare to reference database for taxonomy:
Greengenes
SILVA
RDP
UNITE
High level analyses:
Alpha and beta diversity
QIIME Biobakery
PhyloSeq
Random forests regression
Source tracker



Your most abundant microbes:

Taxonomy	Sample
Genus *Bacteriodes*	59.8%
Genus *Faecaubacterium*	11.4%
Genus *Parabacteroides*	7.3%
Family Lachnospiraceae	3.7%

Your most enriched microbes:

Taxonomy	Sample	Sample	Sample
Genus *Megamonas*	0.88%	0.05%	16x
Genus *Faecaubacterium*	11.36%	6.51%	2x
Family EB1003	0.01%	0.00%	313x
Genus *Sulferella*	3.48%	0.64%	4x

Figure 4.

Once samples are collected, the samples can be put through molecular preparations and DNA sequencing to generate microbiome data. Two common types of protocols are amplicon sequencing and shotgun sequencing. In amplicon sequencing, PCR primers are used to target a specific region of a specific gene, focusing sequencing effort on just those fragments. One of the most widely used protocols targets the V4 region of the 16S rRNA gene.[24] In shotgun sequencing, the DNA in the sample is randomly sheared and sequenced, generating data from many different parts of the genome. The specifics of the molecular protocol used before shotgun sequencing are important for what type of data are being examined, and this type of sequencing can be used, for example, for metagenomics and metatranscriptomics. The initial processing performed on the data after sequencing depends on the type of sequencing performed. For amplicon studies, one common strategy is to upload the data into Qiita[81] and to use Deblur[82] to resolve sequence data into single-sequence variants called suboperational taxonomic units (sOTUs). Taxonomic assignments generally are performed using naive Bayes classifiers such as the RDP classifier,[59] as implemented in the q2-feature-classifier against reference databases such as Greengenes,[83] SILVA,[78] RDP,[79] or UNITE[84] (fungal internal transcribed spacer [ITS]) depending on the amplicon target. Shotgun sequencing of host-associated samples first requires preprocessing to remove either host DNA before analysis. Typically, the shotgun data then are summarized using tools such as Kraken,[75] MEGAN,[85] or HUMAnN2[86] to generate taxonomic or functional profiles, or are assembled with tools such as metaSPAdes[87] and MEGAHIT.[88] For both sequencing methods, higher-level analyses (eg, α and β diversity, taxonomic profiling, and machine learning) subsequently are used to assay patterns of microbiome variation in the context of the study design. Metagenomic assemblies also can be analyzed through platforms such as Anvi'o.[89] SourceTracker,[90] a Bayesian estimator of the sources that make up each unknown community, is useful for classifying microbial samples according to the environment of origin.[91] Citizen Science platforms, such as the American Gut Project,[25] standardize the molecular work and bioinformatic processing to generate a basic summary report of the content of an individuals sample. In the case of the American Gut Project, the samples also are placed into the context of a few other popular microbiome studies through data integration.