Published in final edited form as:

Ann Epidemiol. 2016 May ; 26(5): 311–321. doi:10.1016/j.annepidem.2016.04.005.

Intricacies of assessing the human microbiome in epidemiological studies

Courtney K. Robinson^a, Rebecca M. Brotman^{a,b}, and Jacques Ravel^{a,c}

^a Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD

^b Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD

^c Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD

Abstract

Purpose—In the past decade, remarkable relationships have been documented between dysbiosis of the human microbiota and adverse health outcomes. This review seeks to highlight some of the challenges and pitfalls that may be encountered during all stages of microbiota research, from study design and sample collection, to nucleic acid extraction and sequencing, and bioinformatic and statistical analysis.

Methods—Literature focused on human microbiota research was reviewed and summarized.

Results—While most studies have focused on surveying the composition of the microbiota, fewer have explored the causal roles of these bacteria, archaea, viruses, and fungi in affecting disease states. Microbiome research is in its relatively early years and many aspects remain challenging, including the complexity and personalized aspects of microbial communities, the influence of exogenous and often confounding factors, the need to apply fundamental principles of ecology and epidemiology, the necessity for new software tools, and the rapidly evolving genomic, technological, and analytical landscapes.

Conclusions—Incorporating human microbiome research in large epidemiological studies will soon help us unravel the intricate relationships that we have with our microbial partners and provide interventional opportunities to improve human health.

Keywords

16S ribosomal RNA gene; Metagenomics; Microbiota

Corresponding authors: Rebecca M. Brotman, PhD, MPH, Institute for Genome Sciences, University of Maryland School of Medicine, 801 West Baltimore Street, 6th Floor, Baltimore, MD 21201, phone: (410) 706-6767, fax: (410) 706-1482, ; Email: rbrotman@som.umaryland.edu Jacques Ravel, PhD, Institute for Genome Sciences, University of Maryland School of Medicine, 801 West Baltimore Street, 6th Floor, Baltimore, MD 21201, phone: (410) 706-5674, fax: (410) 706-1482, ; Email: jravel@som.umaryland.edu

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

It is believed that in the human body, microorganisms are more numerous than human somatic and germ cells [1]. Together, the genomes of these microbial mutualists (collectively defined as the metagenome) provide traits and services to humans, and in some cases, are associated with disease pathogenesis [2]. Over the past 10 years, with the advent of high-throughput sequencing technologies, there has been an exponential increase in molecular studies of the human microbiome. Rather than relying solely on bacterial cultivation for identification, partial sequencing of the bacterial 16S rRNA gene has become the standard in cataloguing organisms in biological samples.

If humans are thought of as a composite of microbial and human cells, and the human genetic landscape as an aggregate of the human genome, the microbiota (bacteria, archaea, and lower eukaryotes) and the virome (the collective set of bacteriophages and viruses), then the picture that emerges is one of a human 'supra-organism' [3]. It therefore becomes necessary to consider human health and disease outcomes in the context of our microbial partners. Microbiology is now entering a new era where the focus moves from the properties of single organisms in isolation to the operations of whole communities. The new field of metagenomics involves the genomic characterization of the entire microbial communities and not just cultivation of single organisms.

Molecular methods for interrogating microbial communities have led to a better understanding of the organisms present at specific sites on the human body and their potential roles in human health. The respective microbiota in each body niche can influence a wide variety of health outcomes including obesity [4], brain chemistry [5], ulcerative colitis [6], gynecologic and obstetric health [7], and periodontal disease [8]. Efforts to describe a "core" human microbiome, in the hopes of providing a baseline for comparisons [9], have proven to be challenging because bacterial communities show high inter-subject variability in species composition [10], while functional gene expression is more conserved [9].

With large datasets capturing many dimensions of the microbiota, including diversity, relative abundance and absolute abundance of bacterial taxa, as well as functional measurements of the microenvironment, there are tremendous opportunities for epidemiological studies to describe the microbiota's role in transitions between healthy and disease states. To date, most studies have focused on quantifying the statistical associations between the compositions of the human microbiota with health outcomes, however, fewer have been able to document how microbial changes are part of the causal chain leading to disease [8, 11].

Early studies on microbes were constrained to culture-based methods that were limited by the large numbers of species that resisted cultivation. While cultivation of microbes has improved and the proportion of organisms not yet cultivated is rapidly decreasing, the development of molecular methods for characterizing the microbiota, including marker gene amplicon, metagenomic, and metatranscriptomic sequencing, brought about rapid access to the identification and genomic information of previously uncultivated organisms. Marker

gene amplicon (mainly 16S rRNA gene) sequencing involves interrogating a single gene to identify which species are present. Combined with broad and species-specific quantitative PCR, this approach affords cataloguing species and their abundance in biological samples. For an overview of the human microbiome and 16S rRNA gene-based analyses for characterizations of the human microbiota, as well as terminology in this rapidly evolving field, we refer the reader to an excellent review by Tyler *et al.* [12] as well as an editorial by Marchesi and Ravel [13].

To gain insight into the functional make up of microbial communities, metagenomic sequencing is applied by sequencing all of the DNA recovered from a sample. Analyzing these reads can identify what organisms are present and the community's genomic content and functional potential. Metatranscriptomic sequencing, which surveys expressed genes in a sample, defines the function of the community at the time of sampling. These approaches could be further expanded by looking at the metaproteome [14-16] or the community metabolic outcomes, the metabolome [17-19]. Recent technological advances in high-throughput sequencing has enabled the parallel processing of large number of samples at affordable costs. As a consequence, these methodologies can now be integral to large-scale epidemiological studies.

In this review, we seek to detail what is involved with analyses of the human microbiota from an epidemiological perspective, with specific attention to the associated difficulties in designing, executing, and interpreting studies of the human microbiome. **Figure 1** presents a sample workflow for conducting a 16S rRNA sequencing study, and while the details would differ when conducting a metagenomic or metatranscriptomic study, this flow chart highlights the issues to consider at each step of the process.

Sample collection and storage conditions

One of the first issues that arises when planning epidemiological studies on of the human microbiome is determining collection methods for the samples. Collection should recover samples that are representative of the true microbiota present at the site, while limiting sampling biases and contamination. Less invasive sampling methods encourages recruitment and retention of study participants, and a pilot study can help inform and validate sampling methods. For example, recent studies on the methods for sampling the sinonasal microbiota [20] and intestinal mucosa [21] found the less invasive methods provided samples that had consistent microbiota profiles with samples obtained using classical sampling methods. In contrast, fecal transport swabs recovered less DNA and showed altered microbiota profiles compared to that of fecal material samples [22], stressing the importance of validating collection methods.

An important aspect of sampling strategy also includes sampling frequencies, which if performed in a clinical setting is often limited by the willingness of participants to return to the study site frequently as well as staffing requirements. However, participants are capable and willing to perform self-sampling at home and with high compliance rates [7, 23-29], thus enabling large field-based longitudinal epidemiological studies. Numerous groups have validated the use of self-collected samples compared to clinician-collected samples for

microbiome studies and pathogen detection, as well as confirmed uniformity from repeated sampling at the same sitting [30-33]. The number of samples to be collected at each time point should also be considered. Excessive sampling can be difficult from a human subject perspective, and may in itself disturb the microenvironment thus introducing compounding biases over time, making it potentially difficult to interpret longitudinal patterns of change.

Following sample collection, it is then important to take into consideration methods for sample transport and both short-term and long-term storage. Delays often occur between sampling and final storage because of logistical issues, and it is not always possible to process samples immediately after collection. Numerous studies have evaluated the effect of temperature and duration of storage on fecal samples and have found conflicting results in terms of the effect on microbiota composition based on 16S rRNA gene profiling, with some samples showing little change [22, 34-37] and others showing significant differences [38, 39]. Amies transport media has been a successful choice for preserving fecal [40, 41], vaginal [7, 31, 42], and nasal [43] samples for DNA extraction and sequencing. Samples taken for transcriptomic analysis need to be stored appropriately to minimize RNA degradation, so preservation with guanidine thiocyanate is usually used to prevent nucleases from degrading RNA molecules [44]. RNeasy has been used successfully for recovery of DNA and RNA from fecal samples [38, 44, 45] and saliva [46].

DNA/RNA extraction, 16S rRNA gene amplification, and library preparation

A critical step to microbiome analyses is DNA extraction, as in principle this is where most biases could be introduced, mostly from uneven cell lysis across the microbial community. Cell lysis, typically achieved through enzymatic and/or mechanical manipulations, would ideally work on all cell types equally, resulting in DNA being representative of the composition of the starting material. However, cells can vary in their susceptibility to lysing methods, with some lysing under fairly gentle conditions, and others, particularly Gram-positive organisms or spores, needing much harsher conditions that may result in shearing of DNA from easily-lysed organisms. Several studies have shown the use of mechanical lysis gives the highest bacterial diversity in 16S rRNA gene surveys [47, 48], and performs particularly well in the recovery of Gram-positive organisms in fecal communities [49]. Oral samples extracted using either mechanical or enzymatic lysis steps have shown overall similar microbiota profiles based on 16S rRNA gene amplicon sequencing, but with higher recovery of certain taxa with either method [50]. It is therefore important to consider what types of organisms are expected in a specific sample when choosing an extraction method, and noting that no methods are inherently free of biases [48]. Similar considerations apply to RNA extraction methods.

Of vital importance at every stage of sample manipulation is minimizing the introduction of non-indigenous microbes or DNA. Any contamination from, for example, the lab environment, DNA/RNA extraction kits [51] or PCR reagents [52, 53], can be difficult to distinguish from the microbial content of the samples themselves. The effects of contamination (often present in very low abundance) are generally minimal when dealing with high biomass samples, however samples with low biomass can have so little template DNA that they produce 16S rRNA gene amplicons and metagenomic results that represent

the contaminating DNA and not the sample's true composition [54]. Preparing metagenomic sequencing libraries from low levels of input DNA (a situation encountered with low biomass samples) could result in enrichment of AT-rich DNA during amplification [55]. Dealing with this background amplification becomes a critical matter when working with these kinds of samples [56, 57]. In addition to including proper negative and positive controls to monitor for contamination at each step of the process and maximizing the genomic material used in experiments, Weiss *et al.* also suggests randomizing the order of extractions to control for batch effects that arise from contamination that is unique between different batches or lot numbers of reagents [58].

Phylogenetic analysis

Extracted DNA can be used for the phylogenetic molecular assessment of the composition of the microbiota, either through marker gene amplicons or metagenomic sequencing. Marker gene amplicons sequencing involves the enrichment of a targeted gene that is phylogenetically informative. The sequence of this gene is used to identify what taxa are present in the sample. The most commonly used marker gene is the 16S rRNA gene, which is ubiquitously found in all bacteria and archaea. This gene consists of nine hypervariable regions, the combined sequence of which is unique to bacterial or archaeal taxa and thus can be used for taxonomic classification by comparison to databases. Interspersed between these variable regions are conserved regions that can function as priming sites for “universal” amplification. There are some drawbacks to using the 16S rRNA gene for microbiota analysis studies, including 1) some “universal” primer combinations give poor amplification of certain taxonomic groups, leading to underrepresentation of these organisms [59]; 2) some variable regions lack specificity and may not be able to discriminate between taxa at the species level [60]; and 3) microorganisms contain varying copy numbers of the 16S rRNA gene, making quantification of relative abundance somewhat inaccurate [61]. The ubiquity of 16S rRNA gene, and the high amount of reference sequences deposited in databases, make it the most common target for phylogenetic analyses.

While the full 16S rRNA gene is over 1,500 base pairs long, high-throughput sequencing methods produce reads that are significantly shorter. One to three specific variable regions are targeted for amplification and subsequent taxonomic assignment and analysis. It is important to consider the ability of the targeted variable region(s) to discriminate amongst taxa known to be predominant in the types of samples to be analyzed and minimize the effects of known primer biases. Although primers are designed based on a consensus sequence, some taxa do have mismatches in these consensus regions, potentially leading to their underrepresentation in terms of relative abundance [59]. In addition, the information contained within each variable region can be more or less informative when it comes to taxonomic assignment. For example, the V6 region of 16S rRNA gene performs poorly compared to others in discriminating taxa in human gut samples [62]. Primers for amplification of the V3-V4 or V4 regions are able to detect both bacteria and archaea, so those would be optimal when interested in analyzing the archaeal portion of the microbiota in samples such as stool [63]. An assessment of commonly used primer pairs found that longer 16S rRNA gene amplicons did not necessarily confer better classification; rather, it was the specific target region, depending upon sample origin, that had the biggest impact on

classification [64, 65]. Thus it is recommended to select the PCR primers and the targeted hypervariable regions based on sample types.

The next steps in describing the molecular profile of microbial communities include sequencing of the amplified region or library and quality control of resulting sequence reads. These reads are then used for picking OTUs and/or taxonomic classifications which will be used for downstream analyses. The technical details of this are perhaps outside the scope of this review. Therefore, we have presented a more detailed description of sequencing technologies, sequence read pre-processing, and taxonomic classification in **Appendix A**.

Diversity measures and standardization of protocols

Statistical analysis of 16S rRNA gene sequence data could include the application of ecological concepts such as alpha diversity, an estimate of the mean diversity within a sample, and beta diversity, the comparison of diversity between samples. Alpha diversity describes the richness and evenness of the microbiota in a given sample. Common workflows of 16S rRNA gene sequence analysis include using relative abundances obtained with and without normalizing read counts between samples through subsampling [66, 67]. Work by McMurdie and Holmes has shown that such normalization procedures can lead to overestimates of differentially abundant species across samples and a loss of statistical power [68], instead recommending the use of unrarefied data set and statistical models that account for differences in total read counts between samples [68]. Alpha diversity measures could be used to explore differences between healthy and disease states in epidemiological studies. For example, an increase in Shannon diversity was found in vaginal samples from women diagnosed with bacterial vaginosis [42] as well as Caucasian women who delivered prematurely [69], while a decrease in diversity was observed in fecal samples from individuals with inflammatory bowel disease compared to those without [70]. Interpretation of these differences in ecological metrics should be done with care as it is still unclear how to translate these measures in clinical settings.

Comparisons between samples can be done with multivariate analysis, which can take into account the presence or absence of species and their abundance or phylogeny [65, 71]. Distance between samples can be calculated in a number of ways; Sorensen or Jaccard indexes consider the presence or absence of OTUs, while Bray-Curtis also takes into account OTU abundance. Jensen-Shannon divergence can also be calculated to show the levels of similarity between communities, [72] and has been used successfully in epidemiological studies [72-75]. UniFrac distances consider the phylogeny of member OTUs, and can be calculated either weighted or unweighted for OTU abundances [76]. Visualization of these distances can be accomplished with a clustering approach that produces a dendrogram demonstrating the similarity between samples. Principle coordinates analysis (PCoA) plot single points representing each sample in multiple dimensions which are separated by principle coordinates. Color coding of the samples according to metadata can reveal information about the factors driving similarities between samples, such as geographic location of subjects [77], or recovery time past infection [78].

Clearly there are a number of steps during the analysis process for biases to affect results and it can be challenging to determine if these are severe enough to significantly alter the observed microbiota from its native state. One option for evaluating if this is the case, is processing a “mock microbial community” that comprises known organisms in known quantities and proportions. An investigation of a mock community prepared by the Human Microbiome Project (HMP) compared a range of sample storage temperatures and two DNA extraction methods, and found that the different extractions resulted in microbial community composition and abundance that were statistically different, but provided consistent conclusions [79]. The HMP initially employed four sequencing centers to generate the data, providing an excellent opportunity for the evaluation of center-specific (technical) biases introduced when using the same methods [80]. When Schloss *et al.* [59] analyzed the results from HMP mock communities processed and sequenced at different centers, they found that based on non-metric multidimensional scaling, communities were clustering primarily based on the sequencing center, and secondarily by processing batches. Further, the HMP provided the opportunity to evaluate the effects of different 16S rRNA gene variable regions and bioinformatic analytical approaches on community composition [81]. This work led to standard protocols that could be applied by others and a better understanding of the biases of each platform [81]. One difficulty in this rapidly developing field is that protocols can become quickly outdated and need to be reworked with equivalent rigor when adapting to new sampling devices, storage systems, DNA extraction procedures, and more importantly, sequencing platforms. This is particularly challenging when trying to compare data with previously published studies. When possible, it is recommended to use data that was sequenced on the same platform, using the same variable region, and reprocess the data with current and approved bioinformatics tools, in order to detect any potential biases. The development of standardized protocols, which would likely be body-site specific, but could be utilized by all researchers, would make comparison more reliable. However, the development of such protocols is a challenge in itself [82].

Demystifying clustering methods and their applications

Microbial communities can be grouped according to community composition. These groups are generated using clustering algorithms that also take into account the relative abundances of all taxa detected in a sample. This was first described in the human gut [83], where the term enterotypes was first coined. In the genital tract, the more ecologically correct terminology, and generally applicable, community state types (CST) was used for the characterization of the vaginal microbiota [42, 73, 84]. However the term cervicotype has also been used when specifically analyzing cervical samples [85]. These classifications have been challenged recently as often a blurry line exist between groups and communities which appear to be distributed on a gradient and most samples fall somewhere between the extremes [86]. However, the bioinformatics approach to analyze the data can influence greatly the grouping outcome. Koren *et al.* cited that enterotypes identification depended not only on the structure of the data but also the methods applied to identifying clustering strength [87]. As a result, the decision of how to group communities into different enterotypes (how many? are there outliers? where to break a continuum?) remains open for discussion. It may be that the best clustering categories are those that have biological

relevance or best distinguish risk factors in relation to the health outcome of interest. Future epidemiologic studies will help us to better detail which taxa or communities of bacteria (as in the case of CST) are associated with various disease outcome.

To identify clusters of bacterial communities based on relative abundances of different phylotypes (to generate a CST), Gajer *et al.* computed Jensen-Shannon distances between all pairs of community states (samples) and then generated hierarchical clustering using the Jensen-Shannon distance data and Ward linkage [73]. However, should a large dataset be available, one could build a machine learning algorithm, such as Support Vector Machine (SVM), to make these CST assignments so that they are consistent and independent of the samples that are clustered. Such an algorithm would allow for CST assignments to be comparable across studies using the same 16S rRNA gene primers and sequencing platform. It is important to note that CSTs are not forced, pre-determined or assigned by eye, all common confusions observed in the literature. **Figure 2** shows the dendrogram of the resulting hierarchical clustering of over 3,938 samples processed in our research studies and illustrates how bacterial communities cluster to form CSTs. CSTs are most useful in reducing the complexity of the dataset and allowing epidemiological investigations of disease outcomes with a large number of samples. Data reduction methods such as these may also prove useful in identifying biomarkers associated with disease states, or even susceptibility to diseases before they occur. A recent study incorporated 16S rRNA relative abundance data of gut samples with clinical data to improve the accuracy of predictive models in discriminating between healthy and colorectal cancer patients [88]. Longitudinal study designs are also valuable in facilitating the identification of biomarkers that appear before the onset of disease or detecting changes in the microbiota with fluctuations in symptoms.

Intersection of human microbiome analysis and epidemiology

We have only begun to understand the importance of our human microbiome and how changes in its composition and function can affect our health. Our behaviors, such as smoking, diet, and hygiene do not just affect us, they also affect our microbial partners. Surprisingly little is known about the composition of the human microbiome across the lifespan, how common human activities affect the structure of the microbiota, the correlation with the immunologic microenvironment, or the associations with disease susceptibilities and symptoms for example. Of course, this also presents one of the most challenging aspects of human microbiome research in modeling of microbial communities in the presence of co-factors associated with human behaviors and characteristics.

While the bacterial portion of the microbiota have been a popular area of research, newer work has also focused on characterizing the mycobiome, the fungal portion of the microbiota, and the virome, comprising viruses and bacterial phages. Much like bacterial and archaeal microbiota research has focused on sequencing of the 16S gene, mycobiome research is relying on sequencing part of the 18S rRNA gene and/or ITS (internal transcribed spacer) regions for taxonomic classification. Some of the limitations in analyzing these fungal communities involve disagreements over the optimal 18S or ITS regions for analysis, as well as a lack of robust databases for classification [89]. Virome research is particularly

handicapped by the lack of conserved genetic material among all viruses, so there is no equivalent to the 16S rRNA gene for targeted sequencing and identification of all viruses. Research can be targeted to specific viral groups using target hybridization and enrichment prior to sequencing [90, 91]. Metagenomic sequencing of all DNA from the sample can aid in identifying novel DNA viruses, or known viruses that are unexpected at that body site, but will miss RNA viruses, which would need to be identified through metatranscriptomic analyses. Metagenomic sequencing, for example, has been successfully used to interrogate viral communities such as those inhabiting the respiratory tract in cystic fibrosis and non-cystic fibrosis individuals [92] and the human gut [93]. Virome research could also elucidate the effect of bacteriophages on the human microbiota, with possible applications as a tool to manipulate or modulate the microbiota [94].

Much of the published work on the human microbiome is highly descriptive, demonstrating differences in characteristics of the microbiota between sampling sites of the body [95], over time [73, 96], and between diseased and healthy states [97, 98]. Moving forward, being able to differentiate between correlations of the microbiota with a specific disease state and demonstrating causation is an exciting prospect, one in which an epidemiological approach will play a major role. One example of clear causation is the transplant of fecal materials from healthy donors (i.e., healthy microbiota) to treat patients with recurrent *Clostridium difficile* infection [99, 100]. The fecal microbiota before and after the transplant has been monitored using 16S rRNA gene sequencing for up to a year after treatment, and showed that post-transplant, treated patients experience a lasting increase in species diversity and changes in relative abundance of certain organisms similar to the healthy donor [100]. This experiment provides well-defined proof of concept that altering the microbiota can impact the disease state of a patient. Observational cohort studies that follow cohorts over long periods of time may also be useful in surveying the microbiota before, during, and after disease or dysbiosis, thus identifying if disruption or alteration of the microbiota is the result of disease, or potentially the causal factor to the disruption. While such large-scale studies may be difficult and expensive to run, another alternative that is often underutilized is capitalizing on existing sample repositories that include extensive and frequent sampling. More work needs to be done in centralizing information on existing repositories of both data and clinical samples.

Molecular epidemiology is a discipline that combines molecular approaches with classical epidemiology [101] and within its broad reach includes all the tenets of epidemiology to be applied to human microbiome research. Often overlooked in human microbiome study design are development of strong questionnaires capable of collecting validated information on confounding factors, selection of appropriate controls, adjusting for those intricate human factors as needed, and controlling for correlation between samples collected longitudinally. The field would also benefit from collaborations with behavioral researchers and network epidemiologists to collect better measurements of activities of the core groups studied. Measurement of interactions can be complicated in human microbiome analyses and resorting to basic approaches of stratification of samples in analysis may be necessary. Finally, journals frequently request release of de-identified data and the NIH is also now requiring release of genomic data (<https://gds.nih.gov/03policy2.html>). NIH's database of Genotypes and Phenotypes (dbGaP), which is a controlled-access database governed by a

data user certification protocol (DUC), is one preferred option for human microbiome studies, but other options exist. Public deposit of data has not been routine in epidemiological research and with a push toward greater collaboration and sharing of data, new discoveries will certainly result. A caveat of requiring data release is that older human subject consent forms may not contain specific language related to the release of information to a public or controlled-access database and therefore appropriations need to be met with the local institutional review board. The NIH recently launched the Big Data to Knowledge (BD2K) initiative to support development of new digital tools for analysis of biomedical research. This encourages wider sharing of data between researchers and can lead to new insights into the functional role of the microbiota in human health.

Conclusion

Research on the human microbiome is still in its infancy. Sequencing and other omics' technology continues to rapidly evolve. Methods to extract DNA vary, collection media and devices continue to be re-invented and certainly there is no template to statistically interrogate the complex communities of the microbiome. As we move forward into a relatively new field of inquiry, open access to data, as well as free exchange and comparisons of protocols, will help to solidify the field. We have the expectation that in the future, we will be able to harness the microbiome to improve human health. Therapy in the form of probiotics, prebiotics, using small molecules to control specific microbial biochemical reactions, could be used to manage, modulate or restore the microbiome and maintain homeostasis.

Acknowledgements

This work was funded by the National Institute of Allergy and Infectious Diseases K01-AI080974, U19-AI084044 and R01-AI116799.

Appendix A Appendix A

Sequencing technologies

Early sequencing studies were performed using Sanger sequencing, which yielded long read lengths (>800 bp), but with a low throughput (~ 100 reads per sample) and a higher cost (~ \$100-\$200 per sample). It was quickly replaced by the more economical (<\$10 per sample) and higher throughput (>1,000 reads per sample) second-generation sequencing techniques. While Roche/454 pyrosequencing technology [¹⁰²] has been commonly used, it has recently been replaced by Illumina sequencing technologies that offer higher throughput paired, but shorter (~150-300 bp) reads, and to a lesser extent the Ion Torrent's technology (generate single ~400 bp reads) reads. The latter relies on similar chemistry as the Roche/454 platform, but nucleotide incorporation detection is performed by measuring pH variation change due to the release of a hydrogen ion during ligation using an electronic pH detector [¹⁰³]. A study comparing 16S rRNA gene sequencing on the Illumina and Ion Torrent platform found slight differences in over/underrepresentation of species between them, and premature truncation of reads appears to be an issue with the Ion Torrent reads, leading to additional challenges for the analysis [¹⁰⁴]. PacBio's Single Molecule Real-Time (SMRT)

sequencing reads are increasingly being applied to human microbiome research because of the long (>1.5 kb) reads, hence affording sequencing the entire 16S rRNA gene. However, its adoption is slow because of its higher cost, poor sequence accuracy, and lower throughput [105, 106]. This single molecule long read technology platform might be well-suited to improve metagenome sequence quality by affording better quality genome assemblies, both on its own [107] and in a hybrid approach that incorporates Illumina sequencing [108]. However, its low throughput, hence low sequence sampling, still limit its use to specialized applications. All of these technologies rely on sequencing a mixture of DNA amplicon tagged with unique indexes indicating their origin. Post sequencing, reads are demultiplexed, where the unique index is used to assign each read back to its sample of origin, allowing for hundreds, or even thousands, of samples to be processed in a single sequencing run [109,111].

The Illumina's HiSeq and MiSeq instruments are both capable of paired-end sequencing (that is, can generate sequences from both ends of the target molecule). Therefore, although the read lengths of the HiSeq and MiSeq are only about 150 and 300 base pairs, respectively, overlaps between the forward and reverse reads can be used to assemble both reads to a final sequence of 300 to 500 base pairs. The most common sequencing errors with the Illumina sequencing platform are mismatches rather than insertions or deletions. HiSeq is capable of outputting 600 Gb per run [71] and can process thousands of samples per run [112]. While the MiSeq is not capable of such high throughput, it does offer longer reads (up to 300 bp) and a shorter run time (less than 2 days) than HiSeq (currently about 4 days on the HiSeq 4000) [71]. Studies employing Illumina sequencing have been used to survey the composition of the human microbiota at numerous body sites [113,119].

For a more thorough review of other sequencing technologies, including future technologies in development, and their applications, see Buermans and den Dunnen [105]. Regardless of the sequencing methodology chosen, it should be noted that proprietary technologies can result in a lack of transparency. When sequencing problems occur, it can be very challenging to troubleshoot the problems independently. Microbiome research is a rapidly developing field with major leaps forward in sequence read lengths and decreases in cost over short periods of time. For long-term studies, sequencing methods that were standard at the onset of the study could become outdated or even unavailable by the end of the study. The rapid changes in technology and their disparate availability can make developing standards for the field and comparisons with previous datasets challenging, as each technology has its own set of biases and shortcomings.

Bioinformatics analysis

Sequence pre-processing and quality control

There are several open source pipelines that have been developed to make analysis of 16S rRNA gene amplicon sequences more streamlined and user-friendly. Two of the more popular ones are Quantitative Insights into Microbial Ecology (QIIME) [66] and mothur [67]. Other algorithms or software are available independently, and may require more specialized knowledge to employ. A basic workflow of 16S rRNA gene amplicon sequence analysis is described below.

Because 16S rRNA gene sequences are clustered based upon similarity, sequencing errors could result in reads being misclassified into separate OTUs, leading to an overestimate of bacterial diversity [120, 121]. Thus, removing reads that have low quality scores (using average quality score cut-off), are short (below a given cut-off), and contains mismatches to the primers has been used [120, 122, 123]. Others have applied more conservative strategies where the reads are scanned for low quality regions, after which the read is then trimmed and reassessed for read length. This approach retains as much information as possible [124]. The amplification and sequencing step can lead to chimeric sequences [125, 126], which if not removed could bias the taxonomic assignments by artificially inflating OTU richness estimates [127]. A number of programs are available to detect and remove chimeric sequences, including ChimeraSlayer [128] and UCHIME [129].

Picking OTUs and taxonomic classification

After stringent quality control of the raw sequence reads and because of the large number of sequences obtained with newer technologies, sequences are clustered into operational taxonomic units (OTU) based on sequence similarities. There are several options to accomplish this step, closed- or open-reference-based clustering and *de novo* clustering [130-132]. In reference-based clustering, reads are compared to a database of known sequences (usually full-length sequences for 16S rRNA gene) and classified into taxonomic groups based on identity. This gives taxonomic classification up front and is a good option if working with sample types that are thoroughly covered by your chosen reference database, but can be problematic with unknown sequences that could be unclassified. In closed-reference clustering, reads with no match in the reference database are then excluded, while in open-reference they are included, representing novel diversity not captured by the reference database. *De novo* clustering involves grouping 16S rRNA gene sequences based on similarity, without references. The threshold for similarity is adjustable, with 97% being the most common choice, as it represents the minimum similarity that defines species. There are many clustering algorithms that have been implemented for this purpose [133-136]. A representative sequence of each OTU or consensus sequence is then compared with a database for taxonomic assignment, which is then transitively transferred to all sequence reads forming a given OTU. The size of the OTU represents its relative abundance in the sample. This approach is especially useful when researching previously unstudied environments as it can identify OTUs that are not closely related to known organisms. However, this process is not perfect, and sequence reads forming OTUs could comprise sequences that are not taxonomically identical; similarly, selection of OTU representative sequences could impact study outcomes.

Common reference databases used for 16S rRNA gene sequence taxonomic classification include the Ribosomal Database Project (RDP) [137, 138], SILVA [139], and Greengenes [140]. Relative abundance tables of the proportion of each OTU per sample allow comparisons of the microbiota composition between subjects, subjects' groups and possibly over time. Relative abundance is often used in these analyses, however it can present an imperfect understanding of how samples relate to one another. Information on absolute abundance is important, as communities with dramatic differences in absolute abundance, which might reflect major functional differences, could have similar relative abundance

composition (i.e., a sample having 50% relative abundance of a given OTU could have different implications if its overall bacterial load is 10^3 or 10^7). One option to complement 16S rRNA gene relative abundance, is to use 16S rRNA gene quantitative PCR and measure the total 16S rRNA gene copies (an estimate of bacterial load) [^{141,143}] and an OTU relative abundance to calculate an OTU absolute abundance [¹⁴⁴]. Alternatively, in known samples with limited diversity, targeted qPCR assays could be used to evaluate the absolute abundance of a specific taxa [¹⁴⁵].

List of abbreviations

CST	community state type
HMP	Human Microbiome Project
mRNA	messenger RNA
OTU	operational taxonomic unit
PCR	polymerase chain reaction
qPCR	quantitative PCR
rRNA	ribosomal RNA

References

1. Sender R, Fuchs S, Milo R. Revised estimates for the number of human and bacteria cells in the body. *bioRxiv*. 2016
2. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. *Nature*. 2007; 449(7164):804–10. [PubMed: 17943116]
3. Glendinning L, Free A. Supra-organismal interactions in the human intestine. *Frontiers in cellular and infection microbiology*. 2014; 4:47. [PubMed: 24795867]
4. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Microbial ecology: human gut microbes associated with obesity. *Nature*. 2006; 444(7122):1022–3. [PubMed: 17183309]
5. Messaoudi M, Lalonde R, Violle N, Javelot H, Desor D, Nejdi A, et al. Assessment of psychotropic-like properties of a probiotic formulation (*Lactobacillus helveticus* R0052 and *Bifidobacterium longum* R0175) in rats and human subjects. *British Journal of Nutrition*. 2011; 105(05):755–64. [PubMed: 20974015]
6. Nishikawa J, Kudo T, Sakata S, Benno Y, Sugiyama T. Diversity of mucosa-associated microbiota in active and inactive ulcerative colitis. *Scandinavian Journal of Gastroenterology*. 2009; 44(2):180–6. [PubMed: 18825588]
7. Ravel J, Brotman RM, Gajer P, Ma B, Nandy M, Fadrosh DW, et al. Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. *Microbiome*. 2013; 1(1): 29. [PubMed: 24451163]
8. Curtis MA, Zenobia C, Darveau RP. The Relationship of the Oral Microbiota to Periodontal Health and Disease. *Cell Host & Microbe*. 2011; 10(4):302–6. [PubMed: 22018230]
9. Consortium THMP. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012; 486(7402):207–14. [PubMed: 22699609]
10. Huse SM, Ye Y, Zhou Y, Fodor AA. A Core Human Microbiome as Viewed through 16S rRNA Sequence Clusters. *PloS one*. 2012; 7(6):e34242. [PubMed: 22719824]
11. Ma B, Forney LJ, Ravel J. Vaginal microbiome: rethinking health and disease. *Annu Rev Microbiol*. 2012; 66:371–89. [PubMed: 22746335]

12. Tyler AD, Smith MI, Silverberg MS. Analyzing the human microbiome: a "how to" guide for physicians. *The American journal of gastroenterology*. 2014; 109(7):983–93. [PubMed: 24751579]
13. Marchesi JR, Ravel J. The vocabulary of microbiome research: a proposal. *Microbiome*. 2015; 3:31. [PubMed: 26229597]
14. Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, Halfvarson J, et al. Shotgun metaproteomics of the human distal gut microbiota. *Isme J*. 2009; 3(2):179–89. [PubMed: 18971961]
15. Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, Pan C, et al. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS one*. 2012; 7(11):e49138. [PubMed: 23209564]
16. Kolmeder CA, de Been M, Nikkila J, Ritamo I, Matto J, Valmu L, et al. Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions. *PLoS One*. 2012; 7(1):e29913. [PubMed: 22279554]
17. Nicholson JK, Lindon JC. Systems biology: Metabonomics. *Nature*. 2008; 455(7216):1054–6. [PubMed: 18948945]
18. Holmes E, Wilson ID, Nicholson JK. Metabolic phenotyping in health and disease. *Cell*. 2008; 134(5):714–7. [PubMed: 18775301]
19. Yeoman CJ, Thomas SM, Miller ME, Ulanov AV, Torralba M, Lucas S, et al. A multi-omic systems-based approach reveals metabolic markers of bacterial vaginosis and insight into the disease. *PLoS One*. 2013; 8(2):e56111. [PubMed: 23405259]
20. Bassiouni A, Cleland EJ, Psaltis AJ, Vreugde S, Wormald P-J. Sinonasal microbiome sampling: a comparison of techniques. *PLoS ONE*. 2015; 10(4):e0123216. [PubMed: 25876035]
21. Huse SM, Young VB, Morrison HG, Antonopoulos DA, Kwon J, Dalal S, et al. Comparison of brush and biopsy sampling methods of the ileal pouch for assessment of mucosa-associated microbiota of human subjects. *Microbiome*. 2014; 2(1):5. [PubMed: 24529162]
22. Tedjo DI, Jonkers DMAE, Savelkoul PH, Masclee AA, van Best N, Pierik MJ, et al. The effect of sampling and storage on the fecal microbiota composition in healthy and diseased subjects. *PLoS ONE*. 2015; 10(5):e0126685. [PubMed: 26024217]
23. Brotman RM, Ghanem KG, Klebanoff MA, Taha TE, Scharfstein DO, Zenilman JM. The effect of vaginal douching cessation on bacterial vaginosis: a pilot study. *Am J Obstet Gynecol*. 2008; 198(6):628–7. [PubMed: 18295180]
24. Brotman RM, Melendez JH, Smith TD, Galai N, Zenilman JM. Effect of Menses on Clearance of Y-Chromosome in Vaginal Fluid: Implications for a Biomarker of Recent Sexual Activity. *Sex Transm Dis*. 2010; 37(1)
25. Brotman RM, He X, Gajer P, Fadrosch D, Sharma E, Mongodin EF, et al. Association between cigarette smoking and the vaginal microbiota: a pilot study. *BMC Infect Dis*. 2014; 14:471. [PubMed: 25169082]
26. Anderson DJ, Politch JA, Pudney J, Marquez CI, Snead MC, Mauck C. A quantitative glycogen assay to verify use of self-administered vaginal swabs. *Sex Transm Dis*. 2012; 39(12):949–53. [PubMed: 23191948]
27. Guan Y, Gravitt PE, Howard R, Eby YJ, Wang S, Li B, et al. Agreement for HPV genotyping detection between self-collected specimens on a FTA cartridge and clinician-collected specimens. *Journal of virological methods*. 2013; 189(1):167–71. [PubMed: 23370404]
28. Ndayisaba G, Verwijs MC, van Eeckhoudt S, Gasarabwe A, Hardy L, Borgdorff H, et al. Feasibility and acceptability of a novel cervicovaginal lavage self-sampling device among women in Kigali, Rwanda. *Sex Transm Dis*. 2013; 40(7):552–5. [PubMed: 23965769]
29. Feigelson HS, Bischoff K, Ardini MA, Ravel J, Gail MH, Flores R, et al. Feasibility of self-collection of fecal specimens by randomly sampled women for health-related studies of the gut microbiome. *BMC Res Notes*. 2014; 7:204. [PubMed: 24690120]
30. Menard JP, Fenollar F, Raoult D, Boubli L, Bretelle F. Self-collected vaginal swabs for the quantitative real-time polymerase chain reaction assay of *Atopobium vaginae* and *Gardnerella vaginalis* and the diagnosis of bacterial vaginosis. *Eur J Clin Microbiol Infect Dis*. 2012; 31(4): 513–8. [PubMed: 21789604]

31. Bai G, Gajer P, Nandy M, Ma B, Yang H, Sakamoto J, et al. Comparison of storage conditions for human vaginal microbiome studies. *PLoS One*. 2012; 7(5):e36934. [PubMed: 22655031]
32. Forney LJ, Gajer P, Williams CJ, Schneider GM, Koenig SS, McCulle SL, et al. Comparison of self-collected and physician-collected vaginal swabs for microbiome analysis. *J Clin Microbiol*. 2010; 48(5):1741–8. [PubMed: 20200290]
33. Flores R, Shi J, Gail MH, Gajer P, Ravel J, Goedert JJ. Assessment of the human faecal microbiota: II. Reproducibility and associations of 16S rRNA pyrosequences. *Eur J Clin Invest*. 2012; 42(8): 855–63. [PubMed: 22385292]
34. Wu GD, Lewis JD, Hoffmann C, Chen Y-Y, Knight R, Bittinger K, et al. Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiol*. 2010; 10(1):206–14. [PubMed: 20673359]
35. Fouhy F, Deane J, Rea MC, O’Sullivan Ó, Ross RP, O’Callaghan G, et al. The effects of freezing on faecal microbiota as determined using MiSeq sequencing and culture-based investigations. *PLoS ONE*. 2015; 10(3):e0119355. [PubMed: 25748176]
36. Lauber CL, Zhou N, Gordon JI, Knight R, Fierer N. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. *FEMS Microbiol Lett*. 2010; 307(1):80–6. [PubMed: 20412303]
37. Dominianni C, Wu J, Hayes RB, Ahn J. Comparison of methods for fecal microbiome biospecimen collection. *BMC Microbiol*. 2014; 14:103. [PubMed: 24758293]
38. Flores R, Shi J, Yu G, Ma B, Ravel J, Goedert JJ, et al. Collection media and delayed freezing effects on microbial composition of human stool. *Microbiome*. 2015; 3:33. [PubMed: 26269741]
39. Bahl MI, Bergström A, Licht TR. Freezing fecal samples prior to DNA extraction affects the Firmicutes to Bacteroidetes ratio determined by downstream quantitative PCR analysis. *FEMS Microbiology Letters*. 2012; 329(2):193–7. [PubMed: 22325006]
40. Chiu, CM.; Lin, FM.; Chang, TH.; Huang, WC.; Liang, C.; Wu, WY., et al. Clinical detection of human probiotics and human pathogenic bacteria by using a novel high-throughput platform based on next generation sequencing; *Proceedings Iwbbio 2013: International Work-Conference on Bioinformatics and Biomedical Engineering*; 2013. p. 29-40.
41. Chiu C-M, Huang W-C, Weng S-L, Tseng H-C, Liang C, Wang W-C, et al. Systematic Analysis of the Association between Gut Flora and Obesity through High-Throughput Sequencing and Bioinformatics Approaches. *BioMed Research International*. 2014; 2014(2014):10.
42. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, et al. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci USA*. 2011; 108(Suppl 1):4680–7.
43. Liu CM, Price LB, Hungate BA, Abraham AG, Larsen LA, Christensen K, et al. *Staphylococcus aureus* and the ecology of the nasal microbiome. *Science Advances*. Jun 05.2015 2015 :E1400216. [PubMed: 26601194]
44. Vl ková K, Mrázek J, Kope ný J, Petrželková KJ. Evaluation of different storage methods to characterize the fecal bacterial communities of captive western lowland gorillas (*Gorilla gorilla gorilla*). *Journal of Microbiological Methods*. 2012; 91(1):45–51. [PubMed: 22828127]
45. Nechvatal JM, Ram JL, Basson MD, Namprachan P, Niec SR, Badsha KZ, et al. Fecal collection, ambient preservation, and DNA extraction for PCR amplification of bacterial and human markers from human feces. *Journal of Microbiological Methods*. 2008; 72(2):124–32. [PubMed: 18162191]
46. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, et al. Relating the metatranscriptome and metagenome of the human gut. *Proc Natl Acad Sci USA*. 2014; 111(22):E2329–38. [PubMed: 24843156]
47. Salonen A, Nikkilä J, Jalanka-Tuovinen J, Immonen O, Rajili -Stojanovi M, Kekkonen RA, et al. Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: Effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *Journal of Microbiological Methods*. 2010; 81(2):127–34. [PubMed: 20171997]
48. Yuan S, Cohen DB, Ravel J, Abdo Z, Forney LJ. Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS One*. 2012; 7(3):e33865. [PubMed: 22457796]

49. Santiago A, Panda S, Mengels G, Martinez X, Azpiroz F, Dore J, et al. Processing faecal samples: a step forward for standards in microbial community analysis. *BMC Microbiol.* 2014; 14(1):112. [PubMed: 24884524]
50. Lazarevic V, Gaia N, Girard M, François P, Schrenzel J. Comparison of DNA Extraction Methods in Analysis of Salivary Bacterial Communities. *PLoS ONE.* 2013; 8(7):e67699. [PubMed: 23844068]
51. Mohammadi T, Reesink HW, Vandenbroucke-Grauls CMJE, Savelkoul PHM. Removal of contaminating DNA from commercial nucleic acid extraction kit reagents. *Journal of Microbiological Methods.* 2005; 61(2):285–8. [PubMed: 15722157]
52. Grahn N, Olofsson M, Ellnebo-Svedlund K, Monstein H-J, Jonasson J. Identification of mixed bacterial DNA contamination in broad-range PCR amplification of 16S rDNA V1 and V3 variable regions by pyrosequencing of cloned amplicons. *FEMS Microbiology Letters.* 2003; 219(1):87–91. [PubMed: 12594028]
53. Corless CE, Guiver M, Borrow R, Edwards-Jones V, Kaczmarek EB, Fox AJ. Contamination and Sensitivity Issues with a Real-Time Universal 16S rRNA PCR. *Journal of Clinical Microbiology.* 2000; 38(5):1747–52. [PubMed: 10790092]
54. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 2014; 12(1):87. [PubMed: 25387460]
55. Chafee M, Maignien L, Simmons SL. The effects of variable sample biomass on comparative metagenomics. *Environ Microbiol.* 2015; 17(7):2239–53. [PubMed: 25329041]
56. Aagaard K, Ma J, Antony KM, Ganu R, Petrosino J, Versalovic J. The Placenta Harbors a Unique Microbiome. *Sci Transl Med.* 2014; 6(237):237ra65.
57. Nakatsuji T, Chiang H-I, Jiang SB, Nagarajan H, Zengler K, Gallo RL. The microbiome extends to subepidermal compartments of normal skin. *Nature communications.* 2013; 4:1431.
58. Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R. Tracking down the sources of experimental contamination in microbiome studies. *Genome Biol.* 2014; 15(12):564. [PubMed: 25608874]
59. Schloss PD, Gevers D, Westcott SL. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PloS one.* 2011; 6(12):e27310. [PubMed: 22194782]
60. Kim M, Morrison M, Yu Z. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *Journal of Microbiological Methods.* 2011; 84(1):81–7. [PubMed: 21047533]
61. Kembel SW, Wu M, Eisen JA, Green JL. Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. *PLoS Comput Biol.* 2012; 8(10):e1002743. [PubMed: 23133348]
62. Liu Z, DeSantis TZ, Andersen GL, Knight R. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.* 2008; 36(18):e120–e. [PubMed: 18723574]
63. Takahashi S, Tomita J, Nishioka K, Hisada T, Nishijima M. Development of a Prokaryotic Universal Primer for Simultaneous Analysis of Bacteria and Archaea Using Next-Generation Sequencing. *PLoS ONE.* 2014; 9(8):e105592–9. [PubMed: 25144201]
64. Soergel DAW, Dey N, Knight R, Brenner SE. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *Isme J.* 2012; 6(7):1440–4. [PubMed: 22237546]
65. Hamady M, Knight R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome research.* 2009; 19(7):1141–52. [PubMed: 19383763]
66. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods.* 2010; 7(5):335–6. [PubMed: 20383131]
67. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol.* 2009; 75(23):7537–41. [PubMed: 19801464]

68. McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Comput Biol*. 2014; 10(4)
69. Hyman RW, Fukushima M, Jiang H, Fung E, Rand L, Johnson B, et al. Diversity of the vaginal microbiome correlates with preterm birth. *Reproductive sciences*. 2014; 21(1):32–40. [PubMed: 23715799]
70. Ott SJ, Musfeldt M, Wenderoth DF, Hampe J, Brant O, Fölsch UR, et al. Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease. *Gut*. 2004; 53(5):685–93. [PubMed: 15082587]
71. Di Bella JM, Bao Y, Gloor GB, Burton JP, Reid G. High throughput sequencing methods and analysis for microbiome research. *Journal of Microbiological Methods*. 2013; 95(3):401–14. [PubMed: 24029734]
72. Lin JH. Divergence Measures Based on the Shannon Entropy. *Ieee T Inform Theory*. 1991; 37(1): 145–51.
73. Gajer P, Brotman RM, Bai G, Sakamoto J, Schutte UM, Zhong X, et al. Temporal dynamics of the human vaginal microbiota. *Sci Transl Med*. 2012; 4(132):132ra52.
74. Romero R, Hassan SS, Gajer P, Tarca AL, Fadrosch DW, Nikita L, et al. The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. 2014; 2(1):1–19.
75. Schubert AM, Rogers MA, Ring C, Mogle J, Petrosino JP, Young VB, et al. Microbiome data distinguish patients with *Clostridium difficile* infection and non-*C. difficile*-associated diarrhea from healthy controls. *mBio*. 2014; 5(3):e01021–14. [PubMed: 24803517]
76. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. *Isme J*. 2011; 5(2):169–72. [PubMed: 20827291]
77. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. *Nature*. 2012; 486(7402):222–7. [PubMed: 22699611]
78. David LA, Weil A, Ryan ET, Calderwood SB, Harris JB, Chowdhury F, et al. Gut Microbial Succession Follows Acute Secretory Diarrhea in Humans. *mBio*. 2015; 6(3):e00381–15. [PubMed: 25991682]
79. Hang J, Desai V, Zavaljevski N, Yang Y, Lin X, Satya RV, et al. 16S rRNA gene pyrosequencing of reference and clinical samples and investigation of the temperature stability of microbiome profiles. *Microbiome*. 2014; 2(1):1–15. [PubMed: 24468033]
80. Group NHW, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, et al. The NIH Human Microbiome Project. *Genome research*. 2009; 19(12):2317–23. [PubMed: 19819907]
81. Jumpstart Consortium Human Microbiome Project Data Generation Working G. Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLoS One*. 2012; 7(6):e39315. [PubMed: 22720093]
82. Huttenhower C, Knight R, Brown CT, Caporaso JG, Clemente JC, Gevers D, et al. Advancing the microbiome research community. *Cell*. 2014; 159(2):227–30. [PubMed: 25303518]
83. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. *Nature*. 2011; 473(7346):174–80. [PubMed: 21508958]
84. Zhou X, Bent SJ, Schneider MG, Davis CC, Islam MR, Forney LJ. Characterization of vaginal microbial communities in adult healthy women using cultivation-independent methods. *Microbiology (Reading, England)*. 2004; 150:2565–73. Pt 8.
85. Anahtar MN, Byrne EH, Doherty KE, Bowman BA, Yamamoto HS, Soumillon M, et al. Cervicovaginal bacteria are a major modulator of host inflammatory responses in the female genital tract. *Immunity*. 2015; 42(5):965–76. [PubMed: 25992865]
86. Jeffery IB, Claesson MJ, O'Toole PW, Shanahan F. Categorization of the gut microbiota: enterotypes or gradients? *Nat Rev Micro*. 2012; 10(9):591–2.
87. Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Knight R, et al. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput Biol*. 2013; 9(1):e1002863. [PubMed: 23326225]

88. Zackular JP, Rogers MAM, Ruffin MT, Schloss PD. The Human Gut Microbiome as a Screening Tool for Colorectal Cancer. *Cancer Prevention Research*. 2014; 7(11):1112–21. [PubMed: 25104642]
89. Cui L, Morris A, Ghedin E. The human mycobiome in health and disease. *Genome Med*. 2013; 5(7):63. [PubMed: 23899327]
90. Depledge DP, Palser AL, Watson SJ, Lai IY, Gray ER, Grant P, et al. Specific capture and whole-genome sequencing of viruses from clinical samples. *PLoS One*. 2011; 6(11):e27805. [PubMed: 22125625]
91. Duncavage EJ, Magrini V, Becker N, Armstrong JR, Demeter RT, Wylie T, et al. Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. *J Mol Diagn*. 2011; 13(3):325–33. [PubMed: 21497292]
92. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, et al. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One*. 2009; 4(10):e7370. [PubMed: 19816605]
93. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell*. 2015; 160(3):447–60. [PubMed: 25619688]
94. Scarpellini E, Ianiro G, Attili F, Bassanelli C, De Santis A, Gasbarrini A. The human gut microbiota and virome: Potential therapeutic implications. *Dig Liver Dis*. 2015; 47(12):1007–12. [PubMed: 26257129]
95. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. *Science (New York, NY)*. 2009; 326(5960):1694–7.
96. Zhao J, Schloss PD, Kalikin LM, Carmody LA, Foster BK, Petrosino JF, et al. Decade-long bacterial community dynamics in cystic fibrosis airways. *Proc Natl Acad Sci U S A*. 2012; 109(15):5809–14. [PubMed: 22451929]
97. Hilty M, Burke C, Pedro H, Cardenas P, Bush A, Bossley C, et al. Disordered microbial communities in asthmatic airways. *PLoS One*. 2010; 5(1):e8578. [PubMed: 20052417]
98. Karlsson FH, Tremaroli V, Nookaew I, Bergstrom G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*. 2013; 498(7452):99–103. [PubMed: 23719380]
99. Khoruts A, Dicksved J, Jansson JK, Sadowsky MJ. Changes in the Composition of the Human Fecal Microbiome After Bacteriotherapy for Recurrent *Clostridium difficile*-associated Diarrhea. *J Clin Gastroenterol*. 2010; 44(5):354. [PubMed: 20048681]
100. Song Y, Garg S, Girotra M, Maddox C, von Rosenvinge EC, Dutta A, et al. Microbiota Dynamics in Patients Treated with Fecal Microbiota Transplantation for Recurrent *Clostridium difficile* Infection. *PLoS ONE*. 2013; 8(11):e81330. [PubMed: 24303043]
101. Foxman B, Riley L. Molecular epidemiology: focus on infection. *Am J Epidemiol*. 2001; 153(12): 1135–41. [PubMed: 11415945]
102. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005; 437(7057):376–80. [PubMed: 16056220]
103. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. 2011; 475(7356):348–52. [PubMed: 21776081]
104. Salipante SJ, Kawashima T, Rosenthal C, Hoogestraat DR, Cummings LA, Sengupta DJ, et al. Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Appl Environ Microbiol*. 2014; 80(24):7583–91. [PubMed: 25261520]
105. Buermans HPJ, Den Dunnen JT. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. 2014; 1842(10):1932–41. [PubMed: 24995601]
106. Schloss PD, Westcott SL, Jenior ML, Highlander SK. Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ PrePrints*. 2015; 3(e778v1)

107. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods*. 2013; 10(6): 563–9. [PubMed: 23644548]
108. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol*. 2012; 30(7):693–700. [PubMed: 22750884]
109. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A*. 2011; 108:4516–22. [PubMed: 20534432]
110. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol*. 2013; 79(17):5112–20. [PubMed: 23793624]
111. Fadrosch DW, Ma B, Gajer P, Sengamalay N, Ott S, Brotman RM, et al. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*. 2014; 2(1):6. [PubMed: 24558975]
112. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, et al. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*. 2012; 13(1): 47–58.
113. Walujkar SA, Dhotre DP, Marathe NP, Lawate PS. Characterization of bacterial community shift in human Ulcerative Colitis patients revealed by Illumina based 16S rRNA gene amplicon sequencing. *Gut*. 2014
114. Ursell LK, Gunawardana M, Chang S, Mullen M, Moss JA, Herold BC, et al. Comparison of the vaginal microbial communities in women with recurrent genital HSV receiving acyclovir intravaginal rings. *Antiviral Research*. 2014; 102:87, 94. [PubMed: 24361269]
115. Pearce MM, Hilt EE, Rosenfeld AB, Zilliox MJ, Thomas-White K, Fok C, et al. The female urinary microbiome: a comparison of women with and without urgency urinary incontinence. *mBio*. 2014; 5(4):e01283–14. [PubMed: 25006228]
116. Fitz-Gibbon S, Tomida S, Chiu B-H, Nguyen L, Du C, Liu M, et al. *Propionibacterium acnes* Strain Populations in the Human Skin Microbiome Associated with Acne. *Journal of Investigative Dermatology*. 2013; 133(9):2152–60. [PubMed: 23337890]
117. Hasan NA, Young BA, Minard-Smith AT, Saeed K, Li H, Heizer EM, et al. Microbial Community Profiling of Human Saliva Using Shotgun Metagenomic Sequencing. *Plos One*. 2014; 9(5)
118. Maughan H, Wang PW, Diaz Caballero J, Fung P, Gong Y, Donaldson SL, et al. Analysis of the cystic fibrosis lung microbiota via serial Illumina sequencing of bacterial 16S rRNA hypervariable regions. *PLoS One*. 2012; 7(10):e45791. [PubMed: 23056217]
119. Jervis-Bardy J, Leong LEX, Marri S, Smith RJ, Choo JM, Smith-Vaughan HC, et al. Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. *Microbiome*. 2015; 3(1):19. [PubMed: 25969736]
120. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol*. 2010; 12(1):118–23. [PubMed: 19725865]
121. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods*. 2013; 10(1):57–9. [PubMed: 23202435]
122. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol*. 2007; 8(7):R143. [PubMed: 17659080]
123. Huse SM, Welch DM, Morrison HG, Sogin ML. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology*. 2010; 12(7):1889–98. [PubMed: 20236171]
124. Chou H-H, Holmes MH. DNA sequence quality trimming and vector removal. *Bioinformatics*. 2001; 17(12):1093–104. [PubMed: 11751217]
125. Brakenhoff RH, Schoenmakers JG, Lubsen NH. Chimeric cDNA clones: a novel PCR artifact. *Nucleic acids research*. 1991; 19(8):1949. [PubMed: 2030976]

126. Meyerhans A, Vartanian JP, Wain-Hobson S. DNA recombination during PCR. *Nucleic acids research*. 1990; 18(7):1687–91. [PubMed: 2186361]
127. Lee CK, Herbold CW, Polson SW, Wommack KE, Williamson SJ, McDonald IR, et al. Groundtruthing next-gen sequencing for microbial ecology-biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PLoS One*. 2012; 7(9):e44224. [PubMed: 22970184]
128. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res*. 2011; 21(3):494–504. [PubMed: 21212162]
129. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*. 2011; 27(16):2194–200. [PubMed: 21700674]
130. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol*. 2007; 73(16):5261–7. [PubMed: 17586664]
131. He Y, Caporaso JG, Jiang XT, Sheng HF, Huse SM, Rideout JR, et al. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome*. 2015; 3:20. [PubMed: 25995836]
132. Schloss PD, Handelsman J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol*. 2005; 71(3):1501–6. [PubMed: 15746353]
133. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010; 26(19):2460–1. [PubMed: 20709691]
134. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006; 22(13):1658–9. [PubMed: 16731699]
135. Cai Y, Sun Y. ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic acids research*. 2011; 39(14):e95. [PubMed: 21596775]
136. Wang X, Yao J, Sun Y, Mai V. M-pick, a modularity-based method for OTU picking of 16S rRNA sequences. *BMC bioinformatics*. 2013; 14:43. [PubMed: 23387433]
137. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic acids research*. 2014; 42:D633–42. Database issue. [PubMed: 24288368]
138. Bacci G, Bani A, Bazzicalupo M, Ceccherini MT, Galardini M, Nannipieri P, et al. Evaluation of the Performances of Ribosomal Database Project (RDP) Classifier for Taxonomic Assignment of 16S rRNA Metabarcoding Sequences Generated from Illumina-Solexa NGS. *J Genomics*. 2015; 3:36, 9. [PubMed: 25653722]
139. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*. 2007; 35(21):7188–96. [PubMed: 17947321]
140. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl Environ Microbiol*. 2006; 72(7):5069–72. [PubMed: 16820507]
141. Zhang H, Parameswaran P, Badalamenti J, Rittmann BE, Krajmalnik-Brown R. Integrating High-Throughput Pyrosequencing and Quantitative Real-Time PCR to Analyze Complex Microbial Communities. *High-Throughput Next Generation Sequencing*. 2011; 733:107–28. Chapter 8.
142. Liu CM, Aziz M, Kachur S, Hsueh PR, Huang YT, Keim P, et al. BactQuant: An enhanced broad-coverage bacterial quantitative real-time PCR assay. *BMC microbiology*. 2012; 12: 56–2180-12-56.
143. Brukner I, Longtin Y, Oughton M, Forgetta V, Dascal A. Assay for estimating total bacterial load: relative qPCR normalisation of bacterial load with associated clinical implications. *Diagn Microbiol Infect Dis*. 2015; 83(1):1–6. [PubMed: 26008123]
144. Liu CM, Hungate BA, Tobian AA, Ravel J, Prodder JL, Serwadda D, et al. Penile Microbiota and Female Partner Bacterial Vaginosis in Rakai, Uganda. *mBio*. 2015; 6(3):e00589. [PubMed: 26081632]

145. Koren O, Spor A, Felin J, Fak F, Stombaugh J, Tremaroli V, et al. Human oral, gut, and plaque microbiota in patients with atherosclerosis. *Proc Natl Acad Sci U S A*. 2011; 108(Suppl 1):4592–8. [PubMed: 20937873]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Highlights

- Causal research designs may expand the association between human microbiota and health.
- Standardization of laboratory protocols are likely to aid in interpretation of results.
- Clustering by community state types, or other data reduction mechanisms, may reveal links between the microbiota and health outcomes.
- Deposit of data into accessible databases will further our collective efforts.
- Controlling for correlation with repeated measurements and adjusting for known confounders are key statistical issues in longitudinal studies of the human microbiome.

Major Steps	Considerations
Choosing Approach	Considerations of study design Statistical association versus biological causation
Collection/Storage of Samples	Invasiveness of collection method Validation of self-sampling Temperature, duration of sample storage Quality control in labeling, inventory, and database structure
DNA Extraction	Lysing method Eliminating contamination
Amplification/Sequencing	Primer choice: minimize biases and ensure variable region is compatible with body sites Include appropriate controls Platforms have different read lengths and error rates/types
Pre-processing of Reads	Appropriate level of stringency when removing poor quality sequencing reads Removal of chimeras
Taxonomic Classification	De novo or reference based Cutoff level for similarity Choice of reference database may alter classifications
Diversity Measures/ Standardization of Protocol	Choice of whether to rarefy read counts Several metrics for calculating diversity measures Mock communities useful for standardization of protocols
Data Reduction	Clustering and/or principle coordinates analysis Community state types or taxa-level analyses
Statistical Analysis/ Epidemiological Approach	Control for correlation between repeated measurements Adjust for major confounders, assess for interactions Clinical translation

Figure 1.

Sample workflow for a metagenomic study. This figure presents an example of the major steps involved in conducting a 16S rRNA metagenomic study and highlights major points to consider at each step.

Figure 2.

Heatmap showing the distribution of the top 20 bacterial taxa found in the vaginal microbial communities of 3,938 reproductive-age women (vertical lines) from studies conducted in Ravel and Brotman laboratories. Each listed taxa is color coded by relative abundance. Hierarchical clustering of samples was generated by calculating pair-wise Jensen-Shannon distances between all samples and Ward linkage. The dendrogram above the figure directs assignment of the community state types (CSTs) in the pooled samples. Although only the top 20 taxa are displayed, over 265 taxa have been identified in the human vagina [42] and all taxa are included in assessments of CSTs.