nature mental health

Perspective

https://doi.org/10.1038/s44220-023-00148-3

Bugs as features (part 1): concepts and foundations for the compositional data analysis of the microbiome–gut–brain axis

Received: 11 October 2022

Accepted: 14 September 2023

Published online: 5 December 2023

Check for updates

Thomaz F. S. Bastiaanssen [1,2] ✉, Thomas P. Quinn[3,5] & Amy Loughman[4,5]

There has been a growing acknowledgment of the involvement of the gut microbiome—the collection of microorganisms that reside in our gut—in regulating our mood and behavior. This phenomenon is referred to as the microbiome–gut–brain axis. Although our techniques to measure the presence and abundance of these microorganisms have been steadily improving, the analysis of microbiome data is non-trivial. Here we present a perspective on the concepts and foundations of data analysis and interpretation of microbiome experiments with a focus on the microbiome–gut–brain axis domain. We give an overview of foundational considerations before commencing analysis alongside the core microbiome analysis approaches of alpha diversity, beta diversity, differential feature abundance and functional inference. We emphasize the compositional data analysis paradigm. Furthermore, this Perspective features an extensive and heavily annotated microbiome analysis in R, as a resource for new and experienced bioinformaticians alike.

Microorganisms can be found in large numbers in almost all environments, including in and on the human body. The largest collection of microorganisms on humans can be found in the gut and is referred to as the gut microbiome. According to recent estimates, the human gut microbiome typically consists of around 3×10^{13} microorganisms, weighing approximately 200 g (ref. 1). In terms of genetic diversity, the microbiome outmatches its human host by more than three orders of magnitude, and has co-evolved with their eukaryotic hosts[2]. While the gut microbiome typically refers to microorganisms housed in the large intestine, there are microbial niches throughout the gastrointestinal tract that have relevance for health states, in particular in the oral cavity (for example, tongue, plaque, gingival surfaces) and the small intestine. For completeness, we use 'gut microbiome' or 'gut microbiota' to encompass any and all microbial communities along the gastrointestinal tract.

The microbiome–gut–brain axis

With the advent of high-throughput sequencing, the gut microbiome has become a popular subject of investigation, as evidenced by large scientific endeavors designed to map the human microbiome to health and disease[3–7]. As part of these efforts, it has become increasingly clear that the microbiome is in constant bidirectional communication with the host, and that both systems influence each other on multiple levels. The bidirectional communication between the microbiome and the host brain is referred to as the microbiome–gut–brain axis. There are several ways in which this communication occurs, for example, in the production of neuroactive compounds and metabolites such as short-chain fatty acids, modulation of the immune system and direct stimulation of the vagus nerve[8]. Besides having an important role in gut–brain communication during health and homeostasis, the microbiome has also been found to be affected by psychotropic medication[9,10]. In some cases, the microbiome can even metabolize psychotropic medication such as L-DOPA, which is frequently prescribed for Parkinson's disease[11].

The oral microbiome, typically sampled via saliva, has also been reported to associate with depressive states, risk of dementia, metabolic health, and cardiovascular disease[12–15]. Similar to the large intestinal microbiome–gut–brain axis, there are numerous routes of communication between oral microorganisms and the central nervous

[1]APC Microbiome Ireland, University College Cork, Cork, Ireland. [2]Department of Anatomy and Neuroscience, University College Cork, Cork, Ireland. [3]Independent researcher, Geelong, Victoria, Australia. [4]IMPACT (the Institute for Mental and Physical Health and Clinical Translation), Food and Mood Centre, Deakin University, Geelong, Victoria, Australia. [5]These authors jointly supervised this work: Thomas P. Quinn, Amy Loughman. ✉e-mail: thomazbastiaanssen@gmail.com

system, including direct translocation of microorganisms via facial nerves, the olfactory system and the bloodstream, as well as indirect neuroinflammatory effects via systemic inflammation caused by periodontal infection. The small intestine is more difficult to study; however, it has established roles in carbohydrate metabolism, bile acid deconjugation and micronutrient storage[16]. It is implicated in gastrointestinal pathophysiologies such as environmental enteric dysfunction, pouchitis and irritable bowel syndrome (IBS), of relevance to the pathophysiology[17].

A perspective on microbiome bioinformatics analysis

Microbiome analysis is complex, and the discoveries about methods and biology alike are evolving constantly. This Perspective aims to make microbiome data analysis less daunting by presenting a concise description of the key steps involved. Although there are many reasonable approaches to analysing the microbiome, we set out to provide the reader with at least one such approach. Here we present an overview of the various methods used to analyze, interpret, and visualize microbiome studies. The text below focuses on high-level concepts, but we also include a fully reproducible analysis in Supplementary Data 1, written in R Markdown, that takes our readers through a complete analysis of microbiome data, starting at the feature table.

This two-part Perspective series and the accompanying Supplementary Data 1 were written with an audience specialized in biological psychiatry in mind and many of the examples in this paper reflect this. However, we argue that the points discussed here can be applied to most, if not all, host–microbiome experiments. An overview of a typical microbiome analysis is shown in Fig. 1.

Getting ready for the analysis

Pre-registration

Pre-registration is a main component of reproducible science and is becoming a routine practice[18–20]. We stress here that we are not advocating for the pre-registration of any and all microbiome studies. Indeed, exploratory studies play a crucial role in mapping 'microbial dark matter' and allow for subsequent hypothesis generation. Pre-registration involves documenting hypotheses and an analysis plan for a study before examining data and running the analysis. It is a practical commitment to avoid 'fishing' and the selective presentation of results on the basis of significance, and to mitigate against the known cognitive biases of human reasoning (for example, confirmation bias)[21]. In microbiome science, where the control of the type 1 error rate is critical, and the reproducibility of findings is particularly challenging[22], pre-registration is especially important. Early indications suggest that the practice has reduced publication bias for positive results[23], and can therefore improve the integrity of published research. Pre-registration tools prompt researchers to describe their study and research questions, and then generate a date-stamped document that can be published with a digital object identifier (DOI) either immediately or after a user-defined period of embargo (for example, following publication). The pre-registration document thus serves as a public record of the planned analyses and analytic strategy that can be referenced in resulting publications to affirm that the findings reflect a hypothesis-driven analysis. Pre-processing steps should also be specified a priori where practicable, as these will affect downstream results. There are a number of free tools and guidelines for pre-registration, including the Open Science Framework (osf.io) and As Predicted (aspredicted.org). These are akin to clinical trial registration sites, and are suitable for observational and experimental studies alike. Guidance as to relevant details to include in pre-registration and study design more broadly can be sought from emerging consensus checklists such as Strengthening The Organization and Reporting of Microbiome Studies (STORMS)[24]. It is important to note that within a pre-registration framework, exploratory and post hoc analyses are still entirely valid. Indeed, within a relatively young and rapidly evolving field such as microbiome science, it is appropriate to continue hypothesis generation and exploration. However, exploratory analyses should be presented as such, and should be clearly distinguished from confirmatory hypothesis testing[21]. Pre-registration may include a discussion about the power calculations used to select the sample size, which we discuss in the companion piece to this paper[25]. The special case of experimental designs involving fecal microbiota transplantation is also discussed there. Also, see ref. 26.

Considering potential confounding factors. A key challenge of microbiome research, and in particular observational studies of the human microbiome, is delineating the variable of interest from other factors that influence the ecosystem. Genetics, ethnicity, early life factors such as modes of birth and feeding and stress, habitual diet, environmental exposures, and medication use are just some of the important contributors to the human microbiome, and are also often related to the outcome of interest[9,27–29]. These are always worth considering in microbiome research, especially in the context of causal inference modeling. The most appropriate way to account for these confounding factors is highly context-dependent and can be difficult to determine. Clearly communicated and reproducible methods are key. We discuss three important and related approaches to statistically deal with these factors, a process referred to as deconfounding.

- Linear modeling. One straightforward and common strategy to deal with these factors is to include them as covariates in a statistical model, most frequently a generalized linear model. These flexible models can be thought of as more general versions of frequently used 'named' statistical tests such as the *t*-test, Pearson's correlation coefficient and analysis of variance. Also see the demonstration found in Supplementary Data 1, where we demonstrate and discuss including covariates in a linear model. Relatedly, one could compare two models—an H_0 (null) model just including the covariates used as confounders versus an H_1 model that includes the microbiome feature of interest in addition to those covariates—using a log-likelihood test or similar.

Often however, as a wide variety of factors affect both the microbiome and the brain side of the microbiome–gut–brain axis, it can be unclear how many and which factors to include as covariates.

- Stepwise variable selection. Also referred to as stepwise regression, is often used as a way to statistically determine which factors to include as covariates in a data-driven manner. In brief, this method involves sequentially adding or subtracting covariates from a model and recalculating the test statistic, retaining those covariates that show statistical significance.

We note here that statistical significance does not determine whether a variable is or is not a confounder. Ideally, biological knowledge should inform the inclusion or exclusion of a covariate in a model. The biostatistician should always consider the biological interpretation of including or excluding a covariate from a statistical model. Practically, however, this may not always be feasible, especially in highly complex biological datasets such as from microbiome science.

- Causal inference. A third approach that does require biological interpretation. Although not always the stated aim, causal inference is frequently the underlying motivation for studies of the microbiome. Consider psychiatry, for example. We may aim to estimate the effects that the gut microbiome has on some parameter of brain function, whether it be mood, behavior, cognition or a neurodevelopmental indicator. Other valid study aims might include description (for example, of the microbiome in people experiencing depression) and prediction (for example, can we predict who will develop depression based on their microbial features); however, causal interpretations are often attributed even to these kinds of studies[30]. In the companion piece to this paper,

https://doi.org/10.1038/s44220-023-00148-3

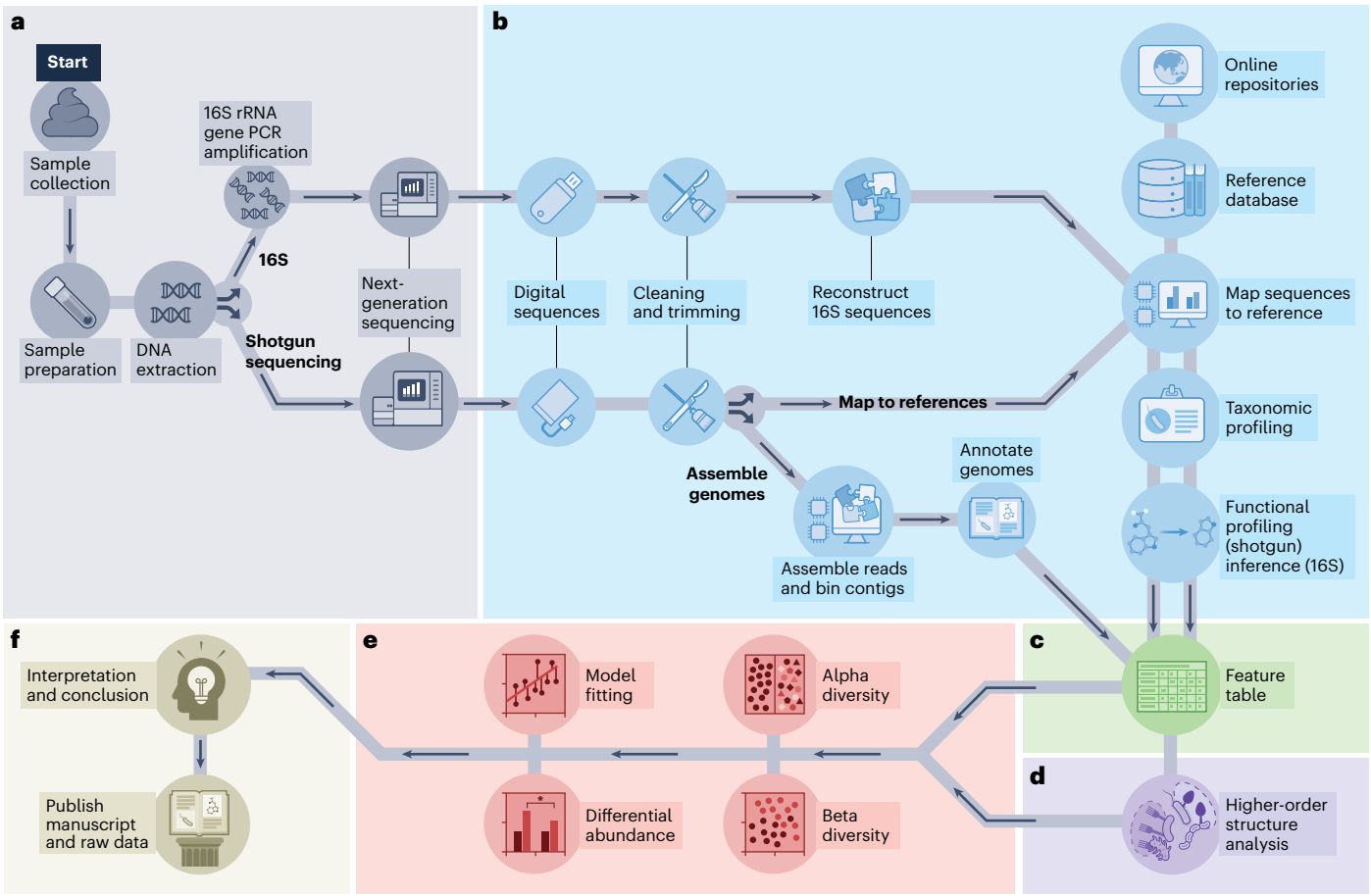

Fig. 1 | Overview of what a typical gut microbiome analysis may look like. **a**, The pre-digital part of the pipeline. Genetic material is isolated and digitized, using either 16S rRNA gene or metagenomic shotgun sequencing. **b**, The digitized reads are annotated based on taxonomy and/or function. This process is distinct between data from 16S and metagenomic shotgun sequencing. In the case of shotgun sequencing, reads can be mapped directly to reference genomes, or reads can be assembled to MAGs, which can in turn be annotated for taxonomy and functional content. Often, both approaches are employed within the same study. **c**, The features are tallied up into feature tables. **d**, Higher-order structures are identified and derived from the feature table. Examples include mesoscale structures such as interaction networks, trophic layers, ecological guilds, and functional modules. **e**, The features of the microbiome are assessed statistically. Special attention should be given here to controlling the false discovery rate. **f**, Finally, the findings are interpreted and presented for peer review. In tandem with publication, raw data should be made available to other researchers by uploading to a repository such as the SRA or ENA. For a more in-depth comparison of sequencing protocols, we refer the reader to specialized reviews[55,56].

we discuss the five phases of causal inference analysis adapted from Ponsonby and present a directed acyclic graph modeled on the example from the Zhu et al. dataset expanded on in the accompanying demonstration[25,31,32].

Even within descriptive or predictive studies, it can be useful to examine whether causal features such as dose response or temporality exist. Causal questions are frequently implied even in cross-sectional and associative human studies, for example, in which the microbiome is not being manipulated, and its effect is therefore not being explicitly measured. For this reason, causal inference principles have broad relevance. Importantly, causal inference is not the same as assigning causality based on an observational study; rather, causal inference seeks to determine whether the data support a causal hypothesis by performing statistical analyses within a causal framework.

The feature table

Although 16S/amplicon and shotgun sequencing differ widely in execution, the type of data that is obtained tends to converge downstream in the analysis. After pre-processing, both 16S- and shotgun-sequencing methodologies yield a feature abundance table. A feature abundance table shows how many observations (that is, counts) there were for each feature (for example, microorganism, function, gene and so on) per sample. Some programs and frameworks, such as marker gene-based operational taxonomic units (mOTUs) and many shotgun-sequencing tools, do not produce count tables but rather (relative) abundance tables (total sum scaling (TSS)-transformed). Fortunately, the compositional methods discussed in Box 1 are still appropriate in these cases. By convention, a feature abundance table should have features as columns and samples as rows. Although many software tools assume this organization, there are notable exceptions, so it is always worth checking the software before proceeding with an analysis. It is tempting to directly correspond a count or abundance to a biological instance or abundance of a feature in a sample, but owing to biases inherent to metagenomic sequencing[33–35], raw abundances should be pre-processed first, for example, via normalization or log-ratio transformation. Sometimes counts and abundances are instead expressed as compositional data, which we discuss in Box 2. See Supplementary Information Sections 1.2 and 1.3 for tools to generate feature tables from 16S- and shotgun-sequencing data, respectively.

Rare features and rarefaction

Before the microbiome analysis starts, it is common to filter out features by removing them entirely from the feature table. Testing fewer

BOX 1

16S rRNA gene/amplicon versus shotgun sequencing

Generally speaking, two methods of microbiome sequencing are widely used. The first is 16S rRNA gene sequencing, also called amplicon sequencing or simply '16S', includes methods where an evolutionarily preserved genomic sequence is targeted and sequenced. The second is metagenomic shotgun sequencing, where all genetic material in a sample is targeted and sequenced.

Although downstream bioinformatics analyses of both types of microbiome sequencing techniques converge, the actual techniques are distinct[57]. When deciding whether to perform 16S or shotgun sequencing in host–microbiome experiments, broadly speaking, it is often preferable to perform shotgun sequencing, as the approach allows for higher-resolution analysis and provides the researcher with direct information on the genetic content of a microbiome sample. Furthermore, shotgun sequencing allows for the reconstruction of uncharacterized genomes, enabling the researcher to investigate novel microorganisms. Shotgun sequencing also enables the user to track a specific microorganism through several samples to perform transmission analysis (see ref. 58 and the section on fecal microbiota transplantation in the companion paper[25]). We also note here that 16S-sequencing results can potentially be biased owing to 16S rRNA gene copy number variation. Copy number variation can be thought of as a multiplicative (scalar) bias of the estimated relative feature abundance. Conveniently, compositional difference between samples is invariant to this type of multiplication by a fixed vector[34,59] (Box 2). Unaccounted copy number variation does, however, bias alpha diversity indices (Supplementary Section 2.1).

There are, however, some scenarios where 16S is preferable. First, the price of 16S is lower than that of shotgun—although this difference is decreasing, and shallow shotgun sequencing has been used as a cost-friendly alternative to 16S[60]. Furthermore, because 16S targets only a gene unique to microorganisms and has a PCR amplification step, the technique is preferable in samples with a low microbial biomass or with a large proportion of non-microbial (host) genetic material such as the tumor microbiome. Notably, both techniques are biased towards detecting specific genetic sequences and thus by extent specific microbial taxa[61]. Such biases are known to occur between metagenomic sequencing experiments of the same type, even between runs in the same laboratory[34,35].

Although wet-lab protocols, including sequencing protocols, are outside the scope of this paper, they greatly impact the quality and content of the resulting microbiome sequencing data. For instance, the microbiome sampling kit itself can contribute a microbiome signature to a measurement[62]. This can be particularly impactful in environments with a low microbial biomass. There is an ongoing effort to standardize microbiome sequencing protocols, for which we refer the reader to specialized reviews[55,56].

See Supplementary Sections 1.1, 1.2 and 1.3 for recommendations on quality control, pre-processing 16S-sequencing data and shotgun-sequencing data to generate feature tables, respectively.

BOX 2

Microbiome data are compositional

Compositional data refers to a type of data that can be described as a set of proportions, percentages, or probabilities, or with a constant or arbitrary sum. Rather than the relative abundance or sizes of the components, the ratios between components hold information in compositional data[63]. Over the past few years, awareness has grown such that microbiome datasets are compositional, which, if ignored, can lead to spurious results. The field of study of how to deal with compositional data is called compositional data analysis (CoDA). There are excellent reviews on CoDA in general and how it relates to the microbiome, in particular, that we encourage our audience to read[41,64,65].

CoDA theory and practice is continually evolving, and a full review of the field is beyond the scope of this paper. In this Perspective, we recommend, at a minimum, performing a CLR or similar transformation (for example, PhILR) on the count data before performing statistical analysis or visualizing the data. Notably, these log-ratio transformations do not require count data as input but could even be performed on expressly relative abundance data such as TSS-transformed feature tables. In contrast, transformations we expressly recommend against here, to account for the effects of compositionality, include log-transformations and TSS on their own.

However, there are three notable exceptions.

- First, alpha diversity should not be done on log-ratio-transformed data. This is due to the nature of the formulas that are used to compute alpha diversity metrics, which all take feature abundance as input[66] (Fig. 2a).
- Second, stacked bar plots should be created using counts normalized to one or to percentages. This transformation is often referred to as TSS. Here, it is worth noting that stacked bar plots depict the proportion of observed reads, which is distinct from the actual sample (relative) abundance[34,35].
- Third, correlating taxa to each other, for example, as part of a network analysis, warrants special attention. This is because one of the properties of compositional data is that features are inherently negatively correlated. Indeed, Karl Pearson warned against applying his namesake Pearson's correlation coefficient on compositional data[67]. Alternatives are available from the propr library[68], the SparCC library[69] and the SPIEC-EASI library[70].

Note that because microbiome count data typically have many zeros and the logarithm of a zero is undefined, the zeroes in microbiome count data must be addressed. Several reasonable solutions have been proposed, but it remains an open question as to which among these solutions performs best (although compare ref. 71, for one benchmark). In the demonstration found in Supplementary Data 1, we employ an approach derived from ref. 71, where zeroes are replaced by two-thirds of the lowest non-zero value. After the CLR transformation, the values of features can take on any value (unlike count data, which cannot be negative). After transformation, classical statistical approaches can be applied as usual.

features reduces the magnitude of the false-discovery-rate adjustment penalty, which in turn helps to increase the statistical power for the remaining tests. Most often, the filtered features represent rare taxa or rare genes. Commonly, features that are detected in only a certain percentage of samples are removed. This is referred to as prevalence filtering. Similarly, features that are detected in only low levels can be dropped. This is referred to as abundance filtering. It is worth considering the underlying reasons why a feature may have low prevalence or

https://doi.org/10.1038/s44220-023-00148-3

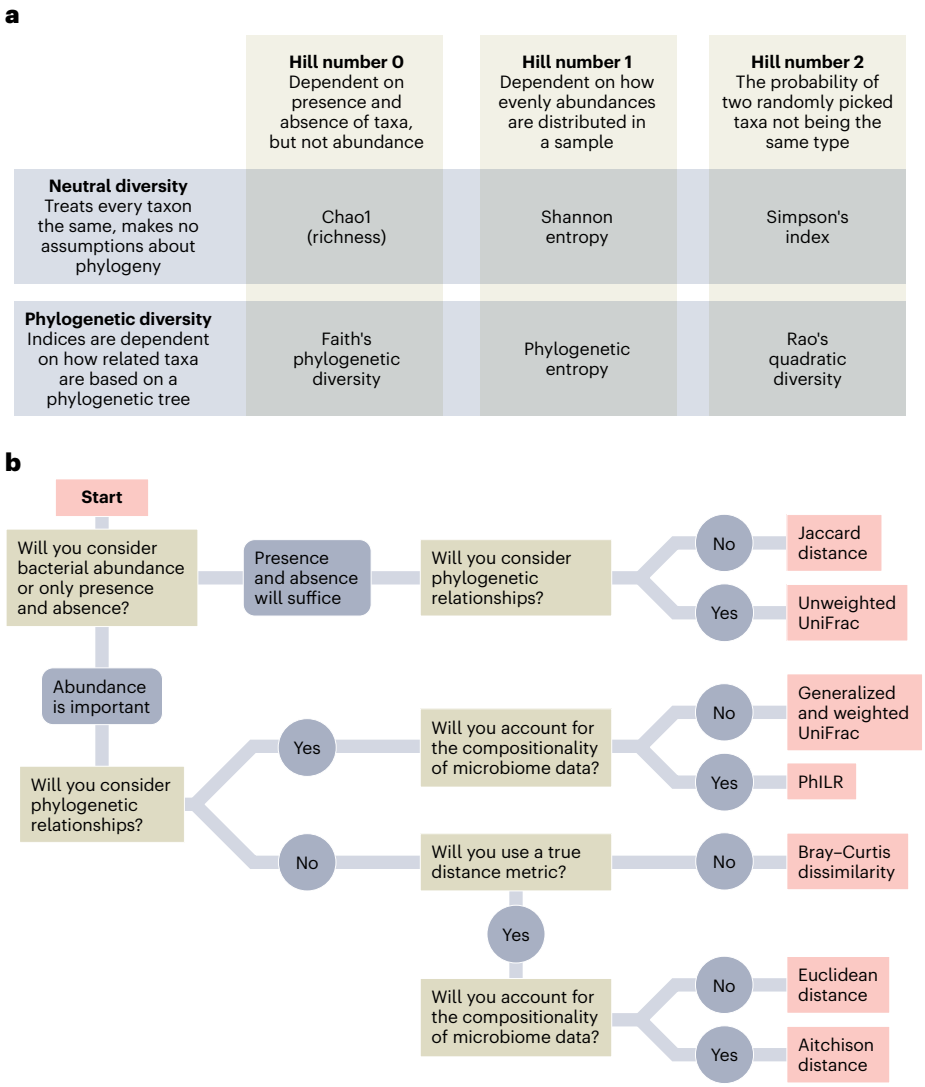

Fig. 2 | Understanding alpha and beta diversity. a, Alpha diversity metrics are related to each other. Commonly used alpha diversity metrics in the microbiome field can be classified along two axes. Here we show the Hill number on the *x* axis and whether the index considers phylogeny on the *y* axis. **b**, Decision tree featuring common beta diversity indices. Some beta diversity indices are more suitable depending on the needs of the researcher. This decision tree recommends an index based on three common criteria: whether one wants to consider abundance, phylogeny and/or true distance.

abundance in your data. Owing to the count-based nature of sequencing data, low-abundance features are less likely to reach the limit of detection and come up as zeroes. It is easy to see how low count abundance, perhaps due to low sequencing depth, can artificially increase feature variance[36]. In other words, the absence of evidence is not the evidence of absence.

Sometimes, one might wish to filter out features based on other metrics, such as variance[37,38]. However, features should not be filtered based on their association with an outcome, as this could bias the test statistics and resulting *P*-value estimates in downstream statistical tests. The total number of observations recorded for each sample in a feature table depends on the sequencing depth of the assay. Rarefaction is the practice of randomly removing observations from a sample until all samples have the same amount of observations. However, it has been described as an unnecessary and potentially counterproductive measure[39]. It is more conventional now to address inter-sample differences in sequencing depth through effective library size normalization or log-ratio transformation[40,41]. One notable exception is diversity analysis as discussed below[42–44]. We argue that, while rarefaction is sometimes justifiable and even recommended, rarefaction should not be seen as the default approach.

Linking the microbiome to host features

Diversity indices

The microbiome is a complex ecosystem. The analysis and visualization of the microbiome can be qualitatively distinct from other high-throughput sequencing data. Although the data arise from a molecular biology assay, several of the statistical approaches used in microbiome analysis originate from other fields, such as ecology. This makes microbiome science a clear beneficiary of interdisciplinary research. Diversity, as popularized in ecology, is a way to quantify and understand variation in microbiome samples. Classically, diversity is separated into two main related types: alpha diversity and beta diversity[45]. Alpha diversity refers to the degree of variation within a sample. Beta diversity refers to the degree of variation between samples. See Fig. 2 for an overview of the different alpha- and beta diversity indices.

We expand on statistical considerations with diversity metrics, including applying statistical models to estimate the effects of

variables on diversity, in Supplementary Information Sections 2.1 and 2.2, respectively.

Alpha diversity. Alpha diversity refers to the degree of complexity within a single microbiome sample. Many different but related alpha diversity indices exist, but their relation is unclear from their names. This can make the underlying principles confusing to understand. It is helpful to classify alpha diversity measures along two axes: the Hill number (0, 1 or 2) and whether it is phylogenetic (yes or no). Regarding the first axis, alpha diversity measures can be understood as being the result of a unifying equation in which a single parameter—called the Hill number—acts to vary the meaning of the equation, and thus define the alpha diversity measure. Every number gives a different alpha diversity metric. In practice, three Hill numbers are most often used: 0, 1 and 2. The number 0 defines richness, or how many different features a sample has. The number 1 defines evenness, or how equally the features in a sample are represented (equivalent to Shannon entropy). The number 2 defines Simpson's index, or the probability that two features picked at random do not have the same name (as a probability, it is bounded by 0 and 1). Regarding the second axis, other phylogenetic-diversity measures, such as Faith's phylogenetic diversity, extend alpha diversity by taking into account the coverage of all features (for example, bacteria) on a phylogenetic tree. Typically, the more of the tree that is represented in a sample, the higher the diversity. Figure 2a illustrates a classification of several popular alpha diversity measures.

Beta diversity. Beta diversity refers to the degree of difference between two microbiomes. It is worth appreciating the assumptions and limitations that come with describing the total difference between two complex ecosystems as a single number. There are many ways to measure the 'difference' between two samples, and each one imparts a unique perspective on the data. In principle, one could use any dissimilarity or distance measure. Three common difference measures are:

- Jaccard's index. This is a similarity measure that simply describes the proportion of unique taxa that are shared between two samples, without taking abundance into account. As such, one could interpret Jaccard's index as the fraction of unique taxa (not abundances) shared by two samples. If two samples have exactly the same microorganism taxa, the Jaccard index will be 1. In the case that two samples share no microbe taxa, the Jaccard index will be 0. Subtracting Jaccard's index from 1 makes it the Jaccard distance measure.
- Euclidean distance. This is the geometric distance derived by applying the Pythagorean theorem, using every microorganism as a separate dimension. It is computed by taking (the square root of) the sum of the squared differences in bacteria abundance. As in geometry, the minimum Euclidean distance is 0 while the maximum is unbounded. Euclidean distance satisfies the triangle inequality, making it useful for certain geometric analyses, such as volatility analysis as discussed below. A related measure called Aitchison distance is the Euclidean distance between log-ratio-transformed data. This distance has a favorable property known as sub-compositional dominance (that is, the removal of a taxa feature will never make two samples appear further apart) and is also equivalent to taking the Euclidean distance between all pairwise log-ratios[40].
- Bray–Curtis dissimilarity. This dissimilarity measure is similar to Jaccard's index in that it ranges from 0 to 1, while also being similar to Euclidean distance in that it is computed from the differences between abundances. Bray–Curtis is calculated by summing the difference in abundance between each microbial taxon, and dividing it by the total microbial abundance of the two samples. Thus, one could interpret Bray–Curtis as the fraction of abundances unshared by two samples (compare with Jaccard distance, which is the fraction of unique taxa unshared by two samples).

The three common difference measures listed above make use of bacteria presence or abundance without considering the phylogenetic relationship between the bacteria. Just as we can make alpha diversity phylogenetic, we can do the same with beta diversity.

- UniFrac. This distance makes use of phylogenetic information to measure the difference between samples. There are (at least) two types. The unweighted UniFrac distance considers the branch lengths of the phylogenetic tree along with microbial presence, and is defined as the sum of branch lengths unshared between the samples divided by the sum of branch lengths present in either sample. This measure has some analogy to Jaccard distance in that an unweighted UniFrac distance of 1 means the two samples share no bacteria taxa in common. The weighted UniFrac distance further considers microbial abundance, and weighs each branch length in the unweighted UniFrac formula by per-sample proportional abundances.
- PhILR. This method uses a log-ratio transformation called the isometric log-ratio (ILR) transformation, which uses a phylogenetic tree to recast the microbiome variables as a series of log-contrasts called 'balances'[46]. PhILR offers two weighting options called taxon weighting and branch weighting. When both are disabled, the PhILR beta diversity is equivalent to Aitchison distance, although its use of phylogeny-based coordinates may yield a more interpretable ordination of the data. The taxon weighting provides a compositionally robust alternative to weighted Jaccard or Bray–Curtis measures, while the branch weighting provides a compositionally robust alternative to UniFrac measures.

Figure 2b illustrates a decision tree that we as the authors use when selecting a beta diversity measure. As with alpha diversity, it is sometimes helpful to compare and contrast the results from multiple measures of beta diversity.

Differential feature abundance

In a nutshell, differential abundance analysis refers to the practice of sequentially testing whether each individual feature (gene, taxon, function and so on) of a feature table is different based on our phenotypes, groups or other metadata. Differential abundance is perhaps one of the most popular microbiome analyses. Like alpha and beta diversities, there are many approaches to measuring differential abundance. In some cases, the features are deconfounded first by regressing out the selected covariates (that is, taking the residuals of a model fitted on the selected covariates; also see 'Considering potential confounding factors' section). Most methods follow the same general pattern.

- Apply a transformation to correct for variation in sequencing depth, compositionality and/or other biases (Box 1).
- Perform a univariate statistical test for each taxon as a dependent variable with the sample metadata as predictors, for instance, by fitting a linear model. This is also a common time to account for confounders by including them as covariates.
- Adjust the *P* values for multiple testing, for example, using Bonferroni, Storey's *q* value, or Benjamini–Hochberg. We expand on multiple testing corrections in the companion piece to this paper[25].

Although this pattern is common for differential abundance analysis, many packages and tools exist to assess differential abundance. Recently, there has been an effort to compare and contrast these tools[47]. There is a striking heterogeneity in the performance of differential abundance tools. We recommend compositional methods combined with linear modeling as a safe, consistent and well-understood default

approach (Box 1). Note that while most differential abundance software do treat taxa as the dependent variables, one could just as well treat taxa as the predictors, as routinely done in machine learning applications.

Functions. While differential abundance is most commonly used on taxonomic count data, it often makes more biological sense to investigate whether any particular functions rather than taxa in the microbiome can explain a phenotype. The way in which one gets to the functional feature table depends on the type of sequencing. In the case of 16S, Piphillin and PICRUSt2 are two options. Both of these tools infer what the metagenome of a sample might look like by mapping 16S sequences to a functional database (for example, Kyoto Encyclopedia of Genes and Genomes (KEGG) or MetaCyc) of marker genes from both fully sequenced microbial genomes and metagenome-assembled genomes (MAGs), then inferring a functional feature table based on the functions present in the reference genomes. In the case of shotgun sequencing, there are two main strategies:

- If we mapped metagenomic shotgun-sequencing reads to a database of reference genomes, the full microbial sequences are already available and we only need to identify genes and collate them, for example, with the same KEGG or MetaCyc database. Tools such Woltka or HUMAnN3 in the biobakery suite are typically used to generate a functional feature table for shotgun data.
- If we rather generate MAGs, genes need to be detected using a program such as Prodigal[48] or Bakta[49] and annotated using a database such as eggNOG, KEGG or MetaCyc. For a user-friendly end-to-end pipeline to generate and annotate MAGs from metagenomic shotgun-sequencing reads, we refer the reader to Metagenome-ATLAS[50].

Regardless of methodology, functional tables tend to contain a large amount of features. Frameworks such as the functional gut–brain modules and gut–metabolic modules encompass microbial pathways that cover specific functional aspects of the microbiome such as the potential to produce neuroactive compounds and to metabolize specific substrates[51,52]. These frameworks enable the interrogation of specific aspects of the microbiome–gut–brain axis.

Functional inference versus annotation. Typically, functions are said to be inferred for 16S and assigned for shotgun sequencing. Strictly speaking, there is inference in both cases. However, in comparing 16S with shotgun sequencing, 16S functional inference can be thought of as a much bigger inferential leap than shotgun. With 16S, we have to first guess the entire genomic content based on a single sequence before inferring function, rather than inferring function directly from the reconstructed or mapped-to genome (as done in shotgun sequencing). This difference is so large that often inference is said to happen only with 16S. In both cases, a functional analysis is constrained by the validity and completeness of the functional database used to assign functional importance to the taxa or genes. A multitude of functional databases is currently available, with KEGG, UniRef90 and MetaCyc being among the most common general ones. Specialized databases, such as those covering antibiotic-resistance genes (ARDB and CARD) or carbohydrate metabolism (CAZy) are also available. These databases all have their own focus and frequently have different and sometimes unclear definitions of what a function entails and how a sequence is assigned to a function. While many databases are largely compatible, converting functional IDs between databases often requires some degree of manual curation, to the detriment of the field. Like taxonomic databases, functional databases are updated frequently and results may be affected as a consequence. Biologically speaking, the functional microbiome is known to be more consistent between hosts than the taxonomic microbiome, meaning that the results of functional analyses might generalize better[53]. As databases expand over time, it is important to report the version number as part of the methods.

Beyond the foundations

The design of host–microbiome experiments and the analysis and interpretation of the resulting data can be a daunting task. In this Perspective, we set out to highlight and explain the foundational concepts to enable the reader to navigate and avoid the most common pitfalls. We have provided and referenced the tools for the reader to customize their own analysis. We do not claim this approach is the only reasonable way to perform microbiome analysis, only that it is a reasonable one. In general, host–microbiome studies would benefit from reporting a characterization of the microbiome data in terms of alpha diversity, beta diversity and the general microbial composition using stacked bar plots or similar. During the statistical analysis of microbiome data, including modeling, correlating, and differential abundance testing, it is important to consider the compositional nature of microbiome data, for example, by first performing a centered log-ratio (CLR) transformation, and to account for the large number of tests performed, for example, by performing the Benjamini–Hochberg procedure. In many cases, studies would also benefit from considering the functional potential of the microbiome rather than limiting analysis to the level of taxonomy.

Microbiome analysis involves techniques and theory from a wide array of fields, including molecular biology, genetics, ecology, and even mathematical geology. In the case of fields assessing host–microorganism interactions, such as the microbiome–gut–brain axis field, expertise from additional fields such as immunology, psychology, psychiatry, pharmacology, neuroscience, and nutrition becomes an additional requirement. In the companion paper, we present multidisciplinary techniques from across these and other fields to enrich and extend the microbiome–gut–brain axis field[25]. Microbiome research is resource intensive, and there is a strong imperative to 'make the most' of collected data via innovative and robust analytic methods. We acknowledge the extensive body of knowledge and expertise required for comprehensive analysis of the microbiome—including but not limited, to the sheer number of specialist software tools needed for preprocessing and analysis. In this, and the following Perspective piece, we strive to assist by showcasing some of these so that the reader and microbiome analyst from any scientific background can navigate the landscape with greater confidence.

This Perspective piece is not without limitations. We acknowledge that no single Perspective can comprehensively cover the entire field of microbiome bioinformatics analysis. Indeed, we chose not to cover in silico metabolic modeling of the microbiome. We also focus on metagenomic analysis of the microbiome, rather than other methods to interrogate the microbiome such as metabolomic, metaproteomics, or metatranscriptomics.

The microbiome field is currently undergoing a phase of rapid growth and development. We anticipate that new tools, databases and approaches will slowly replace the current suite. In particular, we anticipate a movement towards more longitudinal experimental designs, as well as the integration of multiple omics approaches on the same microbiome to more clearly capture the metabolic functional capacity of the microbiome. For some of these extension topics, please refer to the accompanying paper[25]. In terms of statistical analysis, we encourage the ongoing adoption and further development of CoDA-oriented methodology. Open, freely available, and well-documented resources such as Bioconductor and CRAN are, and will continue to be, essential for the development and adoption of new bioinformatics tools and pipelines in the microbiome field, as well as the broader scientific community. Of particular note, we draw attention to the well-curated and well-maintained curatedMetagenomicData package on Bioconductor, which allows researchers to access large human microbiome datasets in feature table form, ready for analysis[54].

Lastly, we highlight that one of the major strengths of science is to build on the previous findings of others. Often in the microbiome field, a large amount of data is gathered and a broad analysis is performed, perhaps linking the microbiome to a host condition. Whenever possible, raw sequencing data paired with anatomized metadata required to reproduce analyses should be made publicly available as it is essential for reliable and robust meta-analyses. Indeed, many journals require researchers to deposit their raw sequences to nucleotide repositories such as the European Nucleotide Archive (ENA), Sequence Read Archive (SRA) and the DNA Data Bank of Japan Sequence Read Archive (DRA). While these types of large-scale studies with broad hypotheses remain valuable to map out the interplay between the microbiome and host, we argue that the field is ready to move towards hypothesis-driven experiments with the intent to uncover specific mechanisms and even tractable aspects of the microbiome to help improve our understanding of both the microbiome and our general health and well-being. These types of specific hypothesis should be formed based on observations from the large-scale exploratory studies and would also benefit from having bioinformaticians and biostatisticians present during the design stage.

References

1. Sender, R., Fuchs, S. & Milo, R. Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol*. **14**, e1002533 (2016).
2. Tierney, B. T. et al. The landscape of genetic content in the gut and oral human microbiome. *Cell Host Microbe* **26**, 283–295 (2019).
3. Peterson, J. et al. The NIH Human Microbiome Project. *Genome Res*. **19**, 2317–2323 (2009).
4. Claesson, M. J. et al. Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488**, 178–184 (2012).
5. Consortium, H. M. P. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–14 (2012).
6. Tigchelaar, E. F. et al. Lifelines deep, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).
7. Integrative, H. et al. The Integrative Human Microbiome Project. *Nature* **569**, 641–648 (2019).
8. Cryan, J. F. et al. The microbiota–gut–brain axis. *Physiol. Rev.* **99**, 1877–2013 (2019).
9. Maier, L. et al. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623–628 (2018).
10. Tomizawa, Y. et al. Effects of psychotropics on the microbiome in patients with depression and anxiety: considerations in a naturalistic clinical setting. *Int. J. Neuropsychopharmacol.* **24**, 97–107 (2020).
11. Maini Rekdal, V., Bess, E. N., Bisanz, J. E., Turnbaugh, P. J. & Balskus, E. P. Discovery and inhibition of an interspecies gut bacterial pathway for levodopa metabolism. *Science* **364**, eaau6323 (2019).
12. Scassellati, C. et al. The complex molecular picture of gut and oral microbiota–brain-–depression system: what we know and what we need to know. *Front. Psychiatry* **12**, 722335 (2021).
13. Simpson, C. A. et al. Oral microbiome composition, but not diversity, is associated with adolescent anxiety and depression symptoms. *Physiol. Behav.* **226**, 113126 (2020).
14. Sureda, A. et al. Oral microbiota and Alzheimer's disease: do all roads lead to rome? *Pharmacol. Res.* **151**, 104582 (2020).
15. Tonelli, A., Lumngwena, E. N. & Ntusi, N. A. The oral microbiome in the pathophysiology of cardiovascular disease. *Nat. Rev. Cardiol.* **20**, 386–403 (2023).
16. Martinez-Guryn, K. et al. Small intestine microbiota regulate host digestive and absorptive adaptive responses to dietary lipids. *Cell Host Microbe* **23**, 458–469 (2018).
17. Kastl Jr, A. J., Terry, N. A., Wu, G. D. & Albenberg, L. G. The structure and function of the human small intestinal microbiota: current understanding and future directions. *Cell. Mol. Gastroenterol. Hepatol.* **9**, 33–45 (2020).
18. Kupferschmidt, K. More and more scientists are preregistering their studies. Should you. *Science Magazine* https://doi.org/10.1126/science.aav4786 (2018).
19. Munafó, M. R. et al. A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).
20. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl Acad. Sci. USA* **115**, 2600–2606 (2018).
21. Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. & Kievit, R. A. An agenda for purely confirmatory research. *Perspect. Psychol. Sci.* **7**, 632–638 (2012).
22. Schloss, P. D. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *mBio* **9**, e00525-18 (2018).
23. Allen, C. & Mehler, D. M. Open science challenges, benefits and tips in early career and beyond. *PLoS Biol*. **17**, e3000246 (2019).
24. Mirzayi, C. et al. Reporting guidelines for human microbiome research: the STORMS checklist. *Nat. Med.* **27**, 1885–1892 (2021).
25. Bastiaanssen, T. F. S., Quinn, T. P. & Loughman, A. Bugs as features (part 2): a perspective on enriching microbiome–gut–brain axis analyses. *Nat. Ment. Health* https://doi.org/10.1038/s44220-023-00149-2 (2023).
26. Ferdous, T. et al. The rise to power of the microbiome: power and sample size calculation for microbiome studies. *Mucosal Immunol*. **15**, 1060–1070 (2022).
27. Dong, T. S. & Gupta, A. Influence of early life, diet, and the environment on the microbiome. *Clin. Gastroenterol. Hepatol.* **17**, 231–242 (2019).
28. Wilson, A. S. et al. Diet and the human gut microbiome: an international review. *Dig. Dis. Sci.* **65**, 723–740 (2020).
29. Yap, C. X. et al. Autism-related dietary preferences mediate autism–gut microbiome associations. *Cell* **184**, 5916–5931 (2021).
30. Hernán, M. A., Hsu, J. & Healy, B. A second chance to get causal inference right: a classification of data science tasks. *Chance* **32**, 42–49 (2019).
31. Zhu, F. et al. Metagenome-wide association of gut microbiome features for schizophrenia. *Nat. Commun.* **11**, 1612 (2020).
32. Ponsonby, A.-L. Reflection on modern methods: building causal evidence within high-dimensional molecular epidemiological studies of moderate size. *Int. J. Epidemiol.* **50**, 1016–1029 (2021).
33. Dillies, M.-A. et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **14**, 671–683 (2013).
34. McLaren, M. R., Willis, A. D. & Callahan, B. J. Consistent and correctable bias in metagenomic sequencing experiments. *eLife* **8**, e46923 (2019).
35. Nearing, J. T., Comeau, A. M. & Langille, M. G. Identifying biases and their potential solutions in human microbiome studies. *Microbiome* **9**, 113 (2021).
36. Erb, I. Power transformations of relative count data as a shrinkage problem. *Inf. Geom.* **6**, 327–354 (2023).
37. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).
38. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
39. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10**, e1003531 (2014).

40. Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. & Pawlowsky-Glahn, V. Logratio analysis and compositional distance. *Math. Geol.* **32**, 271–275 (2000).
41. Quinn, T. P., Erb, I., Richardson, M. F. & Crowley, T. M. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics* **34**, 2870–2878 (2018).
42. Hsieh, T. C. & Chao, A. Rarefaction and extrapolation: making fair comparison of abundance-sensitive phylogenetic diversity among multiple assemblages. *Syst. Biol.* **66**, 100–111 (2017).
43. McKnight, D. T. et al. Methods for normalizing microbiome data: an ecological perspective. *Methods Ecol. Evol.* **10**, 389–400 (2019).
44. Willis, A. D. Rarefaction, alpha diversity, and statistics. *Front. Microbiol.* **10**, 2407 (2019).
45. Sepkoski, J. J. Alpha, beta, or gamma: where does all the diversity go? *Paleobiology* **14**, 221–234 (1988).
46. Silverman, J. D., Washburne, A. D., Mukherjee, S. & David, L. A. A phylogenetic transform enhances analysis of compositional microbiota data. *eLife* **6**, e21887 (2017).
47. Nearing, J. T. et al. Microbiome differential abundance methods produce different results across 38 datasets. *Nat. Commun.* **13**, 342 (2022).
48. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
49. Schwengers, O. et al. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb. Genom.* **7**, 000685 (2021).
50. Kieser, S., Brown, J., Zdobnov, E. M., Trajkovski, M. & McCue, L. A. ATLAS: a snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC Bioinform.* **21**, 257 (2020).
51. Valles-Colomer, M. et al. The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat. Microbiol.* **4**, 623–632 (2019).
52. Vieira-Silva, S. et al. Species–function relationships shape ecological properties of the human gut microbiome. *Nat. Microbiol.* **1**, 16088 (2016).
53. Mehta, R. S. et al. Stability of the human faecal microbiome in a cohort of adult men. *Nat. Microbiol.* **3**, 347–355 (2018).
54. Pasolli, E. et al. Accessible, curated metagenomic data through experimenthub. *Nat. Methods* **14**, 1023–1024 (2017).
55. Bharti, R. & Grimm, D. G. Current challenges and best-practice protocols for microbiome analysis. *Brief. Bioinform.* **22**, 178–193 (2019).
56. Szóstak, N. et al. The standardisation of the approach to metagenomic human gut analysis: from sample collection to microbiome profiling. *Sci. Rep.* **12**, 8470 (2022).
57. Clooney, A. G. et al. Comparing apples and oranges? Next generation sequencing and its impact on microbiome analysis. *PLoS ONE* **11**, e0148028 (2016).
58. Valles-Colomer, M. et al. The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* **614**, 125–135 (2023).
59. Aitchison, J. On criteria for measures of compositional difference. *Math. Geol.* **24**, 365–379 (1992).
60. Hillmann, B. et al. Evaluating the information content of shallow shotgun metagenomics. *mSystems* **3**, e00069-18 (2018).
61. Santiago-Rodriguez, T. M. et al. Metagenomic information recovery from human stool samples is influenced by sequencing depth and profiling method. *Genes* **11**, 1380 (2020).
62. de Goffau, M. C., Charnock-Jones, D. S., Smith, G. & Parkhill, J. Batch effects account for the main findings of an in utero human intestinal bacterial colonization study. *Microbiome* **9**, 6 (2021).
63. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B* **44**, 139–160 (1982).
64. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* **8**, 2224 (2017).
65. Calle, M. L. Statistical analysis of metagenomics data. *Genom. Inform.* **17**, e6 (2019).
66. Hill, M. O. Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**, 427–432 (1973).
67. Pearson, K. Mathematical contributions to the theory of evolution—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. R. Soc. Lond.* **60**, 489–498 (1897).
68. Quinn, T. P., Richardson, M. F., Lovell, D. & Crowley, T. M. propr: an R-package for identifying proportionally abundant features using compositional data analysis. *Sci. Rep.* **7**, 16252 (2017).
69. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
70. Kurtz, Z. D. et al. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, e1004226 (2015).
71. Lubbe, S., Filzmoser, P. & Templ, M. Comparison of zero replacement strategies for compositional data with large numbers of zeros. *Chemometr. Intel. Lab. Syst.* **210**, 104248 (2021).

Acknowledgements

We thank A.-L. Ponsonby for her expert comments on directed acyclic graphs, D. L. Dahly for his insights on statistical analysis and J. F. Cryan for his continued encouragement and excellent advice. APC Microbiome Ireland is a research center funded by Science Foundation Ireland (SFI), through the Irish Governments' national development plan (grant no. 12/RC/2273_P2).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44220-023-00148-3.

Correspondence should be addressed to Thomaz F. S. Bastiaanssen.

Peer review information *Nature Mental Health* thanks the anonymous reviewers for their contribution to the peer review of this work.