Jurburg *et al. Microbiome* (2022) 10:225
https://doi.org/10.1186/s40168-022-01423-8

Microbiome

COMMENT

Open Access

The community ecology perspective of omics data

Stephanie D. Jurburg[1,2,3*], François Buscot[2,4], Antonis Chatzinotas[1,2,3], Narendrakumar M. Chaudhari[2,5], Adam T. Clark[6], Magda Garbowski[2,7], Matthias Grenié[2,3], Erik F. Y. Hom[2,8], Canan Karakoç[1,2,9], Susanne Marr[2,10,11], Steffen Neumann[2,11], Mika Tarkka[2,4], Nicole M. van Dam[2,5,12], Alexander Weinhold[2,5] and Anna Heintz-Buschart[13]

Abstract

The measurement of uncharacterized pools of biological molecules through techniques such as metabarcoding, metagenomics, metatranscriptomics, metabolomics, and metaproteomics produces large, multivariate datasets. Analyses of these datasets have successfully been borrowed from community ecology to characterize the molecular diversity of samples (α-diversity) and to assess how these profiles change in response to experimental treatments or across gradients (β-diversity). However, sample preparation and data collection methods generate biases and noise which confound molecular diversity estimates and require special attention. Here, we examine how technical biases and noise that are introduced into multivariate molecular data affect the estimation of the components of diversity (i.e., total number of different molecular species, or entities; total number of molecules; and the abundance distribution of molecular entities). We then explore under which conditions these biases affect the measurement of α- and β-diversity and highlight how novel methods commonly used in community ecology can be adopted to improve the interpretation and integration of multivariate molecular data.

Keywords: Multivariate statistics, Molecular ecology, Community ecology

One of the most fundamental patterns of scientific discovery is the revolution in thought that accompanies a new body of data [1].

Introduction

The direct characterization and analysis of sampled pools of biomolecules, particularly DNA, RNA, proteins, and metabolites, has fundamentally altered the life sciences and specifically microbiology. Metagenomics and metabarcoding can identify organisms in a sample and detect inter- and intraspecific diversity, while (meta-)transcriptomics, (meta-)proteomics, and metabolomics can characterize the functional responses of biological individuals, populations, or whole communities (Table 1). Despite their different targets, these techniques measure molecular entities in a high-throughput, high-data volume manner and produce similar multivariate data outputs, enabling multivariate, molecular ecology (heretofore MME).

MME techniques measure a wide range of molecular entities within a selected class of molecules. Being generally untargeted, they require less *a priori* knowledge about the biomolecules measured than targeted assays. Interest in MME techniques has grown rapidly over the past two decades, and their application has informed ecological and evolutionary theory. For example, metabarcoding has been used to show how dormancy affects the spatial distribution of microbes [11], and metatranscriptomic analyses have revealed that microbial niche

*Correspondence: s.d.jurburg@gmail.com

[3] Institute of Biology, Leipzig University, Leipzig, Germany
Full list of author information is available at the end of the article

Table 1 MME techniques yield data sets with common structures, and often, limitations

	Common techniques	0-inflated	No N	Compositional
Genomics: The system-wide identification and quantification of DNA sequences and the encoded functions in an organism or population [2].	High throughput sequencing		(+)	
Transcriptomics: The system-wide identification and quantification of the RNA transcripts in an organism or population [3].	High throughput sequencing, microarrays	+	+	+
Proteomics: The use of quantitative protein-level measurements of gene translation to characterize biological processes and decipher the mechanisms of gene expression control [4].	Mass spectrometry	+	+	
Metabolomics: The systematic identification and quantification of metabolites (small molecule substrates, intermediates, products of cell metabolism) in an organism or population [5]..	Nuclear magnetic resonance spectroscopy, mass spectrometry	+	(+)	
Metabarcoding: The large-scale identification and quantification of variation of diversity in an environmental sample in terms of a specific genomic region (DNA) [6].	High throughput amplicon sequencing	++	+	+
Metagenomics: Large-scale identification and quantification of all DNA in an environmental sample [7].	High throughput shotgun metagenomic sequencing	++		+
Metatranscriptomics: Large-scale identification and quantification of all RNA transcripts in an environmental sample [8]	High throughput RNA sequencing, (microarrays)	++		+
Metaproteomics: Large-scale identification and quantification of the entire protein complement from an environmental sample [9].	Mass spectrometry	++	+	
Meta-metabolomics: Large-scale identification and quantification of small molecules from an environmental sample [10].	Nuclear magnetic resonance spectroscopy, mass spectrometry	+	+	

For techniques with "no N," the total number of molecular entities measured contains no biological information. For techniques that produce 0-inflated datasets, the data matrices contain more zeros than non-zero values, while for compositional datasets, the abundance of species is correlated to the technique and contains no biological information. For each limitation, whether it is an issue or a serious issue for each data type is indicated with + or ++, respectively. Data types that face a limitation but it seldom affects the scientific questions asked with these data are indicated with (+)

differentiation in the rhizosphere is both spatially and temporally regulated [12]. Similarly, a combination of proteomics and targeted metagenomics have been used to study adaptation in light of horizontal gene transfer [13].

MME data and conventional community ecology (CE) data are both multivariate and the *sample* × *molecule* matrices produced by MME techniques may be viewed as analogous to the *plot* × *species* matrices produced by community ecologists. Like in community ecology, studies in molecular and microbial ecology are often less concerned with the response of single molecular entities, especially when the identity and function of specific molecular entities are unknown. Rather, they focus on differences or changes in a whole profile or on emergent properties of the community (e.g., metabolic cooperation). Similar to CE studies, MME data are often used to characterize molecular diversity and the relative abundance of molecular entities within samples and across space, time, or experimental treatments.

α-and β-diversity metrics, which describe species diversity in a sample and the difference in diversity between samples, respectively, were developed by CE and are often used to describe changes in the molecular profiles of samples (e.g., Shannon diversity indices, Bray-Curtis dissimilarities; [14–16]). In CE, diversity metrics rely on three key measures [17, 18] (Fig. 1): the absolute numbers of individuals detected in a specific area (N), the total number of species in that area or species richness (S), and the relative abundances of community members (i.e., species abundance distributions, or SAD).

CE measures are affected by biases that are introduced during the molecular sample preparation, molecular analyses, data collection, and processing steps that generate MME data. The techniques that produce MME data are relatively new, and while much work has been devoted to identifying their biases and limitations, how these affect the applicability of CE measures to assess patterns of molecular diversity has received less attention. However, within the context of ecology, the biases and limitations of diversity data have been extensively discussed, resulting in various approaches to overcome them. For example, issues related to detection limits given finite sample sizes are potentially the oldest analytical research topic in

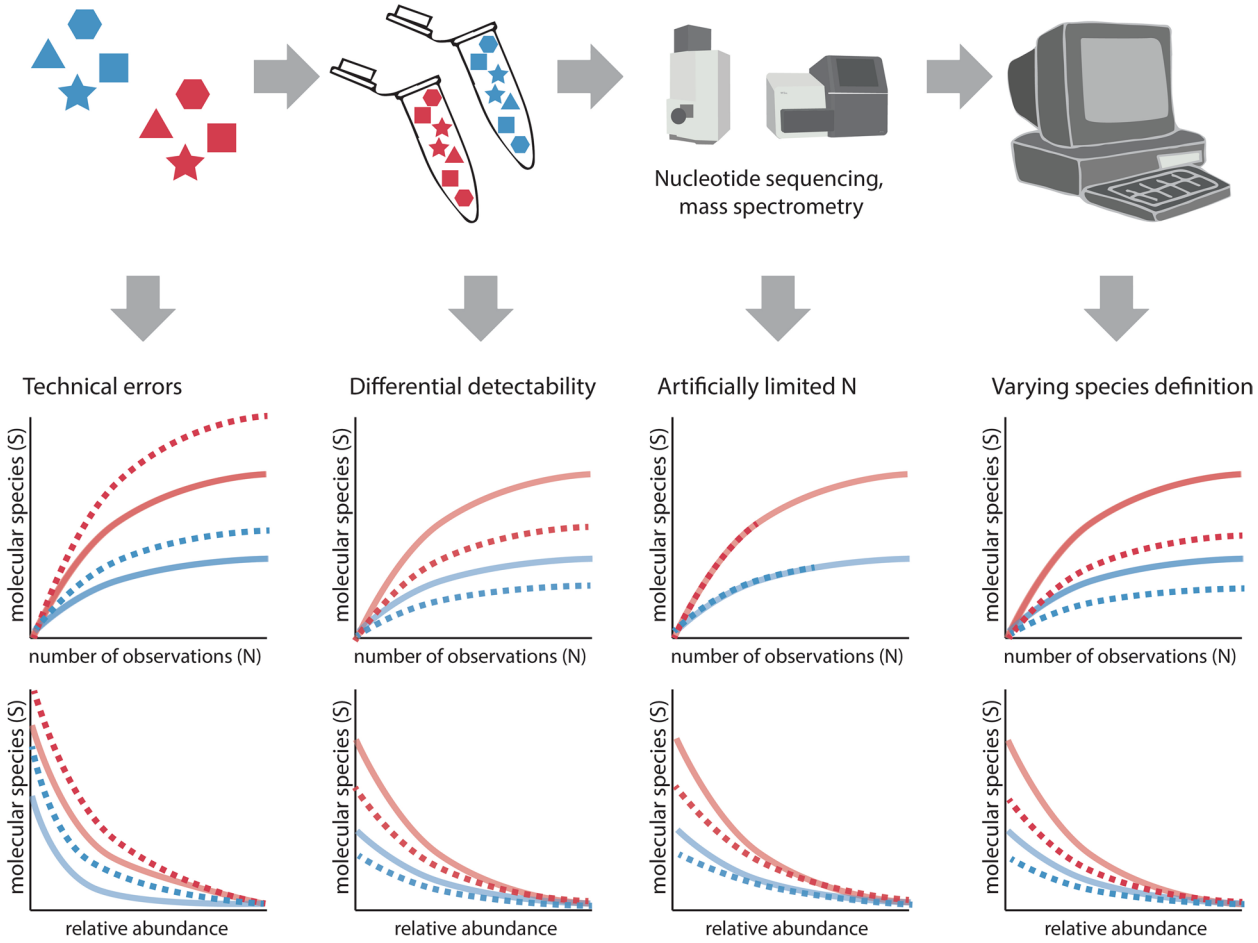

Fig. 1 Sample collection and preparation, data collection, and post-processing are inextricably linked in MME techniques and can all potentially affect estimates of diversity. The effect of researcher choices on S, N (middle row), and SAD (bottom row) during data generation is shown below each step. The true diversity in two samples is shown in red and blue, and the measured diversity is shown as dotted lines. Technical errors during sample collection and storage can increase S (e.g., due to non-specific contamination [19]), resulting in higher estimates for S and steeper SAD (**a**). In contrast, sample preparation can reduce the detectability of certain molecular entities (e.g., during PCR amplification in metabarcoding [20]) resulting in a lower S and flatter SAD (**b**). Technical limitations on the number of observations are imposed by some of the data collection instruments used (e.g., sequencer), placing technical limits on N, and potentially resulting in more even communities (i.e., a flatter SAD) (**c**). During processing, applying a less stringent species definition can result in reduced S and a flatter SAD (**d**)

ecology [21] and have led to the development of methods that can quantify sample completeness (i.e., rarefaction, diversity index estimators [22–24]). Importantly, methods from molecular data analysis and ecology can be used to analyze different kinds of MME data with their specific challenges (Table 1), but they make different assumptions about the underlying distribution of the data [25]. Consequently, their applicability depends both on the raw data and the scientific questions [26].

Here, we explore how detection biases and technical limitations associated with MME data affect the calculation of N, S, and SAD, and how this, in turn, affects the determination of α- and β-diversity. Finally, drawing parallels from other areas of community ecology, we indicate analytical approaches that allow for the robust estimation of diversity metrics for biased or limited MME data.

The components of biodiversity in the context of MME data

Selection of sampling scale and species definition

All components of biodiversity—and the patterns emerging from them—depend on the sampling scale [27, 28]. There is no single "perfect" scale which is optimal for measuring biodiversity, and the relative importance of

different ecological processes is intricately linked to the scale of observation [29]. However, defining the scale of observation in MME analyses is not straightforward, and altering this scale is not always possible. First, the data collection method used (e.g., mass spectrometer, sequencer) dictates how samples must be prepared and indirectly sets limits on the scale of sampling. In low biomass samples, the lower limits can be hard to achieve, leading to biases due to contaminants or artifacts [19]. On the other hand, upper limits can be very small (e.g., µL of a liquid or mg of a solid), so sample preparation must be adapted for the analytical sample to adequately represent the desired scale. Second, during each step of sample preparation, the substrate obtained in the previous procedure is often subsampled, which may further dissociate diversity assessments from the initial sampling scale. Third, sampling strategies that are common in MME techniques can have poorly defined scales (e.g., filtrate, swabs), which complicates the comparison diversity in samples obtained with even slight variations in sampling approaches. Incomparability of sampling scales cannot always be avoided (e.g., if body sites or plant tissues are being compared [30, 31]). The mismatch between the scale of observation and the scale at which the target organisms perceive and function in their environment has received substantial attention in ecology [28, 32] and has recently been discussed in the context of microbes [33], but further research into the spatial scale of influence on MME target molecules is necessary.

The spatial scaling of diversity has been explored in metabarcoding data [34–37], but further research is urgently needed to describe the spatial components of complex samples used for other kinds of MME data, while explicitly acknowledging that scaling up observations is not always possible for MME data and may be constrained by the system of interest (e.g., host-associated systems) and by the technique used (i.e., the laboratory protocols). Additional research is also needed to determine whether small sample sizes limit the ability to detect phenomena that occur at larger scales or that depend on larger organisms [38].

In order to capture microbial diversity at larger scales and reduce the effect of small-scale spatial variability, pooling (i.e., homogenizing multiple smaller samples from a larger area) is a common practice [39]. However, the universal distance decay relationship (DDR, [40]) predicts that samples taken further away from each other in space and time will have less overlap in species, and this is also likely the case for MME data (e.g., sequence-based omics [41, 42]). The spatial distribution of pooled samples can therefore greatly affect (and when unstandardized, bias) diversity estimates [38], resulting in inflated diversity estimates for samples which were pooled from larger areas. Furthermore, MME techniques can target molecules from a wide range of organisms of varying body sizes, but body size affects the distribution of diversity [43–45]. Further research is needed to determine to what extent the DDR applies to MME data (e.g. ecometabolomics of larger organisms), and whether this has implications for sampling designs. For microbial communities, the relative similarity of functional compared to taxonomic profile has been demonstrated [46, 47], but whether DDRs are observable in functional microbial MME data has not been analyzed to date.

While MME techniques often impose limits on the sampling scale, they generally require the researcher to explicitly select a definition of the units of diversity. These units can be species or molecular entities, such as ASVs or protein families. If the definition is not given (as it is, e.g., for identified metabolites), it is usually based on a threshold of molecular similarity of detected molecules to each other or to references. Definitions vary with e.g., 97 to 100% similarity for metabarcoding [48], 95% average nucleotide identity for microbial genomes [49, 50], and a vast range of identity and expected value thresholds for functional units, such as gene families. Importantly, the choice of units and definition directly affects biodiversity measurements (Fig. 1). Within the context of synthesis, archiving raw MME data allows data reusers to reprocess datasets using a single-species definition, allowing cross-study comparisons.

The number of individuals (N)

In CE, for a given sampling scale, N measures the number of individuals in that space. However, MME techniques measure molecules rather than individuals. The two are not always related) e.g., because of differences in cell or body sizes, or in copy numbers of marker genes). Furthermore, the number of detected molecules often reflects a machine's throughput, rather than biological reality [51]. For example, in sequencing-based methods (i.e., (meta-)genomics and (meta-)transcriptomics), the number of observations or sequencing reads per sample (i.e., sequencing depth) reflects the choice of sequencer and the number of samples loaded, rather than the abundance of organisms in the sample [52]. As a consequence for MME data, N often serves only as an indicator of observation effort (e.g., sequencing depth) and is otherwise uninformative.

The decoupling of N and the abundance of molecules in situ creates two limitations. First, it precludes the estimation of the true abundances of molecular entities. Second, uneven observation depths make changes in the abundances of molecular entities (i.e., differential detection) sensitive to the normalization method used [53]. Several statistical approaches have been developed to normalize before detecting changes in

molecular entities across samples through generalized linear models, including DESeq2 [54], metagenomeSeq [55], edgeR [56], LEfSE [57], and voom [58]. However, several normalization methods assume similar SADs and they can bias the comparison of abundances across samples (e.g., in transcriptomic data; [53]). Given the wide range of options, the development of novel tools for comparing workflows and the resulting MME data (e.g., ANPELA in proteomics [59]) and studies comparing different approaches for differential detection [25, 60, 61] are increasingly important. For example, in LC-MS and metabolomics data, spectral data are normalized by external standards, pooled samples, or total biomass [62].

Species richness (S)

CE's species richness in an area (S) can be equated to the number of distinguishable molecular entities in a sample (i.e., molecular richness), which are often defined *ad hoc*, as discussed above. S can be influenced by the environment, dispersal, interactions among organisms, or changes in organisms producing molecules of interest (as in [17]). Assessing the number of molecular entities present is often of interest, with multiple studies having found connections between richness, mechanistic processes, and ecological phenomena. For example, metabarcoding-based assessments of soil bacterial communities revealed that bacterial richness was positively related to carbon decomposition and soil enzymatic activities [63]. Similarly, LC-MS-based assessments of the secondary metabolites of fungi revealed species-specific metabolic richness [64].

The detectability of molecular entities is not evenly distributed (Fig. 1), and this may negatively affect S (and SAD, see below). For all MME techniques, several hundred to thousands of molecular entities may be present in concentrations that span multiple orders of magnitude, which complicates their measurement. Nonrandom differences in the detectability of molecular entities may lead to the consistent underestimation of specific molecular entities, as is the case with biases caused by different primer affinities in metabarcoding [65], or to the underestimation of molecular entities below a certain abundance threshold. For example, the limited dynamic range of a mass spectrometer may cause rare molecular entities to fall below the limit of detection, while very abundant entities may saturate the detector, pushing them above the limit of quantification [66]. Differences in detectability can also be random and result in lower signal-to-noise ratios (e.g., in proteomics data [67]). Mathematical modeling of persistent detection biases has been proposed as a first step to identify where biases arise and to quantify them in metagenomic data [68], while latent variable modeling has been proposed for estimating missing values in proteomics data [69].

Species abundance distributions (SAD)

In CE, species abundance distributions (SAD) describe how abundances vary across species in a community (often expressed as "relative abundances," standardized by total biomass or the total number of individuals in the community [70]). SAD of all ecological communities exhibit a hollow shape, as some species are more abundant than others, but the shape can vary [70]. Flatter SAD indicate a more even community, while hollower SAD indicates strong dominance by certain species. Like community data, MME data generally have few very abundant molecular entities and many with low abundances (e.g., in metabarcoding [71]).

Because SAD are distributions rather than a single metric, comparing SAD across multiple samples is not straightforward, as differences in the relative distribution of taxa (i.e., evenness) can result in different relationships between samples depending on N [72]. This is important for the analysis of SAD, since N in MME tends to relate to technical choices rather than biological reality, and the different numbers of molecules per molecular species further confound the estimation of the number of organisms in the sample [38]. Nevertheless, N affects the other components of biodiversity: the greater the observation effort, the greater S will be, and the greater the number of rare molecular entities that will be found (i.e., a longer-tailed SAD).

When the number of observations is artificially determined by the technique (i.e., uninformative N), an increase in one species may result in an observed decrease of another, even if the absolute abundance of the other molecular entity is unchanged [73]. This limitation makes the data compositional [74], so the observed abundance of each species depends on the abundance of all other species in the sample [52], skewing SAD. Several methods to analyze compositional data [74–77], identify differentially abundant molecular entities [78, 79] or groups of molecular entities [51], and determine causality [80] have been proposed. In the most basic form, compositional analyses use log ratio transformations to individual values or the geometric mean of all values (clr). These were applied, for example, to metabarcoding and metalomic data [81]. However, MME data are often sparse (i.e., molecular entities appear seldom across samples) and zero-inflated (i.e., there are more zeros than expected from the distribution of the observed molecular entities in a sample, Table 1). Because of the large number of zeros, log transformations and ratios do not work [82]. Overcoming this limitation remains an area of active research: the simplest approach to removing zeros

is replacing zeros with fixed, low values that lie below the instrument's detection limit. However, this skews sparse data further [73]. Alternatively, methods that test pairwise ratios of specific molecular entities remove rare features (e.g., ANCOM [83], ANCOM-II [84]). More complex methods use Bayesian inference, but are computationally expensive (i.e., ALDEX2 [85]). For metabolomic data, testing the informative potential across all possible ratios of metabolite pairs is a common and statistically validated practice [86]. To model the relationship between molecular entities and environmental gradients, generalized joint attribute models adapt joint species distribution models to zero-inflated data [87] and are appropriate for MME data.

For many MME questions, rare molecular entities are uninformative, in addition to being more likely observed due to technical error, making them less quantifiable. MME fields have developed different approaches to remove such data. In untargeted metabolomics, molecular entities that are detected in blanks or exhibit high variation across technical replicates are removed from the metabolomic profile [62, 88]. In most modern metabarcoding pipelines, molecules that appear only once are assumed to be indistinguishable from sequencing errors and are removed automatically [89–91], artificially shrinking S [92], N, and resulting in a flatter SAD. However, this may be preferable to including erroneous data. For example, one study found that at most, 40% of molecular entities that appeared once (i.e., singletons) could potentially be artefactual, directly affecting species richness estimation [93]. In many analysis pipelines, rare taxa are commonly filtered to reduce the dimensionality of the data, regardless of the "correctness" of the observation. This is especially true for differential detection methods, which require a minimum abundance to detect differences among samples (e.g., ANCOM-II and DESeq2; [54, 84]).

How SAD curves are used in CE to derive multiple diversity estimates, including evenness and α-diversity for any N, as well as to compare abundance profiles across gradients [94] is discussed below.

Derived diversity components and diversity metrics for MME data

The challenge of analyzing "imperfect" MME data is mirrored by similar issues that commonly arise in CE studies [95]. For example, data from plant or intertidal rock communities are often estimated in terms of "relative cover" (that is, the area covered by each species, divided by the total area covered by any organism). The resulting data are often subject to high error rates, risk undercounting the abundance of rare species, and introduce major statistical challenges like zero inflation [96, 97]. Historically, these problems have severely limited the kinds of analyses that can be applied using cover data. Modern statistical methods partially address them, for example by separately tracking trends in abundance vs. presence/absence data [98], or by modeling the sources of zeroes (e.g., "true" absences vs. "false" absences; [99]).

Estimating the true N, S, and SAD

Issues related to detection limits given finite sample sizes are one of the oldest analytical research topics in ecology [21]. Rarefaction or diversity index estimators can help quantify how complete a particular sample is likely to be and, in some cases, to standardize data across different samples [22, 23]. These are commonly applied to metabarcoding and other sequence-based MME data [100], especially when the aim is to quantify diversity profiles. However, data loss associated with rarefaction remains controversial [101]. More recent methods for extrapolating diversity estimates based on abundance or presence/absence community data [24, 102] allow for the estimation of diversity in MME data without massive losses in data.

In CE, detection errors that result in missing data are often estimated by revisiting a site (i.e., in vegetation surveys), with marked capture-recapture experimental designs, or with *a posteriori* modeling [103]. Several approaches approximate the true S for DNA-based omics (reviewed in [104]), but such approximations have received less attention from other MME methods. Assuming that rare molecular entities (i.e., singletons) represent technical errors produced during sample processing and data collection [105], targeted methods can estimate the true diversity of these rare molecular entities based on the frequencies of more abundant entities [106, 107]. Similarly, diversity estimators such as the jackknife or Chao1 [108] derive an asymptotic diversity estimate based on the number or of singleton or doubleton entities.

Species occupancy models may be used to correct for false positive and false negative errors [109–112] and approximate the true S and SAD, but when the probability of detection is low, species occupancy models require large numbers of technical and biological replicates [113], which are often not possible. Approaches such as joint species distribution modeling and analyses of species-environment associations [114, 115] can be applied to predict the identities of potential missing species and to introduce *post hoc* corrections to incomplete samples. However, this approach draws heavily on *a priori* understanding about species' natural histories as well as historical data from previous studies, and may become more prominent as the molecular world is increasingly well characterized.

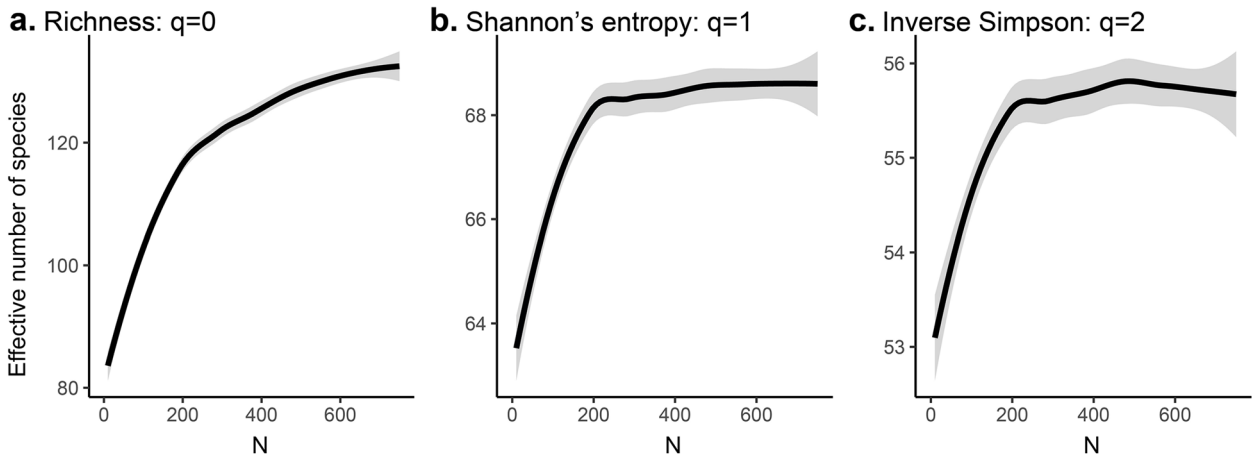

Fig. 2 Measuring α-diversity with Hill numbers. *a*-diversity indices were first unified by ecologist Mark Hill as the inverse of mean proportional abundances in a community [123, 124]. The value of *q* (or order of diversity) describes how this mean is calculated, affecting the sensitivity of diversity indices to rare species. In **a**–**c**, Hill numbers are shown for a metabarcoding data obtained from the fecal sample of an Ecuadorian finch (publicly available in NCBI with accession number SRR6486665 [125]). When q=0, the weighted harmonic mean of species' proportional abundances is measured, and richness is assessed (**a**). When q=1, the weighted geometric mean is measured, and Shannon's entropy is assessed (**b**). When q=2, the weighted arithmetic mean is measured and inverse Simpson's richness is assessed (**c**). All Hill numbers are expressed in units of effective numbers of species, or the number of species that would be expected in a community in which all species are equally abundant

For the estimation of SAD, parametric curves [70] can improve relative abundance estimates. From the histogram of observed relative abundances, different classical models can be fit (reviewed in [70]). The best fitting model can be selected and then compared across samples. The parameters estimated from curve-fitting, such as Fisher's *a*, can serve as a diversity metric that reflects the imbalance between few dominant species versus many rare species. For example, a community with a few highly abundant species and many rare ones will show a low α value compared to a similarly rich community with more even abundances. Fisher's *a* is more sensitive to species of medium abundances and thus more helpful in the context of incomplete sampling [116].

Due to the compositional nature of MME data, less abundant or less detectable components can be pushed below detection thresholds limits which is akin to the classic ecological notion of "veil lines" [21]. Post hoc statistical corrections such as species distribution models attempt to reconstruct information about missing molecular entities based on the observed composition of local or regional species pools [117].

α-diversity

Because of the long tail of rare molecular entities and the variable, uninformative N, computing a spectrum of diversity indices can provide a more complete picture of the diversity profile of a sample [118, 119]. In particular, Hill numbers go beyond single-point diversity estimators for fairer comparison between samples across different observation depths [120], leveraging the accumulation of individuals to produce standardized, continuous diversity estimators (Fig. 2). This continuum of estimates is obtained by varying the exponent q (or order) of the Hill numbers. As q increases, the relative importance of abundant species increases, providing information on whether common or rare species contribute most to *a*-diversity. Hill numbers have been shown to be more robust diversity estimators from molecular data, especially with q>1, as they are less sensitive to rare species and sparse datasets [121]. Importantly, because MME data processing and treatments disproportionately affect rare molecular entity, higher order Hill numbers can more robustly estimate *a*-diversity regardless of the researcher's technical decision-making [122]. An added advantage of Hill numbers is that they can produce confidence intervals along the accumulation of samples, quantifying uncertainty clearly compared to point estimates [24]. Leveraging Hill numbers of q>1 is particularly useful for datasets in which there is a high variation in S across samples (e.g., metabarcoding, metagenomics) and in which the community is not fully characterized [119]. It is important to note, however, that the increased robustness of higher order Hill numbers does result in the loss of information about rare species.

β-diversity

β-diversity metrics are commonly used across MME data to quantify the extent to which the molecular entities observed differ between samples. In general, β-diversity

is assessed by using similarity or dissimilarity metrics, and the choice of metric depends on the data's limitations as well as the research question of interest. For example, in the commonly used metric Euclidean distances (e.g., in principal components analyses), species absences are equally as informative as presences, making them unsuitable for characteristically sparse MME datasets [126]. In contrast, the semimetric Bray-Curtis dissimilarities give more weight to mutual presences and do not consider mutual absences among two samples to be informative.

If the emphasis is on the presence and absence of molecular entities rather than on their relative abundances, MME data can be converted into incidence data, and Sorensen dissimilarities can be calculated [126]. Importantly, incidence-based assessments are heavily affected by rare molecular entities, whose detection (or non-detection) may be artifactual in the case of MME data [127]. Filtering can also bias the measurement of β-diversity in molecular data. For example, in metabarcoding, the practice of removing rare species has recently been shown to artificially decrease α- and increase β-diversity, while reducing the discriminatory power of β-diversity [128].

Similar to α-diversity, β-diversity indices based on Hill numbers have been developed [129, 130], and allow adjusting the relative importance of rare versus abundant species when computing β-diversity through the q parameter. For example, when comparing the similarity between two communities, it is possible to give more weight to shared abundant species than to shared rare species. By varying the value of q, these β-diversity assessments can draw continuous diversity profiles. As q increases, the removal of rare species becomes less central to the calculation of β-diversity.

By permuting the available data to create a random or null expectation of the distributions of entities in a community, null modeling-based approaches can distinguish between changes in S and SAD for a wide range of dissimilarity metrics [131, 132]. Incidence-based β-diversity metrics can be decomposed into nestedness, which quantifies the extent to which samples with smaller numbers of species are subsets of more species-rich samples, and turnover, which quantifies the replacement, or difference of species between samples [133]. Partitioning β-diversity changes can shed light into the ecological processes driving molecular diversity (e.g., in metabarcoding [134]).

When hierarchical information on the similarity among molecular entities is available (e.g., in metabarcoding [135] and proteomics data [136]), phylogenetic β-diversity indices can be used to estimate how the relatedness among molecular entities affects observed community changes [137, 138]. The interpretation of variation in phylogenetic diversity indices can help tease apart evolutionary mechanisms at play [139], reviewed in [140]. For example, phylogenetic clustering can be the result of habitat filtering or of a constrained regional species pool [139]. Integrating phylogenetic β-diversity metrics can shed light into ecological processes (e.g., during secondary succession [141]). Null models can be further extended to account for phylogenetic relationships (i.e., Community Assembly Mechanisms by phylogenetic-bin-based null model analyses or iCamps [142]).

Considering similarities between MME and CE data can provide further avenues of innovation for analysis of β-diversity. For example, the analysis of functional diversity focuses on the diversity of characteristics or functional traits of diversity components [143]. Functional diversity can be estimated either through a species-centered approach, where traits are associated with a specific taxon, or estimated at the community scale through commonly used methods in microbiology [144], providing flexibility in the analysis and interpretation of MME data, which are often associated to a host (e.g., metabolomics). Similar to taxonomic and phylogenetic facets, functional diversity can be estimated through derivations of Hill numbers that can account for possible data treatments [129, 130]. For MME data derived from spectra, extant methods for deriving functional diversity estimates from remote sensing data can be applied [145].

Conclusion: integrating MME data in multi-omics research

The future of microbiome research will likely involve combining various MME techniques (i.e., multi-omics) to determine "who is doing what" [146]. Multi-omics research may also yield new insights that link molecular biology and ecology. For example, combining metatranscriptomics and metaproteomic measurements over time has revealed that on average, archaea produce more proteins per RNA molecule than bacteria [147]. Exploring the gut microbiome of Crohn's disease patients with both metabarcoding and metagenomics showed that metabarcoding data better predicted disease state, whereas metagenomics data were better at classifying treatment response [148]. In another study, combining metagenomic and metatranscriptomic techniques yielded novel insights into carbon cycling in soils [149].

Combining omics with other techniques can increase the specificity of the results and address complex ecological questions, providing new insights into host-microbiome interactions, revealing the trophic structure of a community, and shedding light on the metabolic pathways linking community members. For example, by combining stable isotope fingerprinting (SIF) with classic metaproteomics, direct protein-SIF, allows for the study of the individual physiology and metabolism of microbes

within a community [150]. In one case, transcriptomics, metabolomics, proteomics, and metabarcoding were combined to study host-microbiome interactions in prediabetic individuals, revealing distinct host-microbiome responses between insulin-resistant and insulin-sensitive subjects exposed to viral infections [151].

The greatest challenge to the future of multi-omics is arguably data interpretation [152], as the analyses which inform interpretation require the integration of MME datasets which may have multiple biases. To improve interpretation, it is necessary to consider how these biases arise throughout the data generation pipeline and address them. This includes finding a consensus in experimental design (especially one that allows for different molecules to be extracted simultaneously [153, 154]), collecting necessary metadata, considering challenges in joint sampling and in storage of different molecules with different decay rates, acknowledging different coverage of reference databases [155], and explicitly selecting tools for data integration and interpretation [156]. Further research at fine spatial and temporal scales may improve the discrimination of technical noise and intrinsic variation across MME techniques and inform the development of experimental designs that minimize this noise.

From an ecological standpoint, integrating MME data begins by studying whether different MME data behave similarly across an ecological gradient of interest. As MME data interpretation becomes more advanced and ecological questions become more sophisticated, the joint analysis of multiple MME data matrices will require more advanced statistical methods. Here, statistical advances related to the fourth corner problem, which refers to the difficulty of inferring trait-environment relationships directly from environmental, species abundance, and trait data, may become instrumental [157, 158]. As analytical frameworks increase in complexity to keep up with growing needs for data integration, understanding the limitations of MME data will continue to ensure that data interpretation also improves, both in the specificity and accuracy of conclusions.

Acknowledgements
The authors acknowledge the support of iDiv via the German Research Foundation (DFG FZT 118, 202548816), specifically through sDiv, the Synthesis Centre of iDiv. We would like to thank S. Tem for valuable discussions.

Authors' contributions
All authors were part of the "Imperfect Omics" workshop, held in 2020 at the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, and participated actively in manuscript development and review. S.D.J and A.H.B organized the workshop and framed the manuscript. All authors read and approved the final manuscript

Funding
We acknowledge support by the German Centre for Integrative Biodiversity Research Halle-Jena-Leipzig, funded by the German Research Foundation (FZT 118, 202548816)

Availability of data and materials
Not applicable

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Environmental Microbiology, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany. [2]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany. [3]Institute of Biology, Leipzig University, Leipzig, Germany. [4]Department of Soil Ecology, Helmholtz Centre for Environmental Research- UFZ, Halle, Germany. [5]Institute of Biodiversity, Friedrich Schiller University, Jena, Germany. [6]Institute of Biology, University of Graz, Graz, Austria. [7]Department of Botany, University of Wyoming, Wyoming, USA. [8]Department of Biology and Center for Biodiversity and Conservation Research, University of Mississippi, Oxford, Mississippi, USA. [9]Department of Biology, Indiana University, Indiana, USA. [10]Institute of Biology, Geobotany and Botanical Garden, Martin Luther University Halle Wittenberg, Halle, Germany. [11]Leibniz Institute of Plant Biochemistry, Bioinformatics and Scientific Data, Halle, Germany. [12]Leibniz Institute of Vegetable and Ornamental Crops (IGZ), Großbeeren, Germany. [13]Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, Netherlands.

Received: 3 April 2022 Accepted: 10 November 2022
Published online: 13 December 2022

References

1. Goldenfeld N, Woese C. Biology's next revolution. Nature. 2007;445(7126):369.
2. Group G. Genetics (Macmillan Science Library) (4 Volume set). New York: Macmillan Reference USA; 2002.
3. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science. 2008;321(5891):956–60.
4. Anderson NL, Anderson NG. Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis. 1998;19(11):1853–61.
5. Klassen A, Faccio AT, Canuto GAB, da Cruz PLR, Ribeiro HC, Tavares MFM, et al. Metabolomics: definitions and significance in systems biology. Adv Exp Med Biol. 2017;965:3–17.
6. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004;428(6978):37–43.
7. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 2004;304(5667):66–74.
8. Leininger S, Urich T, Schloter M, Schwark L, Qi J, Nicol GW, et al. Archaea predominate among ammonia-oxidizing prokaryotes in soils. Nature. 2006;442(7104):806–9.
9. Wilmes P, Bond PL. The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. Environ Microbiol. 2004;6(9):911–20.
10. Goodacre R. Metabolomics of a superorganism. The Journal of nutrition. 2007;137(1):259S-66S. https://doi.org/10.1093/jn/137.1.259S.
11. Locey KJ, Muscarella ME, Larsen ML, Bray SR, Jones SE, Lennon JT. Dormancy dampens the microbial distance-decay relationship. Philos Trans R Soc Lond Ser B Biol Sci. 2020;375(1798):20190243.

12. Nuccio EE, Starr E, Karaoz U, Brodie EL, Zhou J, Tringe SG, et al. Niche differentiation is spatially and temporally regulated in the rhizosphere. ISME J. 2020;14(4):999–1014.
13. Bedhomme S, Perez Pantoja D, Bravo IG. Plasmid and clonal interference during post horizontal gene transfer evolution. Mol Ecol. 2017;26(7):1832–47.
14. Quinn RA, Vermeij MJA, Hartmann AC, Galtier d'Auriac I, Benler S, Haas A, et al. Metabolomics of reef benthic interactions reveals a bioactive lipid involved in coral defence. Proc Biol Sci. 2016;283(1829):20160469.
15. Adav SS, Ravindran A, Chao LT, Tan L, Singh S, Sze SK. Proteomic analysis of pH and strains dependent protein secretion of Trichoderma reesei. J Proteome Res. 2011;10(10):4579–96.
16. Batta-Lona PG, Maas AE, O'Neill RJ, Wiebe PH, Bucklin A. Transcriptomic profiles of spring and summer populations of the Southern Ocean salp, Salpa thompsoni, in the Western Antarctic Peninsula region. Polar Biol. 2017;40(6):1261–76.
17. Leibold MA, Chase JM. Metacommunity Ecology, Volume 59 (Monographs in Population Biology, 59). Princeton: Princeton University Press; 2017.
18. He F, Legendre P. Species diversity patterns derived from species-area models. Ecology. 2002;83(5):1185–98.
19. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in low microbial biomass microbiome studies: issues and recommendations. Trends Microbiol. 2019;27(2):105–17.
20. Nichols RV, Vollmers C, Newsom LA, Wang Y, Heintzman PD, Leighton M, et al. Minimizing polymerase biases in metabarcoding. Mol Ecol Resour. 2018;18(5):927-39.
21. Preston FW. The commonness, and rarity, of species. Ecology. 1948;29(3):254.
22. Sanders HL. Marine benthic diversity: a comparative study. Am Nat. 1968;102(925):243–82.
23. Simberloff D. Use of rarefaction and related methods in ecology. In: Dickson KL, Cairns J, Livingston RJ, editors. Biological data in water pollution assessment: quantitative and statistical analyses. 100 Barr Harbor Drive, PO Box C700. West Conshohocken: ASTM International; 1978. p. 150-150–16.
24. Chao A, Gotelli NJ, Hsieh TC, Sander EL, Ma KH, Colwell RK, et al. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecol Monogr. 2014;84(1):45–67.
25. Nearing JT, Douglas GM, Hayes MG, MacDonald J, Desai DK, Allward N, et al. Microbiome differential abundance methods produce different results across 38 datasets. Nat Commun. 2022;13(1):342.
26. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017;5(1):27.
27. Chase JM, Knight TM. Scale-dependent effect sizes of ecological drivers on biodiversity: why standardised sampling is not enough. Ecol Lett. 2013;16(Suppl 1):17–26.
28. Levin SA. The problem of pattern and scale in ecology: the Robert H. MacArthur Award Lecture. Ecology. 1992;73(6):1943–67.
29. Chase JM. Spatial scale resolves the niche versus neutral theory debate. J Veg Sci. 2014;25(2):319–22.
30. Mark Welch JL, Rossetti BJ, Rieken CW, Dewhirst FE, Borisy GG. Biogeography of a human oral microbiome at the micron scale. Proc Natl Acad Sci U S A. 2016;113(6):E791–800.
31. Xu L, Pierroz G, Wipf HM-L, Gao C, Taylor JW, Lemaux PG, et al. Holoomics for deciphering plant-microbiome interactions. Microbiome. 2021;9(1):69.
32. Hurlbert AH, Jetz W. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. Proc Natl Acad Sci U S A. 2007;104(33):13384–9.
33. Kashtan N, Bushong B, Leveau JHJ. It's the economy, stupid: applying (micro)economic principles to microbiome science. mSystems. 2022;7(1):e0103321.
34. Martiny JBH, Eisen JA, Penn K, Allison SD, Horner-Devine MC. Drivers of bacterial beta-diversity depend on spatial scale. Proc Natl Acad Sci U S A. 2011;108(19):7850–4.
35. Wiesel L, Daniell TJ, King D, Neilson R. Determination of the optimal soil sample size to accurately characterise nematode communities in soil. Soil Biol Biochem. 2015;80:89–91.
36. Osborne CA, Zwart AB, Broadhurst LM, Young AG, Richardson AE. The influence of sampling strategies and spatial variation on the detected soil bacterial communities under three different land-use types. FEMS Microbiol Ecol. 2011;78(1):70–9.
37. Penton CR, Gupta VVSR, Yu J, Tiedje JM. Size matters: assessing optimum soil sample size for fungal and bacterial community structure analyses using high throughput sequencing of rRNA gene amplicons. Front Microbiol. 2016;7:824.
38. Jurburg SD, Keil P, Singh BK, Chase JM. All together now: limitations and recommendations for the simultaneous analysis of all eukaryotic soil sequences. Mol Ecol Resour. 2021;21(6):1759–71.
39. Dickie IA, Boyer S, Buckley HL, Duncan RP, Gardner PP, Hogg ID, et al. Towards robust and repeatable sampling methods in eDNA-based studies. Mol Ecol Resour. 2018;18(5):940–52.
40. Nekola JC, White PS. The distance decay of similarity in biogeography and ecology. J Biogeogr. 1999;26(4):867–78.
41. Clark DR, Underwood GJC, McGenity TJ, Dumbrell AJ. What drives study-dependent differences in distance–decay relationships of microbial communities? Glob Ecol Biogeogr. 2021;30(4):811–25.
42. Shade A, Dunn RR, Blowes SA, Keil P, Bohannan BJM, Herrmann M, et al. Macroecology to unite all life, large and small. Trends Ecol Evol. 2018;33(10):731–44.
43. De Gruyter J, Weedon JT, Bazot S, Dauwe S, Fernandez-Garberí P-R, Geisen S, et al. Patterns of local, intercontinental and interseasonal variation of soil bacterial and eukaryotic microbial communities. FEMS Microbiol Ecol. 2020;96(3):fiaa018.
44. Kaspari M, Stevenson BS, Shik J, Kerekes JF. Scaling community structure: how bacteria, fungi, and ant taxocenes differentiate along a tropical forest floor. Ecology. 2010;91(8):2221–6.
45. Zinger L, Taberlet P, Schimann H, Bonin A, Boyer F, De Barba M, et al. Body size determines soil community assembly in a tropical forest. Mol Ecol. 2019;28(3):528–43.
46. Louca S, Jacques SMS, Pires APF, Leal JS, Srivastava DS, Parfrey LW, et al. High taxonomic variability despite stable functional structure across microbial communities. Nat Ecol Evol. 2016;1(1):15.
47. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207–14.
48. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11(12):2639–43.
49. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9(1):5114.
50. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for bacteria and Archaea. Nat Biotechnol. 2020;38(9):1079–86.
51. Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, Paredes R, Noguera-Julian M, Calle ML. Balances: a new perspective for microbiome analysis. mSystems. 2018;3(4):e00053-18.
52. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol. 2017;8:2224.
53. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010;11:94.
54. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
55. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 2013;10(12):1200–2.
56. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
57. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12(6):R60.
58. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29.

59. Tang J, Fu J, Wang Y, Li B, Li Y, Yang Q, et al. ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. Brief Bioinform. 2020;21(2):621–36.
60. Langley SR, Mayr M. Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics. J Proteome. 2015;129:83–92.
61. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013;14(6):671–83.
62. Marr S, Hageman JA, Wehrens R, van Dam NM, Bruelheide H, Neumann S. LC-MS based plant metabolic profiles of thirteen grassland species grown in diverse neighbourhoods. Sci Data. 2021;8(1):52.
63. Delgado-Baquerizo M, Trivedi P, Trivedi C, Eldridge DJ, Reich PB, Jeffries TC, et al. Microbial richness and composition independently drive soil multifunctionality. Funct Ecol. 2017;31(12):2330–43.
64. Maciá-Vicente JG, Shi Y-N, Cheikh-Ali Z, Grün P, Glynou K, Kia SH, et al. Metabolomics-based chemotaxonomy of root endophytic fungi for natural products discovery. Environ Microbiol. 2018;20(3):1253–70.
65. Wasimuddin SK, Ronchi F, Leib SL, Erb M, Ramette A. Evaluation of primer pairs for microbiome profiling from soils to humans within the One Health framework. Mol Ecol Resour. 2020;20(6):1558–71.
66. Armbruster DA, Pry T. Limit of blank, limit of detection and limit of quantitation. Clin Biochem Rev. 2008;29(Suppl 1):S49–52.
67. Lazar C, Gatto L, Ferro M, Bruley C, Burger T. Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. J Proteome Res. 2016;15(4):1116–25.
68. McLaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic sequencing experiments. eLife. 2019;8:e46923.
69. Shen M, Chang Y-T, Wu C-T, Parker SJ, Saylor G, Wang Y, et al. Comparative assessment and novel strategy on methods for imputing proteomics data. Sci Rep. 2022;12(1):1067.
70. McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Benecha HK, et al. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecol Lett. 2007;10(10):995–1015.
71. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A. 2006;103(32):12115–20.
72. Chase JM, McGill BJ, McGlinn DJ, May F, Blowes SA, Xiao X, et al. Embracing scale-dependence to achieve a deeper understanding of biodiversity and its change across communities. Ecol Lett. 2018;21(11):1737–51.
73. Kumar MS, Slud EV, Okrah K, Hicks SC, Hannenhalli S, Corrada BH. Analysis and correction of compositional bias in sparse sequencing count data. BMC Genomics. 2018;19(1):799.
74. Aitchison J. The statistical analysis of compositional data. J R Stat Soc Ser B Methodol. 1982;44(2):139–60.
75. Quinn TP, Richardson MF, Lovell D, Crowley TM. propr: an R-package for identifying proportionally abundant features using compositional data analysis. Sci Rep. 2017;7(1):16252.
76. Quinn TP, Erb I, Gloor G, Notredame C, Richardson MF, Crowley TM. A field guide for the compositional analysis of any-omics data. Gigascience. 2019;8(9):giz107.
77. Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, et al. Establishing microbial composition measurement standards with reference frames. Nat Commun. 2019;10(1):2719.
78. Zhao N, Zhan X, Guthrie KA, Mitchell CM, Larson J. Generalized Hotelling's test for paired compositional data with application to human microbiome studies. Genet Epidemiol. 2018;42(5):459–69.
79. Peng X, Li G, Liu Z. Zero-inflated beta regression for differential abundance analysis with metagenomics data. J Comput Biol. 2016;23(2):102–10.
80. Wang C, Hu J, Blaser MJ, Li H. Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. Bioinformatics. 2020;36(2):347–55.
81. Sisk-Hackworth L, Kelley ST. An application of compositional data analysis to multiomic time-series data. NAR Genom Bioinform. 2020;2(4):lqaa079.
82. Quinn TP, Erb I, Richardson MF, Crowley TM. Understanding sequencing data as compositions: an outlook and review. Bioinformatics. 2018;34(16):2870–8.
83. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis. 2015;26:27663.
84. Kaul A, Mandal S, Davidov O, Peddada SD. Analysis of microbiome data in the presence of excess zeros. Front Microbiol. 2017;8:2114.
85. Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PLoS One. 2013;8(7):e67019.
86. Petersen A-K, Krumsiek J, Wägele B, Theis FJ, Wichmann H-E, Gieger C, et al. On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. BMC Bioinformatics. 2012;13:120.
87. Clark JS, Nemergut D, Seyednasrollah B, Turner PJ, Zhang S. Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data. Ecol Monogr. 2017;87(1):34–56.
88. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc. 2011;6(7):1060–83.
89. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581–3.
90. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. BioRxiv. 2016; 081257.
91. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems. 2017;2(2):e00191-16.
92. Chao A, Kubota Y, Zelený D, Chiu C, Li C, Kusumoto B, et al. Quantifying sample completeness and comparing diversities among assemblages. Ecol Res. 2020;35(2):292–314.
93. Brown SP, Veach AM, Rigdon-Huss AR, Grond K, Lickteig SK, Lothamer K, et al. Scraping the bottom of the barrel: are rare high throughput sequences artifacts? Fungal Ecol. 2015;13:221–5.
94. Collins CD, Holt RD, Foster BL. Patch size effects on plant species decline in an experimentally fragmented landscape. Ecology. 2009;90(9):2577–88.
95. Kellner KF, Swihart RK. Accounting for imperfect detection in ecology: a quantitative review. PLoS One. 2014;9(10):e111436.
96. Dethier MN, Graham ES, Cohen S, Tear LM. Visual versus random-point percent cover estimations: "objective" is not always better. Mar Ecol Prog Ser. 1993;96:93–100.
97. Kercher SM, Frieswyk CB, Zedler JB. Effects of sampling teams and estimation methods on the assessment of plant cover. J Veg Sci. 2003;14(6):899–906.
98. Keim JL, DeWitt PD, Fitzpatrick JJ, Jenni NS. Estimating plant abundance using inflated beta distributions: applied learnings from a lichen-caribou ecosystem. Ecol Evol. 2017;7(2):486–93.
99. Martin TG, Wintle BA, Rhodes JR, Kuhnert PM, Field SA, Low-Choy SJ, et al. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol Lett. 2005;8(11):1235–46.
100. McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KR. Methods for normalizing microbiome data: an ecological perspective. Methods Ecol Evol. 2018;10(3):389-400.
101. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014;10(4):e1003531.
102. Colwell RK, Chao A, Gotelli NJ, Lin SY, Mao CX, Chazdon RL, et al. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. J Plant Ecol. 2012;5(1):3–21.
103. Lele SR, Moreno M, Bayne E. Dealing with detection error in site occupancy surveys: what can we do with a single survey? J Plant Ecol. 2012;5(1):22–31.
104. Bunge J, Willis A, Walsh F. Estimating the number of species in microbial diversity studies. Annu Rev Stat Appl. 2014;1(1):427–45.
105. Bálint M, Bahram M, Eren AM, Faust K, Fuhrman JA, Lindahl B, et al. Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. FEMS Microbiol Rev. 2016;40(5):686–700.
106. Chiu C-H, Chao A. Estimating and comparing microbial diversity in the presence of sequencing errors. PeerJ. 2016;4:e1634.

107. Willis A, Bunge J. Estimating diversity via frequency ratios. Biometrics. 2015;71(4):1042–9.
108. Magurran AE, McGill BJ. Biological diversity: frontiers in measurement and assessment. 1st ed. Oxford: Oxford University Press; 2011.
109. Ficetola GF, Pansu J, Bonin A, Coissac E, Giguet-Covex C, De Barba M, et al. Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. Mol Ecol Resour. 2015;15(3):543–56.
110. Doi H, Fukaya K, Oka S-I, Sato K, Kondoh M, Miya M. Evaluation of detection probabilities at the water-filtering and initial PCR steps in environmental DNA metabarcoding using a multispecies site occupancy model. Sci Rep. 2019;9(1):3581.
111. McClenaghan B, Compson ZG, Hajibabaei M. Validating metabarcoding-based biodiversity assessments with multi-species occupancy models: a case study using coastal marine eDNA. PLoS One. 2020;15(3):e0224119.
112. Brost BM, Mosher BA, Davenport KA. A model-based solution for observational errors in laboratory studies. Mol Ecol Resour. 2018;18(3):580–9.
113. Willoughby JR, Wijayawardena BK, Sundaram M, Swihart RK, DeWoody JA. The importance of including imperfect detection models in eDNA experimental design. Mol Ecol Resour. 2016;16(4):837–44.
114. Gotelli NJ. Null model analysis of species co-occurrence patterns. Ecology. 2000;81(9):2606–21.
115. Tikhonov G, Opedal ØH, Abrego N, Lehikoinen A, de Jonge MMJ, Oksanen J, et al. Joint species distribution modelling with the r-package Hmsc. Methods Ecol Evol. 2020;11(3):442–7.
116. Kempton RA. The structure of species abundance and measurement of diversity. Biometrics. 1979;35(1):307.
117. Lewis RJ, Szava-Kovats R, Pärtel M. Estimating dark diversity and species pools: an empirical assessment of two methods. Methods Ecol Evol. 2016;7(1):104–13.
118. Alberdi A, Gilbert MTP. A guide to the application of Hill numbers to DNA-based diversity analyses. Mol Ecol Resour. 2019;19(4):804–17.
119. Roswell M, Dushoff J, Winfree R. A conceptual guide to measuring species diversity. Oikos. 2021;130(3):321–38.
120. Hsieh TC, Ma KH, Chao A. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). Methods Ecol Evol. 2016;7(12):1451-56.
121. Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, Weitz JS. Robust estimation of microbial diversity in theory and in practice. ISME J. 2013;7(6):1092–101.
122. Mächler E, Walser J-C, Altermatt F. Decision-making and best practices for taxonomy-free environmental DNA metabarcoding in biomonitoring using Hill numbers. Mol Ecol. 2021;30(13):3326–39.
123. Tuomisto H. A consistent terminology for quantifying species diversity? Yes, it does exist. Oecologia. 2010;164(4):853–60.
124. Hill MO. Diversity and evenness: a unifying notation and its consequences. Ecology. 1973;54(2):427–32.
125. Michel AJ, Ward LM, Goffredi SK, Dawson KS, Baldassarre DT, Brenner A, et al. The gut of the finch: uniqueness of the gut microbiome of the Galápagos vampire finch. Microbiome. 2018;6(1):167.
126. Buttigieg PL, Ramette A. A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. FEMS Microbiol Ecol. 2014;90(3):543–50.
127. Deagle BE, Thomas AC, McInnes JC, Clarke LJ, Vesterinen EJ, Clare EL, et al. Counting with DNA in metabarcoding studies: how should we convert sequence reads to dietary data? Mol Ecol. 2019;28(2):391–406.
128. Schloss PD. Removal of rare amplicon sequence variants from 16S rRNA gene sequence surveys biases the interpretation of community structure data. BioRxiv. 2020.
129. Chiu C-H, Jost L, Chao A. Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers. Ecol Monogr. 2014;84(1):21–44.
130. Chao A, Chiu C-H, Jost L. Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers. Annu Rev Ecol Evol Syst. 2014;45(1):297–324.
131. Chase JM, Kraft NJB, Smith KG, Vellend M, Inouye BD. Using null models to disentangle variation in community dissimilarity from variation in α-diversity. Ecosphere. 2011;2(2):art24.
132. Keil P. *Z*-scores unite pairwise indices of ecological similarity and association for binary data. Ecosphere. 2019;10(11):e02933.
133. Baselga A. Partitioning the turnover and nestedness components of beta diversity. Glob Ecol Biogeogr. 2010;19(1):134–43.
134. Dassen S, Cortois R, Martens H, de Hollander M, Kowalchuk GA, van der Putten WH, et al. Differential responses of soil bacteria, fungi, archaea and protists to plant species richness and plant functional group identity. Mol Ecol. 2017;26(15):4085–98.
135. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005;71(12):8228–35.
136. Easterly CW, Sajulga R, Mehta S, Johnson J, Kumar P, Hubler S, et al. metaQuantome: an integrated, quantitative metaproteomics approach reveals connections between taxonomy and protein function in complex microbiomes. Mol Cell Proteomics. 2019;18(8 suppl 1):S82–91.
137. Mouquet N, Devictor V, Meynard CN, Munoz F, Bersier L-F, Chave J, et al. Ecophylogenetics: advances and perspectives. Biol Rev Camb Philos Soc. 2012;87(4):769–85.
138. Tucker CM, Cadotte MW, Carvalho SB, Davies TJ, Ferrier S, Fritz SA, et al. A guide to phylogenetic metrics for conservation, community ecology and macroecology. Biol Rev Camb Philos Soc. 2017;92(2):698–715.
139. Gerhold P, Cahill JF, Winter M, Bartish IV, Prinzing A. Phylogenetic patterns are not proxies of community assembly mechanisms (they are far better). Funct Ecol. 2015;29(5):600–14.
140. Zhou J, Ning D. Stochastic community assembly: does it matter in microbial ecology? Microbiol Mol Biol Rev. 2017;81(4):e00002-17.
141. Jurburg SD, Nunes I, Stegen JC, Le Roux X, Priemé A, Sørensen SJ, et al. Autogenic succession and deterministic recovery following disturbance in soil bacterial communities. Sci Rep. 2017;7:45691.
142. Ning D, Yuan M, Wu L, Zhang Y, Guo X, Zhou X, et al. A quantitative framework reveals ecological drivers of grassland microbial community assembly in response to warming. Nat Commun. 2020;11(1):4717.
143. Mammola S, Carmona CP, Guillerme T, Cardoso P. Concepts and applications in functional diversity. Funct Ecol. 2021;35(9):1869-85.
144. Escalas A, Hale L, Voordeckers JW, Yang Y, Firestone MK, Alvarez-Cohen L, et al. Microbial functional diversity: from concepts to applications. Ecol Evol. 2019;9(20):12000–16.
145. Laliberté E, Schweiger AK, Legendre P. Partitioning plant spectral diversity into alpha and beta components. Ecol Lett. 2020;23(2):370–80.
146. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018;16(7):410–22.
147. Delogu F, Kunath BJ, Evans PN, Arntzen MØ, Hvidsten TR, Pope PB. Integration of absolute multi-omics reveals dynamic protein-to-RNA ratios and metabolic interplay within mixed-domain microbiomes. Nat Commun. 2020;11(1):4708.
148. Douglas GM, Hansen R, Jones CMA, Dunn KA, Comeau AM, Bielawski JP, et al. Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. Microbiome. 2018;6(1):13.
149. Beulig F, Urich T, Nowak M, Trumbore SE, Gleixner G, Gilfillan GD, et al. Altered carbon turnover processes and microbiomes in soils under long-term extremely high CO2 exposure. Nat Microbiol. 2016;1:15025.
150. Kleiner M, Dong X, Hinzke T, Wippler J, Thorson E, Mayer B, et al. Metaproteomics method to determine carbon sources and assimilation pathways of species in microbial communities. Proc Natl Acad Sci U S A. 2018;115(24):E5576–84.
151. Zhou W, Sailani MR, Contrepois K, Zhou Y, Ahadi S, Leopold SR, et al. Longitudinal multi-omics of host-microbe dynamics in prediabetes. Nature. 2019;569(7758):663–71.
152. Misra BB, Langefeld CD, Olivier M, Cox LA. Integrated omics: tools, advances, and future approaches. J Mol Endocrinol. 2019;62(1):R21-45.
153. Sapcariu SC, Kanashova T, Weindl D, Ghelfi J, Dittmar G, Hiller K. Simultaneous extraction of proteins and metabolites from cells in culture. MethodsX. 2014;1:74–80.
154. Roume H, Heintz-Buschart A, Muller EEL, Wilmes P. Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. Methods Enzymol. 2013;531:219–36.
155. Sajulga R, Easterly C, Riffle M, Mesuere B, Muth T, Mehta S, et al. Survey of metaproteomics software tools for functional microbiome analysis. PLoS One. 2020;15(11):e0241503.

156. Pinu FR, Goldansaz SA, Jaine J. Translational metabolomics: current challenges and future opportunities. Metabolites. 2019;9(6):108.
157. Legendre P, Galzin R, Harmelin-Vivien ML. Relating behabior to habitat: solutions to the fourth-corner problem. Ecology. 1997;78(2):547–62.
158. Dray S, Choler P, Dolédec S, Peres-Neto PR, Thuiller W, Pavoine S, et al. Combining the fourth-corner and the RLQ methods for assessing trait responses to environmental variation. Ecology. 2014;95(1):14–21.

Publisher's Note