Best practices for analyzing microbiomes

Rob Knight[1]*, Alison Vrbanac[2]*, Bryn C. Taylor[2]*, Alexander Aksenov[3], Chris Callewaert[4,5], Justine Debelius[4], Antonio Gonzalez[4], Tomasz Kosciolek[4], Laura-Isobel McCall[3], Daniel McDonald[4], Alexey V. Melnik[3], James T. Morton[4,6], Jose Navas[6], Robert A. Quinn[3], Jon G. Sanders[4], Austin D. Swafford[1], Luke R. Thompson[7,8], Anupriya Tripathi[9], Zhenjiang Z. Xu[4], Jesse R. Zaneveld[10], Qiyun Zhu[4], J. Gregory Caporaso[11] and Pieter C. Dorrestein[1,3,4]

***These authors contributed equally to this work**

Correspondence to R.K. robknight@ucsd.edu

Affiliations

[1]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA 92093 USA

[2]Biomedical Sciences Graduate Program, University of California San Diego, La Jolla, CA USA

[3]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA 92093 USA

[4]Department of Pediatrics, School of Medicine, University of California San Diego, La Jolla, CA 92093 USA

[5]Center for Microbial Ecology and Technology, Ghent University, 9000 Ghent, Belgium

[6]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA USA

[7]Department of Biological Sciences and Northern Gulf Institute, University of Southern Mississippi, Hattiesburg, MS USA

[8]Atlantic Oceanographic and Meteorological Laboratory, National Oceanic and Atmospheric Administration, stationed at Southwest Fisheries Science Center, National Marine Fisheries Service, La Jolla, CA, USA

[9]Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA

[10]Division of Biological Sciences, School of Science Technology Engineering and Math, University of Washington Bothell, Bothell, WA, USA

[11]Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA

For publication in *Nature Reviews Microbiology*

Abstract

Complex microbial communities shape the dynamics of various environments, ranging from the mammalian gastrointestinal tract to the soil. Advances in DNA sequencing technologies and data analysis have provided drastic improvements in microbiome analyses, for example, in taxonomic resolution, false discovery rate control and other properties over earlier methods. In this Review, we discuss the best practices for performing a microbiome study, including experimental design, choice of molecular analysis technology, methods for data analysis and the integration of multiple '-omics' data sets. We focus on recent findings that suggest that operational taxonomic unit-based analyses should be replaced for new methods that are based on exact sequence variants, methods for integrating metagenomic and metabolomic data and issues surrounding compositional data analysis, where advances have been particularly rapid. We note that although some of these approaches are new, it is important to keep sight of the classic issues that arise during experimental design and relate to research reproducibility. We describe how keeping these issues in mind allow researchers to obtain more insight from their microbiome data sets.

Introduction

Advances in DNA sequencing technologies have transformed our capacity to investigate the composition and dynamics of complex microbial communities that inhabit diverse environments from mammalian gastrointestinal tracts to deep ocean sediments. These developments have led to vast increases in the number of microbiome studies being performed in many fields of science, from clinical research to biotechnology. With this transformation, researchers are often left holding massive amounts of data and confronted with a bewildering array of computational tools and methods for analyzing their data. Conducting a robust experiment is not trivial in microbiome research, and as with any study, experimental methods, environmental

factors and analysis methods can impact results. Standards for data collection and analysis are still emerging in the field, yet many compelling results can be achieved with current practices.

Microbiome analysis methods and standards are rapidly advancing. In particular, recommendations concerning differential abundance testing, using exact sequence variants rather than operational taxonomic units (OTUs) and performing correlation analysis have evolved quickly in the past two years. We can expect a similar pace of development in several other areas, including metagenomic taxonomy and functional assignment; integration of data sets from multiple sequencing runs; and further improvement in machine learning, compositional data analysis and multi-omics analyses. However, many of the most fundamental issues that concern microbiome studies arise from statistical and experimental design issues. The most important challenge for the field is to integrate new approaches that are unique to microbiome studies, while remembering standard practices that are broadly applicable to all scientific studies.

Although it is impossible to be fully comprehensive in one article, this Review aims to provide straightforward guidelines for designing, executing and analyzing a microbiome experiment, with particular focus on human, model organism and environmental microbiomes. We direct the reader to more specialized reviews on specific topics where these exist.

Experimental Design

Designing an experiment that generates meaningful data is an important first step in your analysis. Typical scientific questions, such as case-control and longitudinal interventions or studies can all be studied in the context of the microbiome. Researchers can identify potential differences in microbial community structure and composition, genetics, or functional variation either between separate communities or over time. Notably, the general approach to microbiome analysis is applicable regardless of sample origin. However, specific details of the analysis may depend on

the sample origin; for example, 16S ribosomal RNA (rRNA) amplicon regions have variable success among different sample types in recapitulating results from metagenomic sequencing data[1].

The other primary considerations when assessing different sample types are experimental design and sample collection. We have observed many confounding issues during human microbiome studies and therefore we emphasize the importance of experimental design when performing these studies, though often many of the same considerations apply to animal models and environmental samples.

Meticulous experimental design is crucial for obtaining accurate and meaningful results from microbiome studies. Many confounding factors, if not controlled, can obscure patterns in microbiome data (Figure 1). Careful curation of metadata, appropriate controls including extraction and reagent blanks, and thoughtful study designs that isolate and interrogate variables of interest are all essential.

First, the scope of the experiment must be defined, and an appropriate experimental design selected for the question of interest. For example, cross-sectional studies are useful for finding differences in microbial communities between different human populations, such as healthy individuals and those with diseases, or individuals living in different geographic regions. However, due to the large variation in the microbiome between individuals and the profound influence of lifestyle[2,3], diet[4], medication[5,6] and physiology, differences between populations may arise from factors other than the disease of interest. For example, initial reports of changes in the microbiome in diabetic individuals were confounded by effects of the drug metformin[5]. Longitudinal studies, especially prospective longitudinal studies that collect baseline samples before disease onset, can help resolve these issues, although they are more expensive. For ease in downstream statistical

analyses, longitudinal studies should plan the timing of sample collection carefully: for human studies, this may mean collecting samples at identical time points for each subject. Interestingly, community instability rather than the specific taxa present at a single time point can be a strong predictor of disease activity[7]. For example, individuals with inflammatory bowel disease (IBD) exhibit greater microbiome fluctuations than control cohorts[7]. Interventional studies, including double blind randomized control studies, are especially useful for identifying specific effects of a course of treatment on the microbiome and disease state. Designing a study with an analysis plan and specific experimental questions to interrogate can help determine the sample size. For example, to test the effects of a new broad-spectrum antibiotic on the mouse gut microbiota, more samples may be required to look at specific taxa shifts compared to assessing how alpha diversity (a quantitative measure of community diversity) changes with antibiotic treatment, as baseline microbiota composition varies between mice. The antibiotic may be expected to decrease alpha diversity in all mice, but it could perturb their microbial community composition in different ways. For any study design, appropriate methods to assess statistical power should be employed in order to discern technical variability and real biological results[8]. However, statistical power and effect size analysis remains a challenge in microbiome research[9]. Some methods that are currently used for power and effect size analysis are based on PERMANOVA[8], Dirichlet Multinomial[10] or random forest analysis[11]. As these methods are further developed to integrate metagenomics, metatranscriptomics, metaproteomics and metabolomics data sets, study design and selection of appropriate sample size will also improve. For specific experimental design considerations, we recommend reviewing the design of other successful studies with similar sample types and desired outcomes. We expand on important considerations for microbiome experimental design below.

Defining controls and exclusion criteria

Defining clear inclusion and exclusion criteria limits confounding covariates. For instance, variability in recovery time from antibiotics among individuals[12] suggests that individuals that were treated with antibiotics in the preceding 6 months should be excluded from most microbiome studies. Similarly, recovery of the skin microbiome after hand washing takes ~2 hours[13].

In case-control experimental designs, controls must be appropriately selected and matched. Age and sex are common control criteria, despite the relatively weak effect of sex on most human microbiomes across body sites[14,15], while other variables such as medication and diet are often more important confounders to control for. The relative effect sizes of these microbiome variables are still emerging[9]. Collection of comprehensive clinical data collection is crucial for identifying confounders that cannot be controlled. This topic has been extensively reviewed in Ref. 16. Environmental studies must also account for similar confounders, as plot-to-plot variation is a widely recognized confounding phenomenon in the ecological literature that should be addressed with nested statistical tests[17].

Animal models

The predominant animal models for studying the microbiome are rodents, such as mice. Other models with varying microbial complexity such as bobtail squid, insects or zebrafish are often useful for studying specific interactions between host and microorganisms (for example, how the microbiome and the host genetics influence each other)[18]. Nevertheless, rodents are often preferred because they are well-characterized and have many physiological similarities to humans. Rodent microbiome studies require particularly careful design. As rodents are coprophagic, cagemate fecal microbiomes become more homogenous over time, so experiments must be replicated across multiple cages to control for cage effects[19]. Parental effects also necessitate

randomizing littermates between cages and allowing for normalization. Single-housing stresses mice[20], and is thus often technically or ethically infeasible. Even genetically identical rodents may differ in their microbiomes due to environmental factors including diet, litter, vendor, shipment and facility[21,22]. Additionally, early life microbial exposures greatly impact the established microbiota and can influence immune system development[23]. Similar considerations apply to other co-housed model organisms, for example, zebrafish[24].

Technical variation

Technical variability among experimental methods ranging from DNA extraction to sequencing is high[25,26]. The same reagent kits must be used for all samples in a study[27], and multiple baseline samples should be collected to assess intrinsic variability among time points in longitudinal studies. Using blanks during sampling, DNA extraction, PCR and sequencing is essential for detecting contamination. Reads that are derived from microorganisms introduced as contaminants or that grow during shipping can sometimes be reduced during analysis[28], though samples should be at -80 °C when possible[29]. For field studies or other situations where freezing is not possible, ambient storage methods, such as 95% ethanol or commercial products such as RNAlater or the OMNIgene Gut kit can be used[30]. Mock communities (reference samples with a known composition) are useful for standardizing analyses[31], as is including the same standard specimens in each DNA sequencing run[32]. In general, reconciling microbiome data that were generated using different methods remains an unsolved challenge.

Depending on the scope of their experiment (which includes the overall experimental design, sample types and source, sequencing method, and other factors that are discussed below), researchers can aim to gain a broad, community-level overview of their samples, a detailed

genomic-level understanding, or even characterization of the functional variation in microbial communities.

Sequencing Targets and Methods

Different methods for surveying microbial communities, including marker gene, metagenome, and metatransciptome sequencing, can produce varying results. All widely-used methods have strengths and weaknesses, so the question, hypothesis, sample type and analysis goals should inform the choice of method (Table 1). Here we discuss the trade-offs between cost, robustness, resolution and difficulty for marker gene, metagenome and metatranscriptome sequencing. We outline the best workflow for each method in Figure 2. To attain a high-level, but low resolution overview, the preferred method is marker gene sequencing. Metagenomic sequencing provides more detail by analyzing the total DNA in a sample, allowing strain-level resolution and detection of genes that can provide information on molecular functions. We also discuss metatranscriptomic sequencing of total RNA, which is used to characterize gene expression in the microbial community.

Marker gene analysis

Marker gene sequencing uses primers that target a specific region of a gene of interest in order to determine microbial phylogenies of a sample. This region typically contains a highly variable region that can be used for detailed identification, flanked by highly conserved regions that can serve as binding sites for PCR primers. Marker gene amplification and sequencing (such as 16S rRNA for bacteria and archaea and internal transcribed spacer (ITS) for fungi) is a well-tested, fast and cost-effective method for obtaining a low-resolution view of a microbial community. This approach works well for host DNA contaminated samples, such as tissue and

low-biomass samples. However, because DNA sequences vary in these primer-amplified regions, primers do not have equal affinity for all possible DNA sequences, and consequently induce bias during PCR amplification. Other sources of inherent bias in marker gene sequencing include variable region selection, amplicon size[33], and the number of PCR cycles[34]. Low-biomass samples are particularly susceptible to bias introduced by over amplification–as the PCR cycle number increases, contaminating microorganisms are increasingly over-represented[35]. Optimizing primer selection can help mitigate bias, but this requires a priori knowledge of microbial community composition to assess taxonomic resolution and coverage of the target community[36]. However, even well-optimized primers are often limited to genus level taxonomic resolution. Marker gene sequencing generally correlates well with genomic content[37,38,39,40,41] and is applicable to the broadest range of sample types and study designs.

Whole metagenome analysis

Metagenomics is the method of sequencing all microbial genomes within a sample. Metagenomic sequencing yields more detailed genomic information and taxonomic resolution than marker gene sequencing alone, but it is relatively expensive to prepare, sequence and analyze the samples. This method captures all DNA present in the sample, including viral and eukaryotic DNA. Given adequate sequencing depth (the number of sequencing reads per sample), taxonomic resolution to species or strain level[42] and the assembly of whole microbial genomes from short DNA sequence reads is possible[43]. However, de novo annotation of functional genes is not possible in such settings. Metagenomic sequencing profiles the functional capacity of an entire community at the gene level[44], moving well beyond the limits of marker gene analysis. However, biases that are introduced by library construction, assembly and reference databases for annotation are less understood than biases that exist in well-characterized marker gene approaches. As the

metagenomics field matures, these annotation steps will continue to be improved and validated. For a comprehensive review on metagenomics, we direct the reader to Ref. 45.

Metatranscriptome analysis

Metatranscriptomics uses RNA sequencing to profile transcription in microbiomes, providing information on gene expression and the active functional output of the microbiome. Metatranscriptomics differs from both marker gene and metagenomic sequencing that sequence DNA in a sample, regardless of cell viability or activity. Although there are methods for depleting relic DNA from dead cells[46], sequencing microbial RNA provides better insight into the functional activity of a microbial community, though it is biased towards organisms with higher rates of transcription. It is worth noting that propidium monoazide (PMA) depletion of relic DNA is an alternative method to identify live microorganisms[47]. Host RNA contamination, particularly the highly abundant rRNAs, is also an important consideration and methods to exclude rRNAs from samples should be considered[48]. RNA must be carefully preserved to avoid degradation in all cases, though certain sample types may warrant specialized protocols for RNA purification. For example, soil samples require removal of enzyme-inhibiting humic substances[49,50]. Despite these technical difficulties, metatranscriptomic data can offer unique insight; transcriptomes vary more within individuals than metagenomes[51], and metatranscriptomics can reveal microbial community response to perturbations, such as xenobiotic exposure[52]. For a comprehensive review on metatranscriptomics analysis of the microbiome, we direct the reader to Ref. 53.

Analyses

Ideally, each microbiome study would analyze samples with all three of the methods discussed above. In most cases, however, there is not enough sample material or enough project

funding for performing all three analyses, and in some cases, the samples might not be amenable to one of the sequencing methods. It is therefore paramount that the researcher chooses the method of sequencing that is most effective for answering their specific questions. If there are no budget constraints, we recommend performing metagenomics rather than marker gene sequencing. However, it is common practice to perform marker gene sequencing to gain a low resolution understanding of the microbial community composition. Next, depending on the focus of the study, the researcher can move on to metagenomic and metatranscriptomic sequencing, though this may require a second study for appropriate sample collection and processing.

Marker gene analyses

As noted above, marker gene approaches are sensitive to technical factors such as primer choice[54], so well-validated protocols such as those used with the diverse sample set in the Earth Microbiome Project should be used[55]. The first step in analyzing marker gene amplicon data is to remove sequencing errors: despite very low sequencing error rates (for example, in Illumina sequencing ~0.1% per nucleotide[56]), most of the apparent sequence diversity arises from sequencing errors[57,58]. Until recently, this problem was addressed by clustering similar sequences into OTUs[59,60]. Clustering sequences into OTUs, termed OTU picking, consolidates similar sequences (usually with a 97% similarity threshold) into single features, merging sequence variants including those introduced by sequence error into a single OTU that can be used in subsequent analysis. However, this method misses subtle and real biological sequence variation, such as single nucleotide polymorphisms (SNPs) that would be consolidated into single OTUs[61]. Oligotyping[62] improves upon traditional OTU picking by including position-specific information from 16S rRNA sequencing to identify subtle nucleotide variation and by discriminating between closely related but distinct taxa. Algorithms such as Deblur[63] and DADA2[64] use error profiles to resolve

sequence data into exact-sequence features (the marker gene sequence) called "sub-OTUs" (sOTUs). The resulting output from these methods is a table of DNA sequences and counts of these different sequences per sample rather than OTU groups. We recommend that these methods replace OTU-based approaches for all applications, except when it is necessary to combine sequence data that were generated using different technologies (that is, Illumina sequencing and 454 pyrosequencing) or with different primer sets, when mapping to a common reference database of full-length sequences is often still needed[65].

One key analysis step is to assign taxonomic names to microbial sequences in the data. Taxonomy is typically assigned by machine learning approaches such as the RDP classifier[66], which uses Naive Bayes models that are trained on oligonucleotide frequencies at the genus level to achieve ~80% accuracy in genus-level assignments. Popular microbiome analysis packages such as QIIME[59] and Mothur[60] provide support for taxonomic classification. In principle, exact matching to reference databases (three of the most characterized and frequently used are Greengenes, RDP, and Silva) should provide better specificity in taxonomic assignment, but the sensitivity of this approach is poor given the large number of unknown taxa. Furthermore, de novo phylogenetic trees that are constructed from short marker gene sequences are typically poorly resolved, so insertion of marker gene sequences into a characterized reference tree that is based on full-length sequences[67] is desirable, given the importance of phylogenetic metrics[68]. "Unclassified" microorganisms should be checked for organelle sequences, and for many studies, chloroplast and mitochondria sequences should be excluded before proceeding with analysis (although for intestinal samples, these sequences can be useful for identifying consumed foods and thus should not be disregarded completely).

Predictive functional profiling[38,39,40,41] is a technique for linking marker gene studies with available microbial genomes to make predictions about metagenomic content and thus the putative biological functions of a microbial community. This analysis generally requires a reference-based OTU table. Methods based on evolutionary models (for example, PICRUSt[39]) provide confidence intervals on these predictions of gene content, which will tend to be wider in regions of the tree distant from reference genome sequences, and narrower where many reference genomes are available. Thus, the availability of sufficient closely related reference genomes is a main factor that influences the accuracy of these results. Another limitation for predictive functional profiling is that some families of bacteria possess a very similar 16S rRNA variable region, despite being phenotypically and genotypically divergent.

Most statistical analyses that are applied to microbiome data that is generated from marker gene sequencing can also be applied to other types of "-omics" analyses, and are described below in the "Higher-level analyses" section.

Metagenome and metatranscriptome analyses

Surveying the complete nucleic acid profile of a sample yields rich information that can be used to investigate a broad range of taxonomic, functional, and evolutionary aspects of microbial communities—even contaminants can provide important details[69]. As with marker gene-based surveys, the analytical methods must be carefully chosen to consider the sample origin and the specific hypotheses under investigation. Here, we discuss the best approaches to perform these analyses.

Read-based profiling takes the unassembled DNA or mRNA sequence reads and compares them against reference databases to assign taxonomy or annotate genes. With the ever-increasing size of modern query datasets and databases, methods are continually being refined to improve the

speed of read-based profiling. Many tools utilize k-mers, assigning taxonomy to short DNA fragments of length "k," such as Kraken[70] or employ the Burrows-Wheeler transform which compresses the database by merging similar sequences (for example, Bowtie2[71] and Centrifuge[72]). For a more comprehensive guide to tool selection, we direct the reader to Ref. 73. 'Marker gene' methods (such as MetaPhlAn2[74] and TIPP[75]) use specific genomic regions for taxonomy assignment, focusing on universal, single-copy elements. Beyond taxonomy assignment, others tools such as HUMAnN244 can also be used for annotating genes and metabolic pathways. Some tools, including MEGAN[76], incorporate both of these functionalities, and can be a preferred method when both annotations are desired. Because each read is considered independently, read-based methods scale efficiently to large, complex data sets, such as soil microbiome data sets. It is important to note that as taxonomic or functional assignment depends on homology between the single read and a reference, database choice is crucial. For well-characterized environments like the human gut, curated genome databases such as RefSeq[77] and protein family databases like Pfam[78] or UniRef[79] increase the accuracy of results and decrease computational costs. For samples from poorly characterized environments, the use of large databases such as NCBI nr and nt and IMG/MG[80] should be considered because the databases are larger, despite the increased computational complexity and decreased assignment specificity. Specialized databases must be used to annotate specific taxonomic or functional categories, such as PHASTER[81] for bacteriophages, Resfams[82] for antibiotic resistance genes and FOAM for environmental samples[83]. Additionally, numerous metagenomic data catalogues are available for many sample types, including Tara for ocean samples[84], the BGI catalogue for mouse gut samples[85] and MetaHit for human gut samples[86].

Another method for analyzing metagenome and metatranscriptome sequencing reads is to assemble the short reads into longer sequences (contigs). These contigs can be further sorted or binned by similarity to assemble partial to full genomes of microorganisms. This allows data exploration beyond taxa and gene annotation, enabling the prediction of multi-gene biosynthetic pathways or even metabolic reconstructions with tools such as antiSMASH[87]. However, assembly-based analyses are not universally applicable; higher biodiversity, the presence of many related strains in samples or low coverage yields fragmented assemblies and can obscure taxa from downstream analyses. For example, soil samples are often difficult to assemble due to the high microbial diversity and uneven distribution[88]. For samples that avoid these complications, metagenome assemblies provide valuable bespoke reference databases for read-based and assembly-based metatranscriptome analyses[89,90], thus recovering the 'microbial dark matter' that is absent in curated databases[91]. Recommended tools for assembly-based analyses include metaSPAdes[92] and MEGAHIT[93]. A comprehensive discussion of these and other tools can be found in Ref. 94. To assemble partial to full genomes of individual microorganisms, contigs are sorted (binned) into separate putative genomes with tools such as MaxBin2[95] and CONCOCT[96], which evaluate nucleotide composition and abundance patterns across samples to perform sorting (binning). To evaluate the quality of these binned and assembled genomes, single-copy gene profiling tools such as CheckM[97] that use common single-copy genes to estimate genome completeness and contamination can be used. Additionally, visualization tools like VizBin[98] display clustering of metagenomic sequences without alignment to a reference database, allowing researchers to visually inspect the sequence clustering of related organisms and assist with evaluating bin quality. Employing integrated workflow tools to automate data processing such as

Anvi'o[99], ATLAS[100], or MetAMOS[101], is highly recommended because assembly-based methods are complex.

In order to compare samples with varying sequencing read counts, various methods of normalization can be employed. Common methods of normalization include: read counts per million (counts are scaled by the total number of reads), transcripts per kilobase million (counts scaled by number of reads and length of reads), and converting the data to relative abundance. Additionally, there are various tools for performing normalization including edgeR[102] and DESeq2[103].

New tools for both read-based and assembly-based approaches are under rapid development. When possible, specific analytical decisions should be made based on performance on well-studied or synthetic datasets (such as the Critical Assessment of Metagenomic Information[104]) that are most similar to the microbial community of interest.

Higher-Level Analyses

Processing microbiome data generates a matrix that relates feature abundance (taxa or genes) to samples. This output is deceptively simple; microbiome data is highly dimensional, often representing thousands of different taxa, and sparse with many zeros present in the matrix, requiring careful statistical treatment to extract meaningful results.

Overall patterns in microbiome variation are typically assessed by alpha and beta diversity. Alpha diversity quantifies feature diversity within individual samples and can be compared across sample groups. For example, when comparing a sample from an individual with a disease to a healthy control, the researcher can use alpha diversity to compare the mean species diversity between the two samples. Measures of species richness (for example, the number of observed species, or Chao1 abundance estimator, which estimates true species diversity) and phylogenetic

measures (Faith's phylogenetic diversity) are sensitive to the number of sequences per sample, whereas measures that combine richness and evenness (Shannon index) are much less so. However, it should be noted that these methods have been evaluated exclusively for 16S rRNA data, and may not apply to other microbiome data types. Beta diversity compares feature dissimilarity between each pair of samples, generating a distance matrix of beta diversity distances between all pairs of samples. Metric selection can influence the results obtained[105,106] and should be chosen with biological data interpretation in mind. Quantitative metrics (Bray-Curtis, Canberra and weighted UniFrac) use feature abundance data in calculations whereas qualitative metrics (binary-Jaccard and unweighted UniFrac) only consider the presence or absence of features. Phylogenetic measures such as UniFrac typically provide interpretable biological patterns[107], though these metrics require a phylogenetic tree and thus cannot be used for direct comparison with "–omics" data that lack trees. Software for performing alpha and beta diversity calculations includes QIIME[59], Mothur[60], and the R package Vegan[108]. The non-parametric permutation tests PERMANOVA and ANOSIM are used for assessing significant beta diversity clustering between groups, but PERMANOVA may perform better on datasets with varying dispersions within groups[109]. Calculation of meaningful alpha and beta diversity measures requires the researcher to control for the sampling effort (that is, the number of sequences per sample obtained), as this can differ by orders of magnitude. The current best solution for UniFrac is rarefaction[110], though for the special case of pairwise differential abundance testing, the full sample set should be used[111].

For visualizing beta diversity data, ordination techniques, such as principal coordinates analysis (PCoA) or principal component analysis (PCA), are commonly used. These methods reduce large and complex distance matrices into a visually manageable two dimensional or three-dimensional representations of sample distances. Samples can then be colored by various metadata

categories to visualize clustering in an unsupervised manner. EMPeror offers an interactive framework for manipulating PCoA plots[112].

Another common analysis approach is to look at differentially abundant microorganisms or functional elements (for example, genes and pathways) in the comparison groups of interest (that is, treatment versus control). Identifying microbial taxa that explain differences between communities is particularly challenging because microbiome data sets are high-dimensional (that is, they include thousands of taxa), sparse and compositional. Compositionality is the crux of the problem[113]; when the proportion of one microorganism increases, the proportions of others must decrease for the proportions to sum to 1. For example, suppose a patient is administered a drug that increases the growth rate in only a single microbial genus, while not affecting the growth of others. Although the other microorganisms are not impacted by the drug, they would have decreased in relative abundance due to the outgrowth of the single microbial genus. This poses challenges for many classical methods, such as parametric statistical tests (for example, Student's t-test and ANOVA), and measures of correlation including Spearman's rank correlation, often leading to completely unacceptable false discovery rates above 90%[110,114,115]. Recently, "compositionally-aware" methods have addressed this problem of compositionality and relative abundance. One approach is to force strong biological assumptions on the statistical test: for example, Lovell's proportionality metric detects only positive correlations[116]. Other tools that are widely applicable and have been optimized for microbiome data, such as SparCC[117] and SPEIC-EASI[118], assume that few species are correlated, so most correlation coefficients are zero. BAnOCC[119] is another tool for addressing the compositionality problem that makes no assumptions about the data. We recommend another approach that does not assume few species are correlated, which is to test for differences between microbial communities using the isometric

log ratio transform (ilr) [*sic*]. The isometric log ratio transform approach controls for false positives due to proportionality by testing for the changes in log ratios between microbial abundances, commonly referred to as balances. Balances can be constructed using prior knowledge such as evolutionary history[107,120,121] or microbial niche differentiation in response to environmental factors such as pH[122]. After the ilr transform is applied, standard statistical tools such as multivariate response, linear regression and classification can effectively test for differences on the balances or log ratios between microorganisms rather than the raw microbial abundances, controlling for compositionally. Other recent methods use absolute quantification to address compositionality by complementing sequencing with microbial cell counts in each sample[123,124].

Machine learning is emerging as an especially useful technique for determining how microbiome data can be used to separate samples based on current state (usually determined by metadata categories, such as 'healthy state' versus 'diseased state')[125,126] or, excitingly, to predict future state[127,128]. For instance, it is possible to model the severity and susceptibility of gingivitis based on an individuals' oral microbiota[127]. Random Forests regression, a machine learning technique, has been effective in many applications, ranging from dating time since death of a corpse[129] to providing a model for determining microbiome maturation in child development[130]. SourceTracker[131], a Bayesian estimator of the microbial sources that make up an unknown community, is useful for classifying microbial samples according to environment of origin[132]. Importantly, machine learning analyses need a substantial sample size and should always be coupled with cross-validation, independent test sets, or other experimental and biological confirmation to ensure robust findings.

Integrating Other 'Omics' Data

Knowing the composition of a microbial community is no longer a sufficient research goal; we want to know the function of the community. Integrating other data types—including marker gene sequencing, metagenomics, metatranscriptomics, metaproteomics, metabolomics and other techniques—for a given study is crucial for a comprehensive understanding of the composition and function of microbial communities. For example, changes in the metabolite profile of a microbial community reflect changes in its biosynthetic activity, mRNA and protein expression, and protein activity[133]. "Multi-omics" analysis integrates chemical and biological knowledge to provide a more complete picture of a biological system and is an active area of research with largely untested methods (Figure 3).

Integrating multi-omics data types is inherently difficult. For example, gene expression and metabolism operate on different timescales[134], and microorganisms produce many metabolites, often only in response to molecular signals from other species[135]. Also, the sparse nature of metagenomic and metabolomic data (where the data matrices are composed mostly of zeros) is much greater than of metaproteomic data and this may pose technical problems for some methods. Although the integration of different "-omics" data sets is a work in progress, tools that integrate these datasets are becoming increasingly available. For example, XCMS Online integrates metabolomic data with metabolic pathways, as well as transcriptomic and proteomic data[136]. Traditional correlation methods such as Pearson and Spearman could enable pairwise correlation between features across "-omics" data sets. However, these are prone to false positives due to the sparsity and high-dimensionality of microbiome and metabolome datasets. Procrustes analysis[137] uses dimensionally-reduced data to test if patterns (distances) between samples in one dataset is observed in the other, essentially correlating ordination spaces rather than individual features (tested using Mantel[138] or PROcrustes randomization TEST). Other methods integrate "-omics"

datasets by not only taking into account the relationships between samples, but also associating samples to particular metadata categories of interest (such as examining healthy versus diseased or control versus treatment groups). These methods include co-inertia analysis, which uses dimensionality reduction to associate sample patterns in two data sets and relevant metadata[139], and partial least-squares[140], related methods such as canonical correlation analysis[141], or robust sparse canonical correlation analysis, which is a variation of the method to deal with sparse "-omics" data[142].

Advanced integrative analysis tools include molecular networking with GNPS[143] to identify metabolites and pathway annotations[144], and general systems biology tools, exemplified by XCMS Online[136]. Increasingly, multi-omics studies are investigating temporal patterns in addition to spatial patterns. Spatial mapping[145], that can now be performed with the tool 'ili[145], adds a powerful dimension to multi-omics studies through visual representations that are readily amenable to human interpretation.

Integration with other "-omics" data can be performed using various statistical methodologies[146]. However, these techniques have been shown to perform suboptimally on microbiome data sets[115]. Furthermore, simply finding correlations in various "-omics" data by itself is only the first step. Establishing causation and correlation across data sets is the next challenge. Box 1 gives an example of the integration of metabolome and microbiome data sets and corresponding approaches to move beyond correlation and determine causation. Correction for multiple comparisons is crucial in multi-omic analyses; data sets can contain thousands of different microorganisms and metabolites, so significant correlations are expected by random chance. Measures to correct significance testing for multiple comparisons include the False Discovery Rate (for example, Benjamini-Hochberg correction) or, for more conservative corrections, the Family-

Wise Error (for example, Bonferroni correction). Using these methods to penalize multiple comparisons in conjunction with statistical models that incorporate sparsity and compositionality[115], false discovery rates in large multi-omic comparisons can be reduced.

Despite these challenges, the future potential for "-omics" data integration is promising. In particular, there are numerous examples where metagenome, metatranscriptome and metabolome data have been successfully integrated, illuminating gene regulation in microbiomes[37] and correlating the presence of microorganisms with metabolites[147]. Such studies have provided insights beyond the capacity of single –omics, such as gut bacterial metabolism of xenobiotics[148] and how antibiotic-induced microbiome depletion creates a favorable metabolomic environment for *Clostridium difficile*[149]. Comparatively, the integration of metaproteomics data with microbiome data is a relatively newer field of investigation, though there are many recent examples of successful integration ranging from identifying biomarkers of Crohn's disease[150] to examining microbial protein production in layers of permafrost[151]. Additionally, tool development for metaproteomics annotations and analysis is ongoing[152,153]. Overall, integrating '-omics' data can provide a more holistic and mechanistic understanding of microbiomes—from DNA identification to functional production of metabolites and proteins—and ideally lead to more actionable scientific insights.

Conclusions

In this Review, we have discussed how all stages of conducting a microbiome study, from designing the experiment to collecting and storing the samples, to obtaining insight from graphical displays of the sequence data, can substantially impact the results and their biological interpretation. As the effects of many of these technical steps are large compared to the real biological variability to be explained, standardization is necessary in order to compare and

combine separate studies, and the first efforts to do this and to provide recommendations and best practices, such as the International Human Microbiome Standards and the Microbiome Quality Control Project (MBQC), are already under way. Including bioinformatics pipelines and controls into these standardization efforts, and in particular using cloud-enabled reproducible computing resources that run open-source code on publicly available data to reproduce scientific claims of publications, is a rapidly emerging area that will bring consistency and comparability to the microbiome field. An important part of such efforts will be spike-in standards (which have already been so important to standardizing microarrays), and standardized biologically realistic samples that can be used to quantify systems-level accuracy in microbiome assays.

This article has focused primarily on DNA-level analyses at the whole-community level, but as expression-level profiling and single-cell profiling techniques continue to advance, many similar considerations will apply to those types of data also. Avoiding the mistakes that have been repeated frequently in other expensive assays, such as inadequate sample size and validation, and employing best practices for standards, sample handling, compositional data analysis and other frequent pitfalls, will accelerate progress in these areas. Using standardized and well-characterized sample sets, such as those developed in MBQC and in the Earth Microbiome Project, can greatly shorten the time needed to understand the value and unique insights provided by a new technique.

As the field trends towards ever-larger data sets, understanding subtle confounding factors long known to epidemiologists and taking more care with longitudinal study designs will become increasingly important. The value of interventional studies over observational studies is considerable, especially when human, animal model and *in vitro* data can be correlated across scales and systems. Increased standardization of techniques and dissemination of methods with

low noise and bias will greatly increase the ability of the microbiome field to deliver on the promise of translatability from lab-scale studies to the clinic, field or natural environment.

Acknowledgements

This review is informed by our work funded by the NIH, NSF, Alfred P. Sloan Foundation, John Templeton Foundation, W. M. Keck Foundation, and hundreds of collaborators on the Human Microbiome Project, American Gut Project, and Earth Microbiome Project.

Competing Interests

The authors declare no competing interests.

Publisher's Note

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

Author Contributions

A.V. and B.C.T. researched the data for the article. A.G., T.K., D.M., J.N., J.G.S. and J. R. Z. substantially contributed to discussion of content. R.K., A.V., B.C.T., A.A., C.C., J.D., L.M., A.V.M., J.T.M., R.Q., L.R.T., A.T., Z.Z.X,, Q.Z. and J.G.C. wrote the article. R.K., A.V., B.C.T., T.K., D.., A.D.S. and P.C.D. reviewed and edited the manuscript before submission.

Table of Contents Blurb

Complex microbial communities shape the dynamics of various environments, ranging from the mammalian gastrointestinal tract to the soil. In this Review, Knight and colleagues discuss the best practices for performing a microbiome study, including experimental design, choice of molecular analysis technology, methods for data analysis and the integration of multiple “-omics” data sets.

Glossary

Alpha Diversity: A measure of within sample diversity

Beta Diversity: A measure of similarity between samples

Coprophagy (coprophagic): The consumption of feces. Many animal species eat feces to more efficiently break down plant matter by digesting the material twice.

Effect size analysis: Quantification of the magnitude of an effect of a particular metadata category (treatment group, sex, sequencing plate) on the data.

Exact sequence variants: For marker gene sequencing, the exact DNA sequence for each read is used instead of OTU clustering.

Faith's Phylogenetic Diversity: An alpha diversity metric which uses a phylogenetic tree to compute sample diversity

False Discovery Rate: A method of understanding the rate of type I errors in null hypothesis testing when performing multiple comparisons

Family-Wise Error: The probability of making one or more Type I errors (false discoveries) when performing multiple hypotheses tests

Humic substances: Produced by biodegrading organic matter, humic substances are the main component of humus (soil).

Isometric Log Ratio Transform: The isometric log ratio (ilr) transform converts a vector of proportions into a vector of log ratios using a tree as a reference. The computed log ratios consist of the difference of mean logarithms of species proportions between adjacent clades within the tree.

k-mers: All possible sequences of length 'k' from a read obtained through DNA sequencing.

Machine Learning: The use of algorithms to learn from and make predictions about data.

Marker genes: Conserved genes (commonly 16S, ITS, and 18S) that typically contain a highly variable region that can be used for detailed identification, flanked by highly conserved regions that can serve as binding sites for PCR primers.

Metadata: Information about the data. In many studies, this is structured as a matrix with samples as rows and metadata categories (Age, Sex, Longitude, Season, Disease-State, Average Monthly Rainfall, etc.) as columns.

Metagenomes: The collection of genetic material from a community of organisms; for example, the genetic material from all microorganisms in the human gut microbiome.

Metatranscriptomes: The total content of gene transcripts from a community of organisms.

Naive Bayes Classifier: A simple probabilistic classifier used in machine learning that is based on applying Bayes' theorem assuming strong independence between the features

Nested statistical tests: Statistical tests that address variables related to the main effect. For example, soil plot would be a nested factor for testing the effects of a fertilizer on the soil microbiota.

Operational Taxonomic Units (OTUs): A group of closely related individuals or sequences (often 97% sequence similarity threshold)

Random Forests regression: A machine learning technique that uses decision trees to perform classification

Reads: Inferred sequences of base pairs in a single DNA fragment.

Shannon Index: A commonly used index to characterize species diversity in a community.

Highlighted References

1. The Human Microbiome Project Consortium. Structure, Function and Diversity of the Healthy Human Microbiome. *Nature*. **486**, 207–214 (2012).

   Significance: First large-scale effort to characterize the healthy human microbiota and commonly used reference database.

2. Qin, J., *et al*. A human gut microbial gene catalog established by metagenomic sequencing. *Nature*. **464**, 59–65 (2010).

   Significance: First large-scale effort to catalog microbial genomes in the human gut with shotgun metagenomic sequencing.

3. Thompson, L. *et al*. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457-453 (2017).

   Significance: Development and implementation of standardized protocols and new analytical methods that enabled a massive comparison of over 100 studies to characterize the Earth's microbial diversity.

4. Caporaso, J. G. *et al*. QIIME Allows Analysis of High-Throughput Community Sequencing Data. *Nat. Methods*. **7**, 335–336 (2010).

   Significance: Widely used software package for microbiome analysis.

5. Schloss, P. D. *et al*. Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).

   Significance: Widely used software package for microbiome analysis.

6. Quince, C., Walker, A.W., Simpson, J.T., Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*. **35**, 833-844 (2017).

   Significance: Comprehensive review on using shotgun metagenomics.

7. Hamady, M., Lozupone, C. & Knight, R. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J*. **4**, 17–27 (2010).

   Significance: Underscores the power incorporating phylogenetic information when comparing microbial communities

8. Forslund, K. *et al*. Disentangling the Effects of Type 2 Diabetes and Metformin on the Human Gut Microbiota. *Nature*. **528**, 262–266 (2015).

   Significance: Excellent example of how study design and metadata collection can influence experimental results.

9. Theriot, C.M., Koenigsknecht, M.J., Carlson, P.E., *et al*. Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection. *Nat. Commun.*, **5**:3114 (2014).

Significance: Great example of 'omics data integration (microbiome and metabolome)

10. Subramanian, S. *et al*. Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature*. **510**, 417–21 (2014).

Significance: Demonstrates power of machine learning with microbiome data by developing a microbiota maturity index.

References

1. Meisel J.S., Hannigan G.D., Tyldsley A.S., *et al*. Skin microbiome surveys are strongly influenced by experimental design. *The Journal of investigative dermatology*. **136**(5):947-956. (2016).

2. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science*. **29**, 560-564 (2016).

3. Noguera-Julian, M. *et al*. Gut Microbiota Linked to Sexual Preference and HIV Infection. *EBioMedicine*. **5**, 135–146 (2016).

4. Wu, Gary D. *et al.* Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science*. **334**, 105–108 (2011).

5. Forslund, K. *et al*. Disentangling the Effects of Type 2 Diabetes and Metformin on the Human Gut Microbiota. *Nature*. **528**, 262–266 (2015).

6. Jackson, M. A. *et al*. Proton Pump Inhibitors Alter the Composition of the Gut Microbiota. *Gut*. **65**, 749–756 (2016).

7. Halfvarson, J. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol*. **2**, 17004 (2017).

8. Kelly, B. J. *et al.* Power and Sample-Size Estimation for Microbiome Studies Using Pairwise Distances and PERMANOVA. *Bioinformatics*. **31**, 2461–2468 (2015).

9. Debelius J., Song S.J., Vazquez-Baeza Y., Xu Z.Z., Gonzalez A., Knight R. Tiny microbes, enormous impacts: what matters in gut microbiome studies? *Genome Biology*. **17**:217 (2016).

10. La Rosa PS, Brooks JP, Deych E, *et al*. Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data. White EP, ed. *PLoS ONE*. **7**(12):e52078 (2012).

11. Knights, D., Costello, E. K. and Knight, R. Supervised classification of human microbiota. *FEMS Microbiology Reviews*. **35**: 343–359 (2011).

12. Dethlefsen, L. & Relman, D. A. Incomplete Recovery and Individualized Responses of the Human Distal Gut Microbiota to Repeated Antibiotic Perturbation. *Proc. Natl Acad. Sci. USA* **108**, 4554–4561 (2011).

13. Fierer, N., Hamady, M., Lauber, C. L., & Knight, R. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc. Natl Acad. Sci. USA* **105**, 17994–17999 (2008).

14. Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., & Knight, R. Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science*. **326**, 1694–1697 (2009).

15. The Human Microbiome Project Consortium. Structure, Function and Diversity of the Healthy Human Microbiome. *Nature*. **486**, 207–214 (2012).

16. McDonald, D., Birmingham, A., & Knight, R. Context and the human microbiome. *Microbiome*. **3**, 52 (2015).

17. Ramette, A. Multivariate analyses in microbial ecology. *FEMS Microbiol. Ecol*. **62**, 142–160 (2007).

18. Kostic, A. D., Howitt, M. R. & Garrett, W. S. Exploring host-microbiota interactions in animal models and humans. *Genes Dev*. **27**, 701–718 (2013).

19. Ridaura, V. K. *et al.* Cultured Gut Microbiota from Twins Discordant for Obesity Modulate Adiposity and Metabolic Phenotypes in Mice. *Science*. **341**, 6150 (2013).

20. Reber, S. O. *et al.* Immunization with a Heat-Killed Preparation of the Environmental Bacterium Mycobacterium Vaccae Promotes Stress Resilience in Mice. *Proc. Natl Acad. Sci. USA* **113**, E3130–E3139 (2016).

21. Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., & Gordon, J. I. Obesity alters gut microbial ecology. *Proc. Natl Acad. Sci. USA* **102**, 11070–11075 (2005).

22. Friswell, M. K. *et al.* Site and Strain-Specific Variation in Gut Microbiota Profiles and Metabolism in Experimental Mice. *PLoS ONE*. **5**, e8584 (2010).

23. Snijders, A. M., Langley, S. A., Kim, Y., *et al*. Influence of early life exposure, host genetics and diet on the mouse gut microbiome and metabolome. *Nature Microbiology*. **2**, 16221 (2016).

24. Stagaman, K., Burns, A. R., Guillemin, K. & Bohannan, B. J. The role of adaptive immunity as an ecological filter on the gut microbiota in zebrafish. *ISME J*. **11**, 1630-1639 (2017).

25. 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. *Nature Biotechnology*. **35**, 1077-1086 (2017).

26. Costea, P.I., Zeller, G. Sunagawa, S., *et al*. Towards standards for human fecal sample processing in metagenomic studies. *Nature Biotechnology*. **35**, 1069-1076 (2017).

27. Salter, S. J. *et al*. Reagent and Laboratory Contamination Can Critically Impact Sequence-Based Microbiome Analyses. *BMC Biol*. **12**, 87 (2014).

28. Amir, A. *et al*. Correcting for Microbial Blooms in Fecal Samples during Room-Temperature Shipping. *mSystems*. **2**, e00199–16 (2017).

29. Fouhy, F. *et al*. The Effects of Freezing on Faecal Microbiota as Determined Using MiSeq Sequencing and Culture-Based Investigations. *PLoS One*. **10**, e0119355 (2015).

30. Song, S. J. *et al*. Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. *mSystems*. **1**, e00021-16 (2016).

31. Jumpstart Consortium Human Microbiome Project Data Generation Working Group. Evaluation of 16S rDNA-Based Community Profiling for Human Microbiome Research. *PLoS ONE*. **7**, e39315 (2012).

32. Chase, J. *et al*. Geography and Location Are the Primary Drivers of Office Microbiome Composition. *mSystems*. **1**, e00022–16 (2016).

33. Walker, A. W., Martin, J. C., Scott, P., Parkhill, J., Flint, H. J., & Scott, K. P. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome*. **3**, 26 (2015).

34. Bonnet, R., Suau, A., Doré, J., Gibson, G. R. & Collins, M. D. Differences in rDNA libraries of faecal bacteria derived from 10- and 25-cycle PCRs. *Int. J. Syst. Evol. Microbiol.* **52**, 757-763 (2002).

35. Salter SJ, Cox MJ, Turek EM, *et al*. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*. **12**:87 (2014).

36. Walters, W. A., Caporaso, J. G., Lauber, C. L., Berg-Lyons, D., Fierer, N., & Knight, R. PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics*. **27**, 1159–1161 (2011).

37. Zaneveld, J. R., Lozupone, C., Gordon, J. I., & Knight, R. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Res*. **38**, 3869–3879 (2010).

38. Okuda, S., Tsuchiya, Y., Kiriyama, C., Itoh, M. & Morisaki, H. Virtual metagenome reconstruction from 16S rRNA gene sequences. *Nat. Commun*. **3**, 1203 (2012).

39. Langille, M. G. *et al*. Predictive Functional Profiling of Microbial Communities Using 16S rRNA Marker Gene Sequences. *Nat. Biotechnol*. **31**, 814–821 (2013).

40. Aßhauer, K. P., Wemheuer, B., Daniel, R., & Meinicke, P. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics*. **31**, 2882–2884 (2015).

41. Jun, S. R., Robeson, M. S., Hauser, L. J., Schadt, C. W., & Gorin, A. A. PanFP: pangenome-based functional profiles for microbial communities. *BMC Res. Notes.* **8**, 479 (2015).

42. Scholz, M. *et al.* Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods*. **13**, 435-438 (2016).

43. Mukherjee S. *et al.* 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.* **35**, 676-683 (2016).

44. Abubucker, Sahar *et al*. Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. *PLoS Comput. Biol*. **8**, e1002358 (2012).

45. Quince, C., Walker, A.W., Simpson, J.T., Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*. **35**, 833-844 (2017).

46. Carini, P., Marsden, P. J., Leff, J. W., Morgan, E. E., Strickland, M. S. & Fierer, N. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat Microbiol*. **2**, 16242 (2016).

47. Emerson JB, Adams RI, Román CMB, et al. Schrödinger's microbes: Tools for distinguishing the living from the dead in microbial ecosystems. Microbiome. 5:86 (2017).

48. Giannoukos, G. *et al*. Efficient and Robust RNA-Seq Process for Cultured Bacteria and Complex Community Transcriptomes. *Genome Biol*. **13**, 3 (2012).

49. Wang Y, Hayatsu M, Fujii T. Extraction of Bacterial RNA from Soil: Challenges and Solutions. *Microbes and Environments*. **27**(2):111-121 (2012).

50. Tveit AT, Urich T, Svenning MM. Metatranscriptomic Analysis of Arctic Peat Soil Microbiota. Voordouw G, ed. *Applied and Environmental Microbiology*. 80(18):5761-5772 (2014).

51. Franzosa, E. A. *et al*. Relating the Metatranscriptome and Metagenome of the Human Gut. *Proc. Natl Acad. Sci. USA* **111**, E2329–E2338 (2014).

52. Maurice, C. F., Haiser, H. J., & Turnbaugh, P. J. Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell*. **152**, 39–50 (2013).

53. Bashiardes S, Zilberman-Schapira G, Elinav E. Use of Metatranscriptomics in Microbiome Research. *Bioinformatics and Biology Insights*. **10**:19-25 (2016).

54. Soergel, D. A. W., Dey, N., Knight, R., & Brenner, S. E. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME J*. **6**, 1440–1444 (2012).

55. Thompson, L. *et al*. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457-453 (2017).

56. Glenn, T. C. Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour*. **11**, 759-769 (2011).

57. 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*. **12**, 118-123 (2010).

58. Reeder, J. & Knight, R. The 'rare biosphere': a reality check. *Nat. Methods*. **6**, 636-637 (2009).

59. Caporaso, J. G. *et al*. QIIME Allows Analysis of High-Throughput Community Sequencing Data. *Nat. Methods*. **7**, 335–336 (2010).

60. Schloss, P. D. *et al*. Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl. Environ. Microbiol*. **75**, 7537–7541 (2009).

61. Callahan, B. J., McMurdie, P. J., & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J*. **11**, 2639–2643 (2017).

62. Eren, A. M., et al. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. and Evol*. **4**, 1111–1119 (2013).

63. Amir, A. *et al*. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems*. **2**, e00191–16 (2017).

64. Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. DADA2: High resolution sample inference from Illumina amplicon data. *Nat. Methods*. **13**, 581–583 (2016).

65. Lozupone, C. A. *et al*. "Meta-Analyses of Studies of the Human Microbiota." *Genome Res*. **23**, 1704–1714 (2013).

66. Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl. Environ. Microbiol*. **73**, 5261–5267 (2007).

67. McDonald, D. *et al*. An Improved Greengenes Taxonomy with Explicit Ranks for Ecological and Evolutionary Analyses of Bacteria and Archaea. *ISME J*. **6**, 610–618 (2012).

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

69. Olm, M. R. *et al*. The source and evolutionary history of a microbial contaminant identified through soil metagenomic analysis. *MBio*. **8**, (2017).

70. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*. **15**, R46 (2014).

71. Langmead, B., & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. **4**, 357-359 (2012).

72. Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res*. **26**, 1721–1729 (2016).

73. McIntyre, A. B. R. *et al*. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol*. **18**, 182 (2017).

74. Truong, D. T., *et al*. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods.* **12**, 902-903 (2015).

75. Nguyen, N., Mirarab, S., Liu, B., Pop, M., & Warnow, T. TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics*, **30**, 3548–3555 (2014).

76. Huson, D. H. *et al*. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput. Biol.* **12**, (2016).

77. O'Leary, N. A. *et al*. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. **44**, D733–D745 (2016).

78. Finn, R. D. *et al*. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res*. **44**, D279–D285 (2016).

79. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*. **31**, 926–932 (2015).

80. Markowitz, V. M. *et al*. IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res*. **40** (2012).

81. Arndt, D. *et al*. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res*. **44**, 1–6 (2016).

82. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J*. **9**, 1–10 (2014).

83. Prestat, E., *et al*. FOAM (Functional Ontology Assignments for Metagenomes): a Hidden Markov Model (HMM) database with environmental focus. *Nucleic Acids Res*. **42**, e145 (2014).

84. Sunagawa, S. *et al*. Structure and function of the global ocean microbiome. *Science*. **348**, 6237 (2015).

85. Xiao, L. *et al*. A catalog of the mouse gut metagenome. *Nat. Biotechnol*. **33**, 1103-1108 (2015).

86. Qin, J., *et al*. A human gut microbial gene catalog established by metagenomic sequencing. *Nature*. **464**, 59–65 (2010).

87. Medema, M. H. *et al*. AntiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res*. **39**, (2011).

88. Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., & Brown, C. T. Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci. USA.* **111**, 4904–4909 (2014).

89. Ye, Y. & Tang, H. Utilizing de Bruijn graph of metagenome assembly for metatranscriptome analysis. *Bioinformatics*. **32**, 1001–1008 (2016).

90. Narayanasamy, S. *et al*. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol*. **17**, 260 (2016).

91. Hug, L. A. *et al*. A new view of the tree of life. *Nat. Microbiol*. **1**, 16048 (2016).

92. Bankevich, A. *et al*. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol*. **19**, 455–477 (2012).

93. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. **31**, 1674–1676 (2014).

94. Vollmers, J., Wiegand, S. & Kaster, A. K. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - Not only size matters! *PLoS ONE*. **12**, (2017).

95. Wu, Y. W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*. **32**, 605–607 (2015).

96. Alneberg, J. *et al*. Binning metagenomic contigs by coverage and composition. *Nat. Methods*. **11**, 1144–1146 (2014).

97. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. **25**, 1043–1055 (2015).

98. Laczny, C. C. *et al*. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome*. **3**, 1 (2015).

99. Eren, A. M. *et al*. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*. **3**, e1319 (2015).

100. White Iii, R. A. *et al*. ATLAS (Automatic Tool for Local Assembly Structures) -a comprehensive infrastructure for assembly, annotation, and genomic binning of metagenomic and metatranscriptomic data. *PeerJ*. Prepr. 1–11 (2017).

101. Treangen, T. J. *et al*. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol*. **14**, R2 (2013).

102. Robinson, M. D., McCarthy, D. J., & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. **26**, 139–140 (2010).

103. Anders, S., & Huber, W. Differential expression analysis for sequence count data. *Genome Biol*. **11**, R106 (2010).

104. Sczyrba, A. *et al*. Critical Assessment of Metagenome Interpretation − a benchmark of computational metagenomics software. *bioRxiv*. 99127 (2017).

105. Barwell, L. J., Isaac, N. J. B., & Kunin, W. E. Measuring β-diversity with species abundance data. *J. Anim. Ecol*. **84**, 1112–1122 (2015).

106. Kuczynski, J., *et al*. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Methods*. **7**, 813–819 (2010).

107. Hamady, M., Lozupone, C. & Knight, R. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J*. **4**, 17–27 (2010).

108. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci*. **14**, 927–930 (2003)

109. Anderson, M.J., Walsh, D.C.I. What null hypothesis are you testing? PERMANOVA, ANOSIM and the Mantel test in the face of heterogeneous dispersions. *Ecol. Monogr*. **83**, 557–574 (2013).

110. Weiss, S. *et al*. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*. **5**, 27 (2017).

111. McMurdie, P. J. & Holmes, S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Comput. Biol*. **10**, (2014).

112. Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A., & Knight, R. EMPeror: a tool for visualizing high-throughput microbial community data. *GigaScience*. **2**, 16 (2013).

113. Aitchison, J. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society*. Series B (Methodological) **44**, no. 2, 139-77 (1987).

114. Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, 27663 (2015).

115. Weiss, S. *et al*. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J*. **10**, 1–13 (2016).

116. Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S. & B??hler, J. Proportionality: A Valid Alternative to Correlation for Relative Data. *PLoS Comput. Biol*. **11**, (2015).

117. Friedman, J. & Alm, E. J. Inferring Correlation Networks from Genomic Survey Data. *PLoS Comput. Biol.* **8**, (2012).

118. Kurtz, Z. D. *et al*. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS Comput. Biol.* **11**, (2015).

119. Schwager, E., Mallick, H., Ventz, S., & Huttenhower, C. A Bayesian method for detecting pairwise associations in compositional data. *PLoS Comput. Biol*. **13**, e1005852 (2017).

120. Washburne, A. D. *et al*. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ*. **5**, e2969 (2017).

121. Silverman, J. D., Washburne, A. D., Mukherjee, S. & David, L. A. A phylogenetic transform enhances analysis of compositional microbiota data. *Elife*. **6**, (2017).

122. Morton, J. T. *et al*. Balance Trees Reveal Microbial Niche Differentiation. *mSystems*. **2**, e00162-16 (2017).

123. Vandeputte, D., Kathagen, G., D'hoe, K. *et al*. Quantitative microbiome profiling links gut community variation to microbial load. *Nature*. **551**, 507-511 (2017).

124. Kleyer, H., Tecon. R., Or, D. Resolving Species Level Changes in a Representative Soil Bacterial Community Using Microfluidic Quantitative PCR. *Front Microbiol*. **8**:2017 (2017).

125. Knights, D., Parfrey, L. W., Zaneveld, J., Lozupone, C. & Knight, R. Human-associated microbial signatures: Examining their predictive value. *Cell Host and Microbe*. **10**, 292–296 (2011).

126. Yazdani, M., Taylor, B.C., Debelius, J., Li, W., Knight, R., Smarr, L. Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease. *IEEE*. 2016 IEEE International Conference on Big Data. 1272–1280. (2016).

127. Huang, S. *et al*. Predictive modeling of gingivitis severity and susceptibility via oral microbiota. *ISME J*. **8**, 1768–1780 (2014).

128. Teng, F. *et al*. Prediction of Early Childhood Caries via Spatial-Temporal Variations of Oral Microbiota. *Cell Host Microbe*. **18**, 296–306 (2015).

129. Metcalf, J. L. *et al*. Microbial community assembly and metabolic function during mammalian corpse decomposition. *Science*. **351**, 158–62 (2016).

130. Subramanian, S. *et al*. Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature*. **510**, 417–21 (2014).

131. Knights, D. *et al*. Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods*. **8**, 761–763 (2011).

132. Lax, S. *et al*. Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science*. **345**, 1048–1052 (2014).

133. Roume, H. *et al*. A biomolecular isolation framework for eco-systems biology. *ISME J*. **7**, 110–121 (2013).

134. Nicholson, J. K. & Lindon, J. C. Systems biology: Metabonomics. *Nature*. **455**, 1054–6 (2008).

135. Wang, R. & Seyedsayamdost, M. R. Hijacking exogenous signals to generate new secondary metabolites during symbiotic interactions. *Nat. Rev. Chem*. **1**, 21 (2017).

136. Huan, T. *et al.* Systems biology guided by XCMS Online metabolomics Addressing reproducibility in single- laboratory phenotyping experiments. *Nat. Publ. Gr.* **14**, 461–462 (2017).

137. Hurley, J. R. & Cattell, R. B. The procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behav. Sci.* **7**, 258–262 (1962).

138. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Research*. **27** (2): 209–220 (1967).

139. Doledec, S. & Chessel, D. Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshw. Biol.* **31**, 277–294 (1994).

140. Boulesteix, A. & Strimmer, K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform.* **8**, 32–44 (2007).

141. Witten, D. M. & Tibshirani, R. J. Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. *Stat. Appl. Genet. Mol. Biol.* **8**, 1–27 (2009).

142. Wilms I., Croux C. Robust sparse canonical correlation analysis. *BMC Systems Biology*. **10**:72 (2016).

143. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).

144. Dhanasekaran, A. R., Pearson, J. L., Ganesan, B. & Weimer, B. C. Metabolome searcher: a high throughput tool for metabolite identification and metabolic pathway mapping directly from mass spectrometry and using genome restriction. *BMC Bioinformatics*. **16**, 62 (2015).

145. Protsyuk, Ivan. *et al.* 3D molecular cartography using LC-MS combined with Optimus and `ili software. *Nat. Protocols*. **13**, 134-154 (2018).

146. McHardy, I. H. *et al*. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome*. **1**, 17 (2013).

147. Whiteson, K. L. *et al*. Breath gas metabolites and bacterial metagenomes from cystic fibrosis airways indicate active pH neutral 2,3-butanedione fermentation. *ISME J*. **8**, 1247–1258 (2014).

148. Maurice, C.F., Haiser, H.J., Turnbaugh, P.J. Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell*. **152**(1-2):39-50 (2013).

149. Theriot, C.M., Koenigsknecht, M.J., Carlson, P.E., *et al*. Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection. *Nat. Commun.*, **5**:3114 (2014).

150. Erickson, A. R. *et al*. Integrated Metagenomics/Metaproteomics Reveals Human Host-Microbiota Signatures of Crohn's Disease. *PLoS One*. **7**, e49138 (2012).

151. Hultman, J., Waldrop, M.P., Mackelprang, R., *et al.* Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature*. **521**, 208-212 (2015).

152. Jagtap, P.D., Blakely, A., Murray, K., et al. Metaproteomic analysis using the Galaxy framework. *Metaproteomics*. **15**,20, 3553-3565 (2015).

153. Cheng K., Ning Z., Zhang X., et al. MetaLab: an automated pipeline for metaproteomic data analysis. *Microbiome*. **5**:157 (2017).

154. Ríos-Covián, D. *et al*. Intestinal Short Chain Fatty Acids and their Link with Diet and Human Health. *Front. Microbiol*. **7**, 185 (2016).

155. Balskus, E. P. Colibactin: understanding an elusive gut bacterial genotoxin. *Nat. Prod. Rep*. **32**, 1534–40 (2015).

156. Quinn, R. A. *et al*. Microbial, host and xenobiotic diversity in the cystic fibrosis sputum metabolome. *ISME J*. **95384**, 1–16 (2015).

157. Fang, H., Huang, C., Zhao, H. & Deng, M. CCLasso: Correlation inference for compositional data through Lasso. *Bioinformatics*. **31**, 3172–3180 (2015).

158. Lê Cao, K. A., González, I. & Déjean, S. IntegrOmics: An R package to unravel relationships between two omics datasets. *Bioinformatics*. **25**, 2855–2856 (2009).

159. Wikoff, W. R. *et al*. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proc. Natl. Acad. Sci. USA*. **106**, 3698–703 (2009).

160. Yilmaz, P., *et al*. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specificaitons. *Nat. Biotech*. **29**, 415-420 (2011).

161. Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D., & Knight, R. Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res*. **35**, e120 (2007).

162. The Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease. *Cell Host Microbe*. **16**, 276–289 (2014).

163. Korem, T. *et al*. Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science*. **349**, 1101–1106 (2015).

164. Sangwan, N., Xia, F., & Gilbert, J. A. Recovering complete and draft population genomes from metagenome datasets. *Microbiome*. **4**, 8 (2016).

165. Bikel, S. *et al*. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. *Comput. Struct. Biotechnol. J*. **13**, 390–401 (2015).

166. Sultan, M. *et al*. Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics*. **15**, 675 (2014).

167. Peano, C. *et al*. An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microb. Inform. Exp*. **3**, 1 (2013).

Box 1. Metabolomics and the microbiome

Microbially produced metabolites influence host physiology, can shape microbial community dynamics and are involved in both health and disease. These metabolites can have both beneficial (for example, short-chain fatty acids (SCFAs)[154]) and detrimental effects on the host (for example, the genotoxin colibactin[155]). However, identifying a metabolite as sourced from the microbiome is particularly challenging. Even more challenging is identifying which microorganism or collection of microorganisms produced or modified a particular metabolite. Here are several strategies to address this problem:

1. **Compare metabolites from natural samples to those from cultured isolates of microbiome-isolated microorganisms.** One useful approach is matching tandem mass spectrometry data from cultured isolates to clinical or environmental samples, showing that a particular metabolite signature can be sourced from the cultured microorganism[156].
2. **Map metabolites detected in a microbiome sample to paired genome or metagenomic data.** Some metabolites are unique to particular microbial taxa. Detection of these metabolites in a natural sample can enable determination of their likely source by mining paired genomic data for genes known to produce that metabolite. For example, 2,3-butanedione, a unique fermentation product, is a microbial metabolite produced by *Streptococcus* spp. Detection of this metabolite in clinical samples along with the biosynthetic genes, facilitates mapping of reads to the biochemical pathway back to the genome of the organism of origin[147].
3. **Build co-occurrence networks of microorganisms and metabolites**. Co-occurrence or correlation methods associate microorgaisms with metabolite features. This is an active area of research, but available algorithms that have been optimized for detecting correlations between microorganisms in sparse microbiome data include SparCC[117], CCLasso[157], and others[115,158]. However, this approach warrants caution because of the high false discovery rates across the large multivariate datasets.
4. **Germ free versus specific pathogen free murine models**. These comparisons identify metabolites from the microbiome as metabolites detected in colonized mice but not in uncolonized mice are likely produced by microorganisms. Gnotobiotic mice (mono-colonized or with defined communities) help identify specific microorganisms that produce metabolites of interest[159].

Box 2. Good working practices

It is crucial for microbiome analyses to be reproducible. Similar microbiome studies can often have conflicting results, and without proper documentation of sample collection, data processing, and analysis methods, it is difficult to re-examine the data and reconcile these differences. As the field evolves, it will be necessary to revisit early experiments and potentially re-analyze the data with updated tools. Reproducibility is paramount for this process to be possible and efficient. When collecting samples, details of the collection process should be recorded in the experimental metadata to ensure that as much variability as possible is accounted for. Additionally, the Genome Standards Consortium minimum information standards (MIxS) for marker genes (MIMARKS) and metagenomes (MIMS)[160] should be adhered to. These unified standards enable comparisons across data sets. During bioinformatics processing, researchers should track all of the commands that they ran and all software versions that they used, and deposit their raw data and metadata in public repositories. We recommend using tools such as Jupyter Notebooks (http://jupyter.org) or R Markdown (https://rmarkdown.rstudio.com/) to facilitate this, and then storing the notebooks in a revision control management system such as GitHub (https://github.com). Some software packages, such as QIIME 2[59] (https://qiime2.org) and Galaxy (https://usegalaxy.org/) automatically track this information for researchers through an integrated data provenance tracking system. Qiita (http://qiita.microbio.me) and EBI (http://www.ebi.ac.uk/) are powerful meta-analysis and data archiving tools, respectively, and when combined allow a researcher to analyze their microbiome data in the context of tens of thousands of other samples, which enables the data to be re-used by future researchers.

Box 3. Considerations for different microbiomes

Although microbiome data analysis methods are widely applicable to many sample types and environments, experimental design, and method selection require careful consideration for different sample types. First, one must consider the composition of the sample and feasibility of use for different methods. For samples that are heavily contaminated with non-microbial DNA, such as tissue, shotgun metagenomic sequencing may not be feasible without non-microbial DNA depletion. Depending on the experimental question, samples heavily contaminated with relic DNA from dead microorganisms, such as soil, may require physical removal of relic DNA by propidium monoazide[46] or other methods prior to DNA extraction. The amount of sample to collect is also determined by sample type. Whereas a high biomass fecal sample may only require a swab, samples with low microbial density may necessitate larger volumes and potentially concentration for sufficient DNA extraction. For example, ocean microbiome samples are usually large volumes of water run though a filter to trap and concentrate the target organisms prior to DNA extraction[84]. Though in all cases appropriate controls should be included, low biomass environments, such as blood, spinal fluid or laboratory clean rooms, particularly necessitate controls that have gone through the entire sampling process to fully characterize contaminants. DNA contaminants can be found in numerous reagents, including swabs, DNA extraction kits and PCR reagents[27]. Furthermore, the method of sample preservation is both dictated by analysis method and sample type. For example, metatranscriptomics requires an RNAse inhibitor and metabolomics requires sample preservation that does not interfere with metabolite extraction or data collection.

In addition to sampling considerations, study design and metadata collection also require careful tailoring to sample type and environment. For example, animal studies require an evaluation of co-housing cage effects and should stratify experimental groups into multiple cages. Fresh samples should be collected and the mouse of origin should be recorded in the metadata. Environmental samples require collection of metadata related to environmental conditions, such as pH, salinity, elevation, and depth for soil samples. The manner of collection is highly dependent on sample type and cannot be detailed for all possible samples in this Review. We recommend consulting well-validated protocols related to the sample type of interest. In any case, methods of collection, preservation, and storage should remain consistent across all samples in a study to avoid introducing confounding variation. Sample composition can be affected by outgrowth of certain microorganisms during storage at room temperature[28].

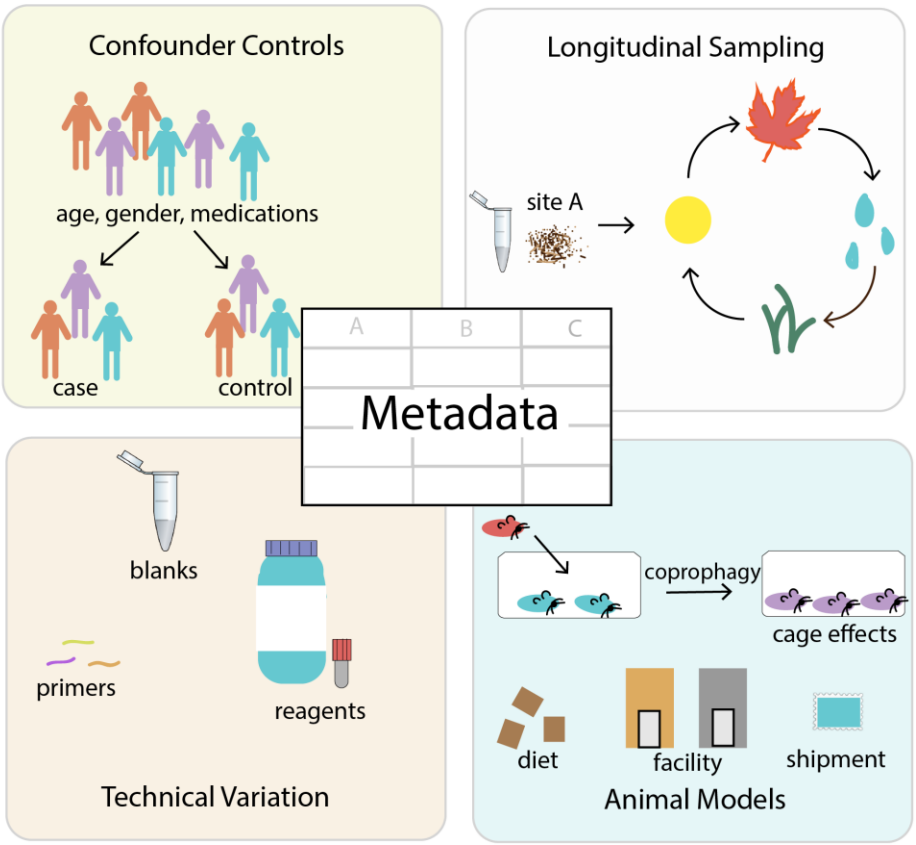

Figure 1. **Experimental design considerations for microbiome experiments**.

Conducting a robust microbiome experiment warrants careful attention to numerous factors. Stratification by potential confounders (for example, age, gender, diet, lifestyle factors and medications) can help resolve differences in microbiota between groups of interest which might otherwise be masked by a confounder-effect[5]. Longitudinal studies are especially powerful as they both control for confounding factors and allow for the assessment of community stability[7]. Similar considerations apply to animal studies, though the additional impact of coprophagy must be addressed in experimental design. For all studies, standardizing technical factors and sample processing is essential to control for variation introduced by 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, is crucial for data interpretation; without metadata, it is difficult to draw meaningful conclusions from sequencing data.

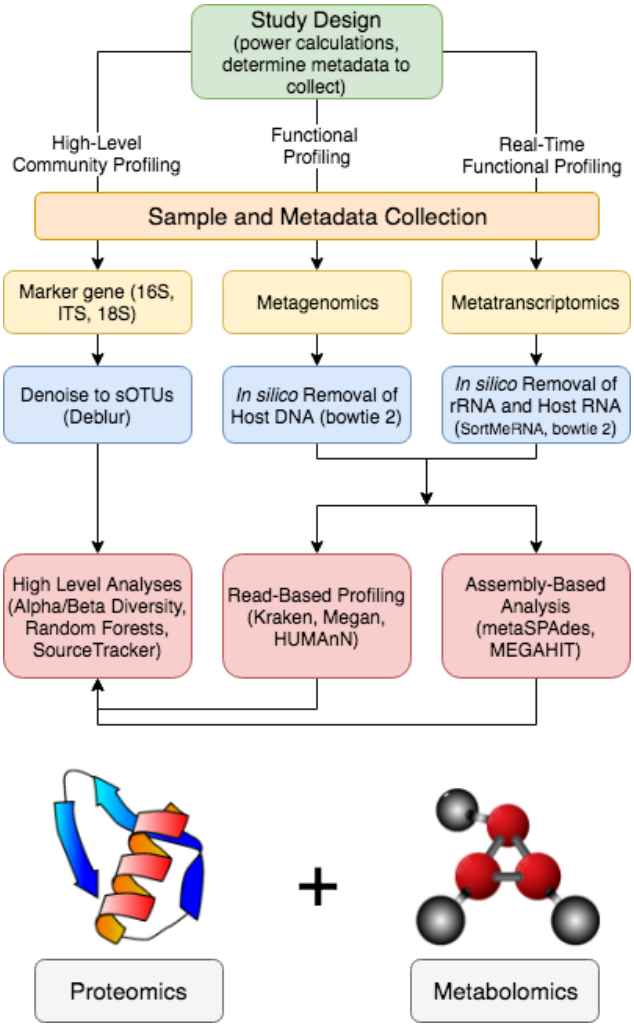

Figure 2. **Best workflow for 16S ribosomal RNA, metagenomic, and metatranscriptomic sequencing.**

After careful design and sample collection, microbiome data is generated from 16S ribosomal RNA (rRNA), metagenomic or metatranscriptomic sequencing. After performing 16S rRNA sequencing, we recommend using Deblur[63] to resolve sequence data into single-sequence variants called "sub-operational taxonomic units (sOTUs)." Although DADA2 and Deblur achieve the similar results, Deblur is an order of magnitude faster than DADA2, is parallelizable, and shows greater stability (that is, it obtains the same sOTUs across different samples)[63]. Metagenomics and metatranscriptomics first require pre-processing to remove either host DNA or rRNA and host RNA. The resultant sequencing data can be analyzed by either read-based profiling using state-of-the-art tools such as Kraken[70], Megan[76], or HUMAnN[44], or by assembly-based analyses, with tools such as metaSPAdes[92] and MEGAHIT[93]. For each of these three methods, higher level analyses (for example, alpha and beta diversity, taxonomic profiling and machine learning) are subsequently used to find overall patterns in microbiome variation. Random Forests regression has been effective in many applications, ranging from dating time since death of a corpse[129] to providing an index for microbiome maturation[130]. SourceTracker[131], a Bayesian estimator of the sources that make up each unknown community, is useful for classifying microbial samples according to environment of origin[132].

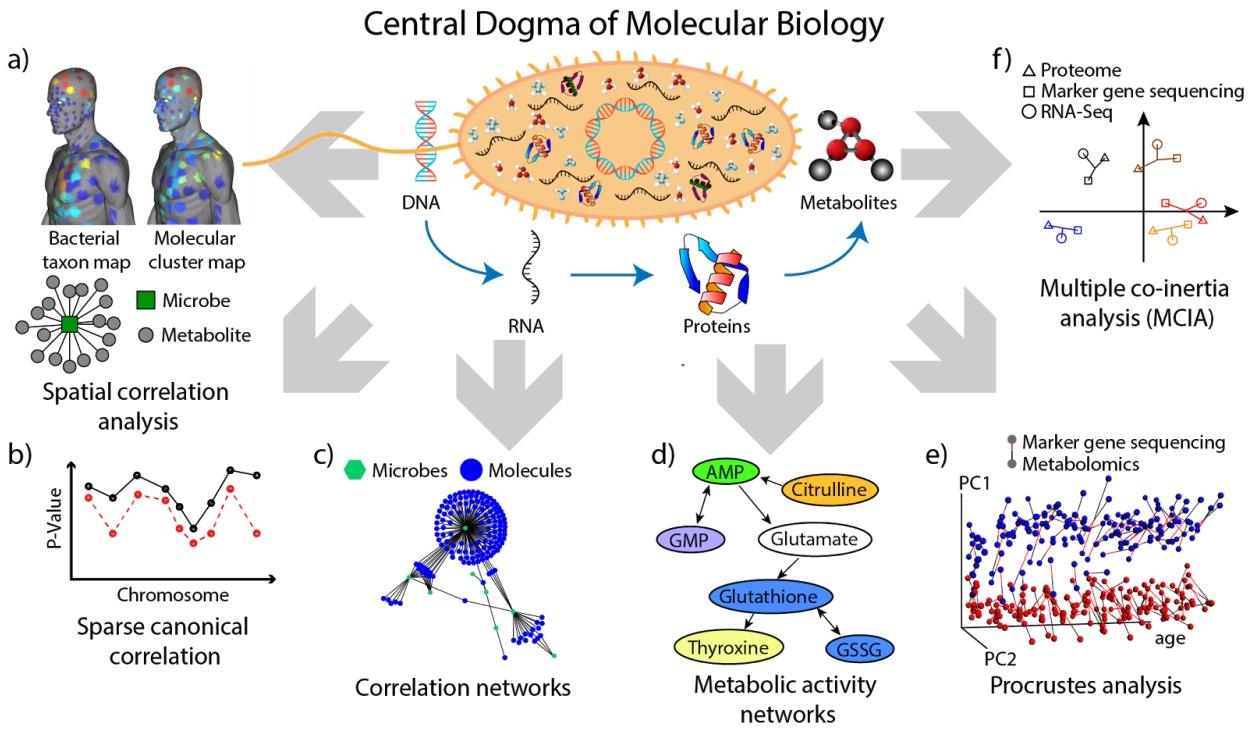

Figure 3. **Integrating "-omics" data with microbiome data.**

The central dogma of molecular biology of progression from genes to downstream metabolic products is reflected by the compendia of corresponding "-omes" co-occurring within the cell. Linking the knowledge from different "-omics" studies constitutes the multi-omics analysis. Panels around the cell represent some integration examples of various "-omics" data with marker gene sequencing: a) Three-dimensional visualization of mapped molecular and microbial (or any other) features aids our understanding of spatial correlation thereof. b) Sparse canonical correlation analysis[141] identifying linear combinations of the two sets of variables that are highly correlated with each other. c) Correlation network analysis shows clustering of a particular microorganism with metabolites that are potentially produced and/or processed by it. d) Metabolic activity networks help to predict microbial community structure and function by mathematical modelling of the molecular mechanisms of particular organism(s). e) Procrustes analysis enables the direct comparison of different "-omics" data sets with the same internal structure on a single PCoA plot to reveal trends in the data. f) Multiple co-inertia analysis (MCIA) enables multidimensional comparisons through graphical representation, so that the similarity of different "-omics" data can be more easily understood.

Table 1: Pros and cons of genomic analyses for evaluating microbial communities.

Method	Pros	Cons
Marker gene analysis	• Quick, simple and inexpensive sample preparation and analysis[55,59] • Correlates well with genomic content[37-41] • Amenable to low-biomass and highly host-contaminated samples • Large existing public data sets for comparison[16,55,160]	• No live, dead or active discrimination • Subject to amplification biases[34] • Choice of primers and variable region magnifies biases[33,54,159] • Requires a priori knowledge of microbial community[36] • Resolution typically limited to genus level at best • Functional information is limited [39,40]
Whole metagenome analysis	• Can directly infer the relative abundance of microbial functional genes; microbial taxonomic and phylogenetic identity to species and strains level is attainable for known organisms[42] • Does not assume knowledge of microbial community (i.e., captures phages, viruses, plasmid, microbial eukaryotes, etc.) • No PCR-related biases • Can estimate in situ growth rates for target organisms with sequenced genomes[161] • Can allow assembly of population-averaged microbial genomes[43,162] • Can be mined for novel gene families	• Relatively inexpensive, laborious and complex sample preparation and analysis • Contamination from host-derived DNA and organelles may obscure microbial signatures • Viruses and plasmids are not typically well annotated by default pipelines • Deep sequencing depths are typically required relative to other methods • No live, dead or active discrimination • Population-averaged microbial genomes tend to be inaccurate owing to assembly artefacts
Metatranscriptome analysis	• Can estimate which microorganisms in a community are actively transcribing when paired with marker gene analysis • Inherently discriminates between active live organisms versus dormant or dead microorganisms and extracellular DNA • Captures dynamic intra-individual variation[51] • Directly evaluates microbial activity, including responses to intervention and event exposure[52]	• Most expensive, laborious and complex sample preparation and analysis[163] • Host mRNA contamination and rRNA must be removed[48,164,165] • Requires careful sample collection and storage • Data are biased towards organisms with high transcription rates • Requires paired DNA sequencing to decouple transcription rates from bacterial abundance changes.