Received: 4 January 2021 | Revised: 31 March 2021 | Accepted: 13 April 2021

DOI: 10.1111/1755-0998.13401

MOLECULAR ECOLOGY RESOURCES WILEY

NEWS AND VIEWS

OPINION

All together now: Limitations and recommendations for the simultaneous analysis of all eukaryotic soil sequences

Stephanie D. Jurburg[1,2] | Petr Keil[1,3,4] | Brajesh K. Singh[5] | Jonathan M. Chase[1,3]

[1]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

[2]Institute of Biology, Leipzig University, Leipzig, Germany

[3]Department of Computer Science, Martin Luther University, Halle-Wittenberg, Halle, Germany

[4]Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Praha–Suchdol, Czech Republic

[5]Hawkesbury Institute for the Environment, and Global Centre for Land-Based Innovation, Western Sydney University, Penrith, NSW, Australia

Correspondence
Stephanie D. Jurburg, German Centre for Integrative Biodiversity Research (iDiv), Leipzig, Germany
Email: s.d.jurburg@gmail.com

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: FZT 118

Abstract

The soil environment contains a large, but historically underexplored, reservoir of biodiversity. Sequencing prokaryotic marker genes has become commonplace for the discovery and characterization of soil bacteria and archaea. Increasingly, this approach is also applied to eukaryotic marker genes to characterize the diversity and distribution of soil eukaryotes. However, understanding the properties and limitations of eukaryotic marker sequences is essential for correctly analysing, interpreting, and synthesizing the resulting data. Here, we illustrate several biases from sequencing data that affect measurements of biodiversity that arise from variation in morphology, taxonomy and phylogeny between organisms, as well as from sampling designs. We recommend analytical approaches to overcome these limitations, and outline how the benchmarking and standardization of sequencing protocols may improve the comparability of the data.

KEYWORDS
community ecology, microbiome, sequencing, soil

1 | INTRODUCTION

Next generation sequencing technologies have revolutionized microbial ecology, revealing the extensive diversity of bacteria and archaea in our planet (Bates et al., 2011; Thompson et al., 2017), and providing insights into their ecology (Fierer & Jackson, 2006; Martiny et al., 2006). The popularity of amplicon sequencing, where a section of a universal marker gene is amplified and sequenced, has soared over the past decade. In soil, amplicon sequencing of the 16S rRNA gene has been especially useful, as soil prokaryotes remain largely uncultured but are extremely diverse, and perform key ecosystem functions (Steen et al., 2019).

Soil eukaryotes, including protists, worms, arthropods, fungi, plant roots, and others, have received comparatively less attention. This is due to technological difficulties associated with sampling, including the complexity and heterogeneity of the soil matrix (Orgiazzi et al., 2016). Consequently, data on the diversity and distributions of soil meso- and macrofauna are limited (Cameron et al., 2019), largely because the identification of these organisms is body size-specific, labour-intensive, and requires a deep knowledge of organisms' morphologies or specific biochemistry (Orgiazzi et al., 2016). Nevertheless, soil eukaryotes are essential to soil functions, as both consumers and ecosystem engineers (Thakur et al., 2019).

Amplicon sequencing has become an increasingly attractive alternative for the identification of soil eukaryotes (Pawlowski et al., 2020). Universal marker genes including the ITS region, as well as the 18S rRNA, mitochondrial 16S rRNA, and COI genes, have been used to assess the global diversity of fungi (Tedersoo et al., 2014; Větrovský et al., 2019), protists (Oliverio et al., 2020), nematodes, microarthropods (Wu et al., 2011), and rotifers (Robeson et al., 2011) without the

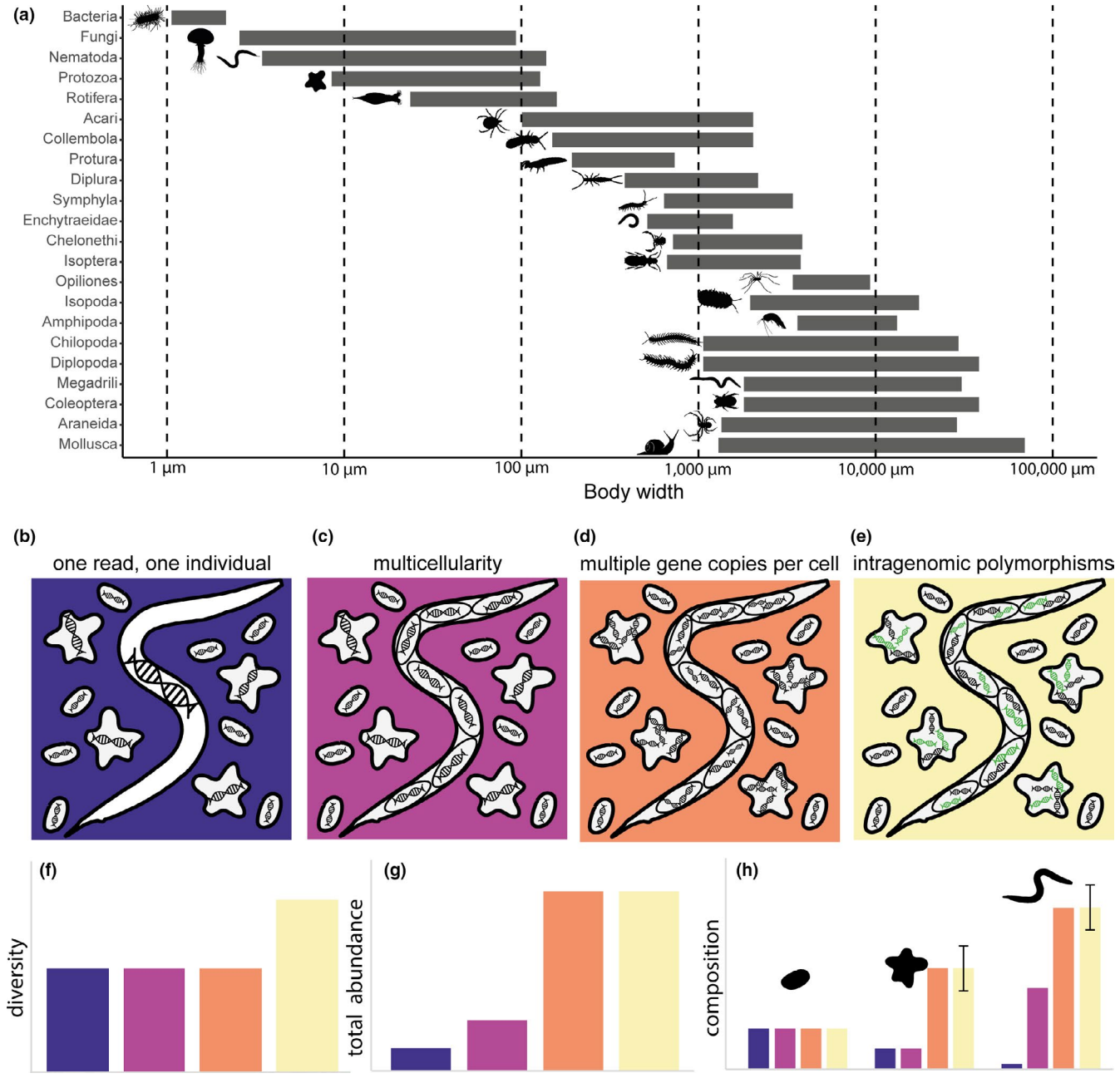

FIGURE 1 Issues arising from variation in morphology. (a) Soil biota comprise a broad range of sizes (adapted from Swift et al., 1979). (b) The ideal scenario for amplicon sequencing-based studies is that a sequenced read is equivalent to an organism, or proportional to its abundance. (c) The multicellularity of eukaryotes results in an overestimation of the abundance according to body size, while (d) the abundance of organisms with multiple copies of the marker gene per cell will also be overestimated. Finally, (e) organisms belonging to a single species, but that contain multiple, different copies of the marker gene (intragenomic polymorphisms) may be estimated as several species. These biases affect estimates of (f) diversity, (g) the total abundance of organisms, and (h) the abundances of specific species. In the case of intragenomic polymorphisms, compositional data may be biased by the incorrect classification of different sequences from a single organism as multiple species, indicated by error bars

need for prior isolation of target organisms (Pawlowski et al., 2020). These markers allow researchers to study several groups simultaneously, filling gaps in soil biodiversity data and serving as the basis for synthesis efforts (Orgiazzi et al., 2015). Such efforts are underway both at regional and global scales (e.g., Bastida et al., 2020; Delgado-Baquerizo et al., 2020; Ramirez et al., 2014) and are becoming increasingly important to ecological research (Compson et al., 2020).

As amplicon sequencing of eukaryotic markers becomes standard practice (Pawlowski et al., 2020), it is important to understand how the variation among taxa, sampling, and sequencing affect the analysis and interpretation of amplicon sequences derived from eukaryotic markers (Bent & Forney, 2008; Nekola & White, 1999; Ruppert et al., 2019; Taberlet et al., 2018). Here, we explore how the wide range of morphological and phylogenetic variation, as well as

sampling practices, compromise the comparability of diversity and community composition among taxa, studies, and sampling designs from amplicon sequencing data. Our aim is to identify the challenges and limitations associated with this approach, as well as to provide recommendations on how to produce and analyse amplicon sequencing data so that the data are reusable and results interpretable.

2 | PROBLEMS

Morphological, taxonomic, phylogenetic, and sampling variation may all bias the quality of amplicon sequencing data of eukaryotes. In turn, this alters estimates of their α-diversity (i.e., richness, evenness), β-diversity (turnover), abundance, and composition.

2.1 | Morphological variation

Soil eukaryotes range in body size from unicellular protists to multicellular organisms (i.e., earthworms and snails; Figure 1a), all of which may contribute DNA to a soil sample. The study of ecological communities with amplicon sequencing relies on the assumption that the genes belonging to each individual in the community are amplified proportionally to their abundance in the community (ideally, one read: one organism; Figure 1b). All life deviates from this assumption, but with their variable morphologies, soil eukaryotes deviate from one read: one organism in several ways, and to a greater extent than prokaryotes, which we overview below.

First, multicellularity disrupts estimates of relative abundances, as these become confounded with the organisms' sizes (Elbrecht & Leese, 2015; Figure 1c). Second, the variable number of copies of a marker gene is exacerbated in eukaryotes. While bacterial cells may contain up to 15 copies of the 16S RNA gene, protists may contain between 1 and 400,000 copies of the 18S rRNA gene (Kirchman, 2018; Figure 1d). In bacteria, this variable gene copy number can be corrected using bacterial sequence databases (although is discouraged for soil prokaryotes, which are poorly represented in sequence databases; Louca et al., 2018); however, eukaryotic sequence databases are considerably sparser (Geisen et al., 2019). These two phenomena obfuscate the relationship between the number of gene copies detected from a sample and the abundance of organisms (i.e., number of individuals) in the community, leading to potential overestimation of the abundance of larger individuals or those with the most copies of the marker gene (Geisen et al., 2019; Figure 1g–h). Studies which focus on groups with known cell numbers by isolating the organisms prior to sequencing have approached this limitation by modelling relative copy numbers per individuals (e.g., in nematodes; Darby et al., 2013) but this does not work for the majority of soil fauna that are highly variable in body size. Third, a single eukaryotic cell may have multiple, different copies of a gene (intragenomic polymorphisms, Figure 1e), as has been shown for protists, nematodes, and fungi (Bik et al., 2013; Thornhill & Santos, 2007; Wu et al., 2016). While multicellularity and multiple gene copies per cell lead to an overestimation of abundance, intragenomic polymorphisms can result in inflated estimates of α-diversity including richness, or the number of taxa (Figure 1f). These polymorphisms can emerge quickly (i.e., over 400 generations in a nematode population; Bik et al., 2013). Furthermore, the number of marker genes per cell may vary within an individual. For example, the number of mitochondria in a cell depend on the cell's function (Veltri et al., 1990), and this may also result in skewed estimates of individual abundances.

2.2 | Taxonomic and phylogenetic variation

To work accurately, the marker gene or region of choice must be sufficiently conserved on either flank of the DNA segment so that primers can capture all versions of the segment; but it must also be adequately variable in the centre of the segment to classify species according to variations in the DNA sequences. Such an ideal universal marker does not exist, as life exhibits a wide range of morphological and phylogenetic variation, and increased universality of a marker generally comes at the cost of taxonomic resolution. While the 16S rRNA gene is widely used to classify prokaryotes into taxonomic units, no such consensus exists among eukaryotes, and extant markers must consider several hurdles that arise from this variation.

Primer mismatches, in which a primer does not match the DNA template and fails to amplify it, occur selectively (Elbrecht & Leese, 2012; Tedersoo et al., 2016), resulting in an underestimation of α-diversity (Figure 2b,e). Primer mismatches result in the systematic exclusion of certain clades (Nichols et al., 2018), making diversity estimates primer-specific. For example, soil invertebrate communities exhibit different compositions depending on whether the 28S rRNA gene or the COI gene is sequenced (Dopheide et al., 2019), and diversity estimates may depend on the choice of marker gene (e.g., 18S rRNA or COI gene; Tang et al., 2012), or on the target region selected within the marker gene (e.g., in the 18S rRNA gene; Leasi et al., 2018).

The ability of a marker gene to detect taxonomic α-diversity is further obscured by the relationship between trait- and gene-based taxonomy. On the one hand, varying rates of evolution between different clades result in different taxonomic resolutions. For example, morphological differentiation in recently radiated lineages may be more apparent than genetic differences (Eberle et al., 2020). These may result in the classification of all members of a clade as a single species (Tang et al., 2012; Tedersoo et al., 2016; Figure 2c,e), the underestimation of α-diversity (Leasi et al., 2018), and the potential underestimation of β-diversity. On the other hand, marker genes may be better able to identify cryptic species (Fonseca, 2018).

The operational taxonomic units (OTUs) defined by sequences often do not match established taxonomic frameworks for eukaryotes (Figure 2d), and provide a different paradigm for quantifying diversity and composition than morphology-based assessments (Shade et al., 2018). A cutoff of 97% or 100% similarity in the 16S rRNA gene is generally used for prokaryotes (but see Mysara et al., 2017); however,

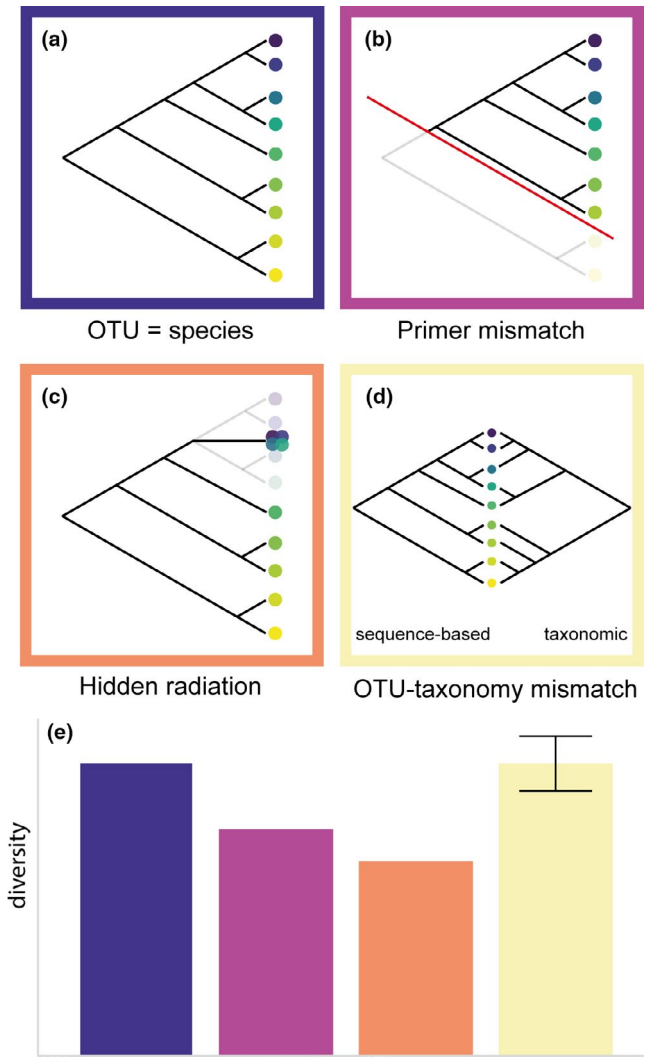

FIGURE 2 Issues arising from variation in phylogeny and taxonomy. (a) The ideal scenario for amplicon sequencing-based studies is that an operational taxonomic unit (OTU) is equivalent to a taxonomic species, and that all species are detected with sequencing; however (b) primer mismatches result in the systematic exclusion of certain branches of the phylogenetic tree which are not captured by the selected primer set, resulting in the omission of present taxa. (c) Differing evolutionary rates among clades may result in the clustering of several species into one OTU, and (d) the taxonomic classification of species may greatly differ from the sequence-based classification. (e) All three phenomena may bias estimates of biodiversity; however, the difference between taxonomy and sequence-based classifications may compound the misestimation of diversity, indicated by grey lines

no such consensus exists for eukaryotes. The taxonomic level at which the community is analysed greatly affects estimates of α-diversity (i.e., in nematode communities; Dell'Anno et al., 2015; Figure 2e). Due to the variable rates of evolution among eukaryotic taxa, there is probably no universally applicable species cutoff (Mysara et al., 2017). Defining taxonomic units at the level of single nucleotide variations—the strictest possible definition for the sequenced amplicons—is a viable, and increasingly popular alternative (Callahan et al., 2017; Edgar, 2018). However, this level of resolution may confound biological variation with sequencing-related artefacts, or capture population-level variation, affecting the ecological interpretations. Whether a single marker region can provide sufficient resolution to accurately characterize community α-diversity has been questioned (Rodriguez-R et al., 2018), and the degree to which universal markers capture diversity is marker-specific (Ficetola et al., 2020).

2.3 | Sampling variation

One of the greatest incentives for using amplicon sequencing is the potential to facilitate synthesis and the comparison among distinct groups of eukaryotic taxa. Spatial structuring has long been recognized in ecological sampling designs in microbial ecology (e.g., spatially explicit designs; Yergeau et al., 2007), and the study of the global distribution of microbes (Fierer & Jackson, 2006; Green et al., 2004; Martiny et al., 2006). The spatial scale of sampling is critical for comparability across studies (Dickie et al., 2018), but has received less attention. The spatial scale of sampling (Dungan et al., 2002) is characterized by volume or area of samples taken (their grain), the spatial extent of a study, and the distance between samples (Figure 3). All three aspects of spatial scale are seldom documented in studies of soil biota, and the homogenization of multiple, randomly selected samples within a plot is common.

Two fundamental spatial patterns in ecological communities are (i) the distance decay of compositional similarity (DDS, Nekola & White, 1999; Figure 3), and (ii) the taxa-area (or taxa-volume) relationships (TAR, Green et al., 2004; Woodcock et al., 2006; Figure 3). These illustrate why sampling choices can strongly affect the resulting estimates of diversity and community composition. For example, due to DDS, samples that are taken more closely together will typically have more similar compositions, and thus have lower total diversity, than samples that are taken further apart. It is therefore essential to take the grain, distance, and extent of the samples into consideration when comparing across samples which were collected in disparate ways. However, this information is seldom considered (Dickie et al., 2018). In addition, the abundance of soil biota varies over time, but temporal patterns of soil eukaryote diversity are poorly understood (Bálint et al., 2018; Briones, 2018).

The laboratory methodologies used prior to sequencing may also bias soil diversity assessments. Estimates of diversity are influenced by the laboratory protocol used for DNA extraction (Santos et al., 2017). Current protocols for eukaryotic sequencing are nearly identical to those for prokaryotic sequencing (e.g., the Earth Microbiome Project's protocols; Thompson et al., 2017), despite the much wider range of body sizes in soil eukaryotes (Figure 1). The suggested amount of the soil sample (less than 1 g in most commercial DNA extraction kits) is

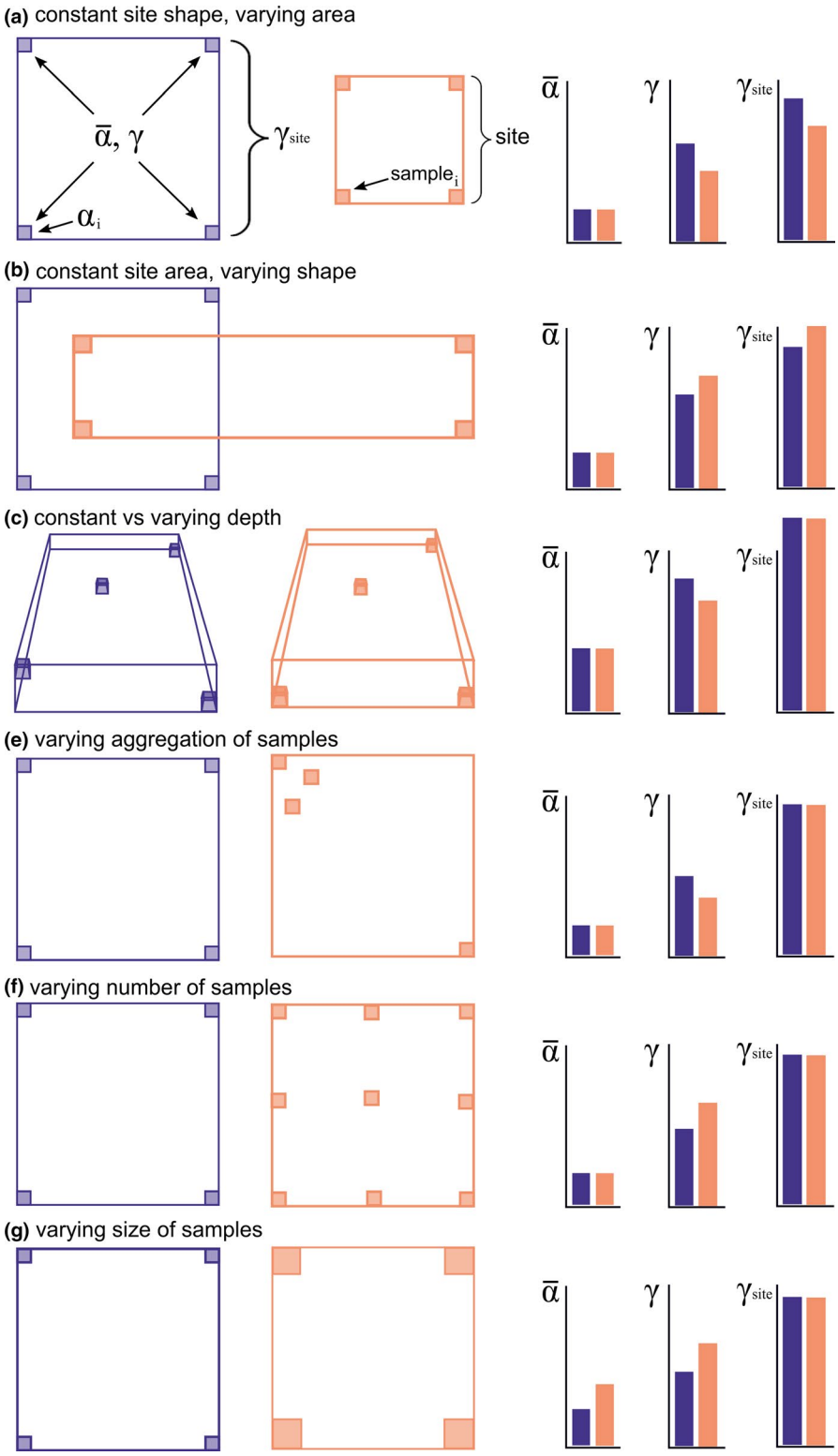

FIGURE 3 Spatial sampling issues that affect average species richness per sample ($\bar{\alpha}$), total richness across all samples (γ), and total richness of an entire site, i.e. both within and outside of the samples (γ_{site}). All of the expected effects stem from two ubiquitous empirical patterns: the increase of number of taxa with increasing area or volume (taxa-area or taxa-volume relationship) and the Tobler's law (a.k.a. distance decay of similarity) that states that closer locations are more similar in their taxonomic composition that distant ones

also much smaller than that traditionally used in morphological assessments of soil fauna, and this may result in estimates of α-diversity that are lower and have higher between-replicate variability (e.g., in nematodes; Wiesel et al., 2014). Indeed, the finding that body size positively correlates with random variation in community structure (Zinger et al., 2019) may be due to the patchiness that arises from observing large organisms with relatively small samples (De Gruyter et al., 2019). Further studies sampling larger volumes (e.g., extracellular DNA extraction, Taberlet et al., 2012; Zinger et al., 2016) are necessary to determine the extent to which β-diversity is inflated in eDNA data targeting larger soil organisms.

3 | MOVING FORWARD

Despite the long list of biases inherent to eukaryotic amplicon sequencing, certain precautions can be taken to mitigate their impact in ecological studies. We propose a three-step approach to using eukaryotic amplicon sequencing data to study soil communities, and recommendations for future experimental designs.

3.1 | Stratify analyses

We advocate for the separation of eukaryotic amplicon sequencing data by group (heretofore stratification) as a way to deal with the wide range of organisms assessed with this technique (Graham et al., 2018). This approach is related to using different marker groups to target specific groups (Oliverio et al., 2020; Tedersoo et al., 2016). Stratification ensures that detection biases do not propagate to the rest of the community and that observations remain comparable within groups. For example, intragenomic polymorphisms of 18S in nematodes can inflate nematode diversity estimates (Dell'Anno et al., 2015). Here, stratifying the data may ensure that the resulting bias does not affect other soil eukaryotes. There is no consensus on the optimal grouping, and this requires careful consideration. Historically, soil biota have been grouped according to physical traits, most notably body size (Orgiazzi et al., 2016). Size-based stratification has a phylogenetic component and may overcome errors associated with body size (i.e., multicellularity), but can ignore finer problems derived from genetic differences, such as primer mismatches. Alternatively, stratifications based strictly on phylogeny may more comprehensively account for errors (e.g., Zinger et al., 2019). Growing evidence suggests that traits, including body size, are phylogenetically conserved across the tree of life (Blomberg et al., 2003; Martiny et al., 2015), and size data may not be available a priori. However, most research on trait conservatism in eukaryotes has focused on plants and vertebrates, and whether relevant morphological and genetic features are conserved in soil eukaryotes (dominated by fungi, protists and invertebrates) requires investigation. Whether stratification is necessary may depend on the research question, as analysing organisms with diverse body sizes with amplicon sequencing may be equivalent to assessing the community through estimates of relative biomass, rather than individual abundances (Elbrecht & Leese, 2015; Schenk et al., 2019; Yoccoz et al., 2012), once accounting for the difference in marker gene copies per cell.

3.2 | Rarefy separately

In addition to the standard issues associated with the ecological analysis of observational data, sequencing data can be further distorted by the amplification and sequencing processes. Amplification artificially and exponentially increases the number of reads in the original sample, and the sequencer used imposes its own limits on the number of reads. Amplicon sequencing data must therefore be standardized prior to statistical analyses (Quinn et al., 2019). However, no consensus on the best methodology for standardizing amplicon sequencing data exists, and the optimal method depends on the ecological questions of interest (McKnight et al., 2019) and the characteristics of the data (Weiss et al., 2017). We advocate for rarefaction (see Gotelli & Colwell, 2001), which randomly resamples observations to the same depth as a sensible compromise. Rarefaction outperforms most bioinformatics methods in compositional analyses (McKnight et al., 2019), and deals well with small sample sizes and variable read depths (Weiss et al., 2017).

When cataloguing all eukaryotes simultaneously, biases arising from morphological and phylogenetic variation may interact to further distort estimates of diversity. We suggest rarefying phylogenetic groups separately, as many of the characteristics which bias abundance estimates (i.e., marker gene copies per cell, organism size, and taxonomic resolution of the marker gene of choice) are phylogenetically conserved (Briones, 2014; Martiny et al., 2015). Consider, for example, a comparison between the diversity in two adjacent soil samples, one which captured a segment of earthworm tissue and a second one which did not. In rarefying both samples to the same observation depth, a high proportion of the reads in the first sample may belong to the earthworm. Consequently, diversity will be underestimated in the first sample relative to the second. Stratifying data may be a good starting point, however further benchmarking is necessary to determine the optimal grouping to reduce biases. Performing rarefactions separately for each group allows the adjustment of the minimum amount observations of individuals for each group, and prevents the propagation of biases across different groups.

3.3 | Consider presence/absence instead of abundance

One way to address the mismatch between number of reads and abundance is to work with binary presence/absence (incidence) data instead of abundances. This approach has been recommended for certain types of ecological data and questions (Beentjes et al., 2018; Delgado-Baquerizo et al., 2020; Elbrecht & Leese, 2015; Ficetola et al., 2015). Incidence data are more common, easier to collect, and potentially more comparable (Alberdi & Gilbert, 2019), as they suffer less from the mismatch between reads and abundances which arises from variable numbers of marker gene copies per cell and cells per organism (Figure 1, Lamb et al., 2019).

Many fundamental ecological variables are derived from incidences, and are robust and practically useful (i.e., species richness, incidence-based β-diversity, and their scaling relationships). For example, in traditional ecology, simple incidence-based number of species (richness) can be predictive of biomass productivity and other ecosystem services (Tilman et al., 2014). It has also been shown that the amount of information about an ecological process can be, in some instances, higher in incidence data than in abundance data

(Bastow Wilson, 2012; Dale et al., 2001). This is because incidences and abundances can be driven by different ecological processes (Orrock et al., 2000; Potts & Elith, 2006; Wenger & Freeman, 2008). For example, while extreme winter temperatures or soil type may be a limiting factor determining presence/absence, the actual abundance may depend on variation in microclimates and microhabitats, which cannot be studied with the common destructive sampling practices (Fierer, 2017). In other words, variation of abundances can mask important ecological correlations at the incidence level (Bastow Wilson, 2012).

While converting amplicon sequence data is the most conservative approach to analysing diversity through amplicon sequencing, it is not perfect. Incidences can inflate the importance of rare OTUs (Deagle et al., 2019), which may be artefacts of sequencing errors or an overly-strict OTU similarity threshold. Furthermore, incidence data may not match research objectives, such as when defining a core diet or microbiome (Alberdi & Gilbert, 2019) or when the measurement of interest is taxon biomass, rather than abundance. Therefore, it is important to ensure that abundance data are retrievable for future reuse. The decision whether to work with abundances or incidences thus depends on the specific hypotheses and research questions.

Using a continuum of diversity measures differentially weighted by abundances, such as Hill numbers (Alberdi & Gilbert, 2019; Hill, 1973) can be a useful alternative. This allows weighting OTUs according to their rarity, offering a mathematically elegant continuum between incidence- and abundance-based measures of biodiversity (Chiu & Chao, 2016). This approach has already been used to quantify diversity in environmental DNA (Calderón-Sanou et al., 2020).

3.4 | Standardize sampling

One particularly underappreciated source of methodological variation is how soil samples are arranged spatially, since biodiversity and community composition vary with area, volume, and distance (Figure 3). A community can be defined for any extent, and all extents may be ecologically meaningful (Chase et al., 2018; Wiesel et al., 2014). However, communities defined at different spatial scales (both grain and extent) cannot be directly compared (Figure 3). In soils, combining multiple subsamples from a defined plot, or pooling, is common, as is insufficient reporting of sampling metadata (Dickie et al., 2018). These two practices make accounting for experimental differences or sampling scale a posteriori impossible. The Earth Microbiome Project (Thompson et al., 2017) has pioneered the large-scale standardization of laboratory protocols and the recording of standard environmental metadata. We argue that additional metadata of the spatial and temporal components of sampling should be reported for each sample. This includes the distance among samples, precise reporting of the spatial location, volume, extent, and grain of sampling, and the time of sample collection. With such information, it will be possible to account for differences in sampling strategy statistically, to compare samples across studies, and study the relationship between sample volume, organism size, and the patterns of diversity detected in the heterogeneous soil matrix.

3.5 | Account for imperfect detection and false positives statistically

Many of the issues described above concern imperfect detection (i.e., the detection of OTUs that are not present in the sample, or the failure to detect present OTUs; Figures 1 and 2).

Several statistical methods have been developed to deal with the complexities of microbiome data (e.g., SparCC, isometric log ratio transforms, and machine learning algorithms, see Knight et al., 2018). One additional solution is occupancy modelling (Guillera-Arroita, 2017), a powerful toolset that accounts for the biases caused by imperfect detection (Figure 4), by adjusting for false presences and false absences (Lahoz-Monfort et al., 2016). Occupancy models are hierarchical statistical models that explicitly separate the observed (biased) detections and the unobserved true presences/absences (i.e., occupancy) or abundances. This is possible by having separate submodels for the true occupancy and the observation process (Figure 4). For the model to work, the data must be informative about the detection process, for example via repeated sampling in time or spatial subsamples (Willoughby et al., 2016). Importantly, there are also variants that can estimate abundance (Kery & Andrew Royle, 2010), work for multispecies (Iknayan et al., 2014), estimate multiple facets of biodiversity (Broms et al., 2015), and can be designed to account for the complex spatial structuring (Chen & Ficetola, 2019). Occupancy modelling also offers an extension that considers the false positive rate (i.e., the rate at which OTUs which were not present in the sample are detected, Lahoz-Monfort et al., 2016; Royle & Link, 2006). However, complex occupancy models must be informed by further experimental work (e.g., from additional PCR calibrations, using confirmation designs, or in the form of Bayesian priors and plausible limits on the probabilities; see Lahoz-Monfort et al., 2016 for a review). The first promising steps have been made towards the application of multispecies occupancy models to various amplicon sequencing data (Doi et al., 2019; Ficetola et al., 2015; McClenaghan et al., 2020), and we argue that this is a good starting point for data analysis of soil eukaryotes.

4 | FUTURE DIRECTIONS: BENCHMARKING AND STANDARDIZING

While sequencing technologies for eukaryotes can be adopted from prokaryote-based techniques, benchmarking and standardization remain to be done. Empirical studies have shown how sample size (Wiesel et al., 2014), extraction method (Griffiths et al., 2018), and the primers used (Schenk et al., 2019) influence diversity estimates in isolated nematode communities, but less is known about how these procedures affect the diversity estimates

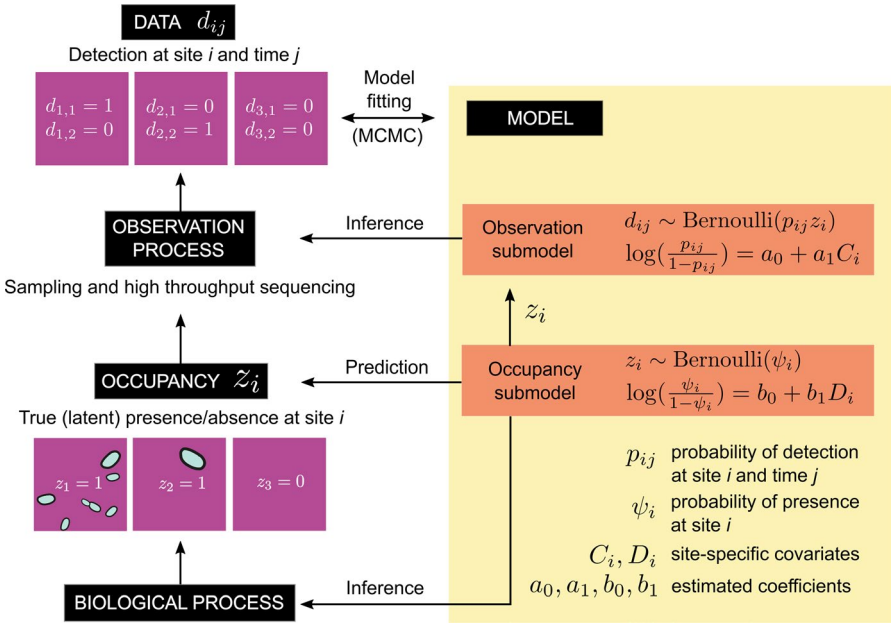

FIGURE 4 Example of a simple occupancy model that accounts for imperfect detectability when estimating the presence/absence of a single species at three sites (based on Kéry, 2010 and Guillera-Arroita, 2017). The model has two parts: (i) an occupancy submodel, which is an ordinary logistic regression of presence/absence against a covariate, and (ii) an observation submodel that estimates detectability of the species thanks to repeated sampling at each site. Both submodels are fitted at the same time, usually by Markov Chain Monte Carlo (MCMC) or maximum likelihood. Royle and Link (2006) and Lahoz-Monfort et al. (2016) provide generalizations of this model that also account for false positives

of other groups of soil fauna (Marquina et al., 2019). One exception is the effect of DNA extraction and storage on diversity assessments, which has received considerable attention (Delmont et al., 2011; Guerrieri et al., 2020; İnceoğlu et al., 2010; Zinger et al., 2016). Benchmarking is a formidable challenge, but it is necessary for soil eukaryotic biodiversity assessments. The accuracy of different universal gene markers needs to be compared for each phylogenetic subgroup of soil biota, especially since most studies currently use single markers for universal diversity assessments. The affinity of any primer pair is likely to be unbalanced across the tree of life, and marker regions differ in their coverage of taxa (Ficetola et al., 2020). One alternative is to simultaneously sequence multiple marker regions, a practice that increases diversity estimates, more closely approximates morphology-based assessments (Meyer et al., 2020), and may become increasingly accessible as the cost of high throughput sequencing continues to decrease (Eberle et al., 2020). Long-read metagenomic shotgun sequencing may also serve to compare diversity assessments performed with different gene segments, and may help uncover novel biota (Eloe-Fadrosh et al., 2016) and further biases, such as those associated with DNA extraction (Jacquiod et al., 2016). Additionally, the sensitivity of the resulting sequence fragments for assigning species identities needs to be determined and reported within each phylogenetic group. In a first step, this can be done in silico using extant genetic repositories of fully sequenced individuals. Such benchmarking efforts are essential to characterizing and quantifying biases, within taxonomic groups as well as across all eukaryotes targeted (Elbrecht & Leese, 2015; Thomas et al., 2016), and may aid in selecting the appropriate phylogenetic grouping for stratifying the data prior to analyses.

Assessing soil communities using amplicon sequencing involves countless choices (e.g., sample size, marker gene, primers), which may affect the resulting output, and its comparability to other data. Another, easier way to ensure comparability is with the development of a standard protocol, such as that proposed by the Earth Microbiome Project (Thompson et al., 2017). However, the standardization of methods across studies may never be perfect, and the continued development of several protocols may maximize the discovery rate. Here, the best solution is to ensure that experimental methods, potential biases and deviations from the standard protocol (for example varying sample size or primer sequence) are always reported in the metadata. Guidelines for the deposition of comprehensive experimental metadata have been proposed (Yilmaz et al., 2011), but these do not require the standardized reporting of spatial sampling designs. The archiving of laboratory protocols (Rambold et al., 2019) may offer a more comprehensive paradigm for data reusability. If sufficient metadata are available, then data integration, meta-analyses, and comparison among studies are possible, as methodological biases can be modelled and accounted for a posteriori using statistical models and meta-analytical machinery (Gerstner et al., 2017).

As the cost of next generation sequencing continues to plummet and its throughput continues to rise, amplicon sequencing will probably become an integral part of soil ecology, filling long-standing gaps in the field and improving our understanding of belowground biota (Cameron et al., 2019). Data created by amplicon sequencing

may be integrated to inform ecological syntheses, and will serve as the record of soil biodiversity for future generations, aiding in the study of the effects of global change on the diversity and dynamics of soil biota. Ensuring that these data remain comparable in the long term is paramount for the both the present and the future of soil ecology.

ACKNOWLEDGEMENTS

This work was funded by the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, which is funded by the German Research Foundation (DFG FZT 118). BKS visit to iDiv was funded by Humboldt Research Award and Australian Research Council (DP190103714). We would like to thank S. Tem for the valuable discussions, and the anonymous reviewers for their constructive feedback. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ORCID

Stephanie D. Jurburg https://orcid.org/0000-0002-7701-6030

REFERENCES

Alberdi, A., & Gilbert, M. T. P. (2019). A guide to the application of Hill numbers to DNA-based diversity analyses. *Molecular Ecology Resources*, *19*(4), 804–817. https://doi.org/10.1111/1755-0998.13014

Bálint, M., Pfenninger, M., Grossart, H.-P., Taberlet, P., Vellend, M., Leibold, M. A., Englund, G., & Bowler, D. (2018). Environmental DNA time series in ecology. *Trends in Ecology and Evolution*, *33*(12), 945–957. https://doi.org/10.1016/j.tree.2018.09.003

Bastida, F., Eldridge, D. J., Abades, S., Alfaro, F. D., Gallardo, A., Garcia-Velazquez, L., Garcia, C., Hart, S. C., Perez, C. A., Santos, F., Trivedi, P., Williams, M. A., & Delgado-Baquerizo, M. (2020). Climatic vulnerabilities and ecological preferences of soil invertebrates across biomes. *Molecular Ecology*, *29*, 752–761. https://doi.org/10.1111/mec.15299

Bastow Wilson, J. (2012). Species presence/absence sometimes represents a plant community as well as species abundances do, or better. *Journal of Vegetation Science*, *23*(6), 1013–1023. https://doi.org/10.1111/j.1654-1103.2012.01430.x

Bates, S. T., Berg-Lyons, D., Caporaso, J. G., Walters, W. A., Knight, R., & Fierer, N. (2011). Examining the global distribution of dominant archaeal populations in soil. *The ISME Journal*, *5*(5), 908–917. https://doi.org/10.1038/ismej.2010.171

Beentjes, K. K., Speksnijder, A. G. C. L., Schilthuizen, M., Schaub, B. E. M., & Van der Hoorn, B. B. (2018). The influence of macroinvertebrate abundance on the assessment of freshwater quality in the Netherlands. *Metabarcoding and Metagenomics*, *2*, 1–8. https://doi.org/10.3897/mbmg.2.26744

Bent, S. J., & Forney, L. J. (2008). The tragedy of the uncommon: Understanding limitations in the analysis of microbial diversity. *The ISME Journal*, *2*(7), 689–695. https://doi.org/10.1038/ismej.2008.44

Bik, H. M., Fournier, D., Sung, W., Bergeron, R. D., & Thomas, W. K. (2013). Intra-genomic variation in the ribosomal repeats of nematodes. *PLoS One*, *8*(10), 1–8. https://doi.org/10.1371/journal.pone.0078230

Blomberg, S. P., Garland, T. Jr., & Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution*, *57*(4), 717–745. https://doi.org/10.1111/j.0014-3820.2003.tb00285.x

Briones, M. J. I. (2014). Soil fauna and soil functions: A jigsaw puzzle. *Frontiers in Environmental Science*, *2*. doi: 7https://doi.org/10.3389/fenvs.2014.00007

Briones, M. J. I. (2018). The serendipitous value of soil fauna in ecosystem functioning: The unexplained explained. *Frontiers in Environmental Science*, *6*, 149. https://doi.org/10.3389/fenvs.2018.00149

Broms, K. M., Hooten, M. B., & Fitzpatrick, R. M. (2015). Accounting for imperfect detection in Hill numbers for biodiversity studies. *Methods in Ecology and Evolution*, *6*(1), 99–108. https://doi.org/10.1111/2041-210X.12296

Calderón-Sanou, I., Münkemüller, T., Boyer, F., Zinger, L., & Thuiller, W. (2020). From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices? *Journal of Biogeography*, *47*(1), 193–206. https://doi.org/10.1111/jbi.13681

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, *11*(12), 2639–2643. https://doi.org/10.1038/ismej.2017.119

Cameron, E. K., Martins, I. S., Lavelle, P., Mathieu, J., Gottschall, F., Guerra, C. A., & Patoine, G. (2019). Global gaps in soil biodiversity data. *Nature Ecology & Evolution*, *2018*(7), 1042–1043. https://doi.org/10.1038/s41559-018-0573-8.Global

Chase, J. M., McGill, B. J., McGlinn, D. J., May, F., Blowes, S. A., Xiao, X., Knight, T. M., Purschke, O., & Gotelli, N. J. (2018). Embracing scale-dependence to achieve a deeper understanding of biodiversity and its change across communities. *Ecology Letters*, *21*(11), 1737–1751. https://doi.org/10.1111/ele.13151

Chen, W., & Ficetola, G. F. (2019). Conditionally autoregressive models improve occupancy analyses of autocorrelated data: An example with environmental DNA. *Molecular Ecology Resources*, *19*(1), 163–175. https://doi.org/10.1111/1755-0998.12949

Chiu, C. H., & Chao, A. (2016). Estimating and comparing microbial diversity in the presence of sequencing errors. *PeerJ*, *4*, e1634. https://doi.org/10.7717/peerj.1634

Compson, Z., McClenaghan, B., Singer, G. A. C., Fahner, N. A., & Hajibabaei, M. (2020). Metabarcoding from microbes to mammals: Comprehensive bioassessment on a global scale. *Frontiers in Ecology and Evolution*, *8*. Article No: 581835. https://doi.org/10.3389/fevo.2020.581835

Dale, M. B., Salmina, L., & Mucina, L. (2001). Minimum message length clustering: An explication and some applications to vegeration data. *Community Ecology*, *2*(2), 231–247. https://doi.org/10.1556/ComEc.2.2001.2.11

Darby, B. J., Todd, T. C., & Herman, M. A. (2013). High-throughput amplicon sequencing of r RNA genes requires a copy number correction to accurately reflect the effects of management practices on soil nematode community structure. *Molecular Ecology*, *22*(21), 5456–5471. https://doi.org/10.1111/mec.12480

De Gruyter, J., Weedon, J. T., Bazot, S., Dauwe, S., Fernandez-Garberí, P.-R., Geisen, S., De La Motte, L. G., Heinesch, B., Janssens, I. A., Leblans, N., Manise, T., Ogaya, R., Löfvenius, M. O., Peñuelas, J., Sigurdsson, B. D., Vincent, G., & Verbruggen, E. (2019). Patterns of local, intercontinental and interseasonal variation of soil bacterial and eukaryotic microbial communities. *FEMS Microbiology Ecology*, *96*(3), 1–12. https://doi.org/10.1093/femsec/fiaa018

Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., & Eveson, J. P. (2019). Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? *Molecular Ecology*, *28*(2), 391–406. https://doi.org/10.1111/mec.14734

Delgado-Baquerizo, M., Reich, P. B., Trivedi, C., Eldridge, D. J., Abades, S., Alfaro, F. D., Bastida, F., Berhe, A. A., Cutler, N. A., Gallardo, A., García-Velázquez, L., Hart, S. C., Hayes, P. E., He, J.-Z., Hseu, Z.-Y., Hu, H.-W., Kirchmair, M., Neuhauser, S., Pérez, C. A., … Singh, B. K. (2020). Multiple elements of soil biodiversity drive ecosystem functions across biomes. *Nature Ecology & Evolution*, *4*, 210–220. https://doi.org/10.1038/s41559-019-1084-y

Dell'Anno, A., Carugati, L., Corinaldesi, C., Riccioni, G., & Danovaro, R. (2015). Unveiling the biodiversity of deep-sea nematodes through metabarcoding: Are we ready to bypass the classical taxonomy? *PLoS One*, *10*(12), e0144928. https://doi.org/10.1371/journal.pone.0144928

Delmont, T. O., Robe, P., Clark, I., Simonet, P., & Vogel, T. M. (2011). Metagenomic comparison of direct and indirect soil DNA extraction approaches. *Journal of Microbiological Methods*, *86*(3), 397–400. https://doi.org/10.1016/j.mimet.2011.06.013

Dickie, I. A., Boyer, S., Buckley, H. L., Duncan, R. P., Gardner, P. P., Hogg, I. D., Holdaway, R. J., Lear, G., Makiola, A., Morales, S. E., Powell, J. R., & Weaver, L. (2018). Towards robust and repeatable sampling methods in eDNA-based studies. *Molecular Ecology Resources*, *18*, 940–952. https://doi.org/10.1111/1755-0998.12907

Doi, H., Fukaya, K., Oka, S., Sato, K., Kondoh, M., & Miya, M. (2019). Evaluation of detection probabilities at the water-filtering and initial PCR steps in environmental DNA metabarcoding using a multispecies site occupancy model. *Scientific Reports*, *9*(1), 1–8. https://doi.org/10.1038/s41598-019-40233-1

Dopheide, A., Tooman, L. K., Grosser, S., Agabiti, B., Rhode, B., Xie, D., Stevens, M. I., Nelson, N., Buckley, T. R., Drummond, A. J., & Newcomb, R. D. (2019). Estimating the biodiversity of terrestrial invertebrates on a forested island using DNA barcodes and metabarcoding data. *Ecological Applications*, *29*(4), e01877. https://doi.org/10.1002/eap.1877

Dungan, J. L., Perry, J. N., Dale, M. R. T., Legendre, P., Citron-Pousty, S., Fortin, M.-J., Jakomulska, A., Miriti, M., & Rosenberg, M. S. (2002). A balanced view of scale in spatial statistical analysis. *Ecography*, *25*(5), 626–640. https://doi.org/10.1034/j.1600-0587.2002.250510.x

Eberle, J., Ahrens, D., Mayer, C., Niehuis, O., & Misof, B. (2020). A plea for standardized nuclear markers in metazoan DNA taxonomy. *Trends in Ecology & Evolution*, *35*, 336–345. https://doi.org/10.1016/j.tree.2019.12.003

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*, *34*(14), 2371–2375. https://doi.org/10.1093/bioinformatics/bty113

Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass - sequence relationships with an innovative metabarcoding protocol. *PLoS One*, *8*(10). Article No: e0130324. https://doi.org/10.1371/journal.pone.0130324

Eloe-Fadrosh, E. A., Ivanova, N. N., Woyke, T., & Kyrpides, N. C. (2016). Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nature Microbiology*, *1*(4), 1–4. https://doi.org/10.1038/nmicrobiol.2015.32

Ficetola, G. F., Boyer, F., Valentini, A., Bonin, A., Meyer, A., Dejean, T., Gaboriaud, C., Usseglio-Polatera, P., & Taberlet, P. (2020). Comparison of markers for the monitoring of freshwater benthic biodiversity through DNA metabarcoding. *Molecular Ecology*, 1–14. https://doi.org/10.1111/mec.15632

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., Gielly, L., Lopes, C. M., Boyer, F., Pompanon, F., Rayé, G., & Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, *15*(3), 543–556. https://doi.org/10.1111/1755-0998.12338

Fierer, N. (2017). Embracing the unknown: Disentangling the complexities of the soil microbiome. *Nature Reviews Microbiology*, *15*(10), 579–590. https://doi.org/10.1038/nrmicro.2017.87

Fierer, N., & Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Science*, *103*(3), 626–631. https://doi.org/10.1073/pnas.0507535103

Fonseca, V. G. (2018). Pitfalls in relative abundance estimation using eDNA metabarcoding. *Molecular Ecology Resources*, *18*(5), 923–926. https://doi.org/10.1111/1755-0998.12902

Geisen, S., Briones, M. J. I., Gan, H., Behan-Pelletier, V. M., Friman, V.-P., de Groot, G. A., Hannula, S. E., Lindo, Z., Philippot, L., Tiunov, A. V., & Wall, D. H. (2019). A methodological framework to embrace soil biodiversity. *Soil Biology and Biochemistry*, *136*, 107536. https://doi.org/10.1016/j.soilbio.2019.107536

Gerstner, K., Moreno-Mateos, D., Gurevitch, J., Beckmann, M., Kambach, S., Jones, H. P., & Seppelt, R. (2017). Will your paper be used in a meta-analysis? Make the reach of your research broader and longer lasting. *Methods in Ecology and Evolution*, *8*(6), 777–784. https://doi.org/10.1111/2041-210X.12758

Gotelli, N. J., & Colwell, R. K. (2001). Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, *4*(4), 379–391. https://doi.org/10.1046/j.1461-0248.2001.00230.x

Graham, C. H., Storch, D., & Machac, A. (2018). Phylogenetic scale in ecology and evolution. *Global Ecology and Biogeography*, *27*(2), 175–187. https://doi.org/10.1111/geb.12686

Green, J. L., Holmes, A. J., Westoby, M., Oliver, I., Briscoe, D., Dangerfield, M., Gillings, M., & Beattie, A. J. (2004). Spatial scaling of microbial eukaryote diversity. *Nature*, *432*(7018), 747–750. https://doi.org/10.1038/nature03034

Griffiths, B. S., de Groot, G. A., Laros, I., Stone, D., & Geisen, S. (2018). The need for standardisation: Exemplified by a description of the diversity, community structure and ecological indices of soil nematodes. *Ecological Indicators*, *87*, 43–46. https://doi.org/10.1016/j.ecolind.2017.12.002

Guerrieri, A., Bonin, A., Münkemüller, T., Gielly, L., Thuiller, W., & Francesco Ficetola, G. (2020). Effects of soil preservation for biodiversity monitoring using environmental DNA. *Molecular Ecology*. 1–13. https://doi.org/10.1111/mec.15674

Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: Advances, challenges and opportunities. *Ecography*, *40*(2), 281–295. https://doi.org/10.1111/ecog.02445

Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, *54*(2), 427–432. https://doi.org/10.2307/1934352

Iknayan, K. J., Tingley, M. W., Furnas, B. J., & Beissinger, S. R. (2014). Detecting diversity: Emerging methods to estimate species diversity. *Trends in Ecology & Evolution*, *29*(2), 97–106. https://doi.org/10.1016/j.tree.2013.10.012

İnceoğlu, Ö., Hoogwout, E. F., Hill, P., & van Elsas, J. D. (2010). Effect of DNA extraction method on the apparent microbial diversity of soil. *Applied and Environmental Microbiology*, *76*(10), 3378–3382. https://doi.org/10.1128/AEM.02715-09

Jacquiod, S., Stenbæk, J., Santos, S. S., Winding, A., Sørensen, S. J., & Priemé, A. (2016). Metagenomes provide valuable comparative information on soil microeukaryotes. *Research in Microbiology*, *167*(5), 436–450. https://doi.org/10.1016/j.resmic.2016.03.003

Kéry, M. (2010). *Introduction to WinBUGS for ecologists: Bayesian approach to regression, ANOVA, mixed models and related analyses*. Academic Press.

Kery, M., & Andrew Royle, J. (2010). Hierarchical modelling and estimation of abundance and population trends in metapopulation designs. *Journal of Animal Ecology*, *79*(2), 453–461. https://doi.org/10.1111/j.1365-2656.2009.01632.x

Kirchman, D. L. (2018). *Processes in microbial ecology*. Oxford University Press.

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., Melnik,

A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., … Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, *16*(7), 410–422. https://doi.org/10.1038/s41579-018-0029-9

Lahoz-Monfort, J. J., Guillera-Arroita, G., & Tingley, R. (2016). Statistical approaches to account for false-positive errors in environmental DNA samples. *Molecular Ecology Resources*, *16*(3), 673–685. https://doi.org/10.1111/1755-0998.12486

Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., & Taylor, M. I. (2019). How quantitative is metabarcoding: A meta-analytical approach. *Molecular Ecology*, *28*(2), 420–430. https://doi.org/10.1111/mec.14920

Leasi, F., Sevigny, J. L., Laflamme, E. M., Artois, T., Curini-Galletti, M., de Jesus Navarrete, A., Di Domenico, M., Goetz, F., Hall, J. A., Hochberg, R., Jörger, K. M., Jondelius, U., Todaro, M. A., Wirshing, H. H., Norenburg, J. L., & Thomas, W. K. (2018). Biodiversity estimates and ecological interpretations of meiofaunal communities are biased by the taxonomic approach. *Communications Biology*, *1*(1), 112. https://doi.org/10.1038/s42003-018-0119-2

Louca, S., Doebeli, M., & Parfrey, L. W. (2018). Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome*, *6*(1), 41. https://doi.org/10.1186/s40168-018-0420-9

Marquina, D., Esparza-Salas, R., Roslin, T., & Ronquist, F. (2019). Establishing arthropod community composition using metabarcoding: Surprising inconsistencies between soil samples and preservative ethanol and homogenate from Malaise trap catches. *Molecular Ecology Resources*, *19*(6), 1516–1530. https://doi.org/10.1111/1755-0998.13071

Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., Morin, P. J., Naeem, S., Øvreås, L., Reysenbach, A.-L., Smith, V. H., & Staley, J. T. (2006). Microbial biogeography: Putting microorganisms on the map. *Nature Reviews Microbiology*, *4*(2), 102–112. https://doi.org/10.1038/nrmicro1341

Martiny, J. B., Jones, S. E., Lennon, J. T., & Martiny, A. C. (2015). Microbiomes in light of traits: A phylogenetic perspective. *Science*, *350*(6261), aac9323-1–aac9323-8. https://doi.org/10.1126/science.aac9323

McClenaghan, B., Compson, Z. G., & Hajibabaei, M. (2020). Validating metabarcoding-based biodiversity assessments with multi-species occupancy models: A case study using coastal marine eDNA. *PLoS One*, *15*(3), e0224119. https://doi.org/10.1371/journal.pone.0224119

McKnight, D. T., Huerlimann, R., Bower, D. S., Schwarzkopf, L., Alford, R. A., & Zenger, K. R. (2019). Methods for normalizing microbiome data: An ecological perspective. *Methods in Ecology and Evolution*, *10*(3), 389–400. https://doi.org/10.1111/2041-210X.13115

Meyer, A., Boyer, F., Valentini, A., Bonin, A., Ficetola, G. F., Beisel, J.-N., Usseglio-Polatera, P. (2020). Morphological vs. DNA metabarcoding approaches for the evaluation of stream ecological status with benthic invertebrates: Testing different combinations of markers and strategies of data filtering. *Molecular Ecology*. 1–18. https://doi.org/10.1111/mec.15723

Mysara, M., Vandamme, P., Props, R., Kerckhof, F.-M., Leys, N., Boon, N., Raes, J., & Monsieurs, P. (2017). Reconciliation between operational taxonomic units and species boundaries. *FEMS Microbiology Ecology*, *93*(4). 1–12. https://doi.org/10.1093/femsec/fix029

Nekola, J. C., & White, P. S. (1999). The distance decay of similarity in biogeography and ecology. *Journal of Biogeography*, *26*(4), 867–878. https://doi.org/10.1046/j.1365-2699.1999.00305.x

Nichols, R. V., Vollmers, C., Newsom, L. A., Wang, Y., Heintzman, P. D., Leighton, M., & Shapiro, B. (2018). Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources*, *18*(5), 927–939. https://doi.org/10.1111/1755-0998.12895

Oliverio, A. M., Geisen, S., Delgado-Baquerizo, M., Maestre, F. T., Turner, B. L., & Fierer, N. (2020). The global-scale distributions of soil protists and their contributions to belowground systems. *Science Advances*, *6*(4), 1–11. https://doi.org/10.1126/sciadv.aax8787

Orgiazzi, A., Bardgett, R. D., & Barrios, E. (2016). *Global soil biodiversity atlas*. European Commission.

Orgiazzi, A., Bonnet, M., Panagos, P., Arjen, G., Groot, D., & Lemanceau, P. (2015). Soil biodiversity and DNA barcodes: Opportunities and challenges. *Soil Biology and Biochemistry*, *80*, 244–250. https://doi.org/10.1016/j.soilbio.2014.10.014

Orrock, J. L., Pagels, J. F., McShea, W. J., & Harper, E. K. (2000). Predicting presence and abundance of a small mammal species: The effect of scale and resolution. *Ecological Applications*, *10*(5), 1356–1366. https://doi.org/10.2307/2641291

Pawlowski, J., Apothéloz-Perret-Gentil, L., & Altermatt, F. (2020). Environmental DNA: What's behind the term? Clarifying the terminology and recommendations for its future use in biomonitoring. *Molecular Ecology*, *29*(22), 4258–4264. https://doi.org/10.1111/mec.15643

Potts, J. M., & Elith, J. (2006). Comparing species abundance models. *Ecological Modelling*, *199*(2), 153–163. https://doi.org/10.1016/j.ecolmodel.2006.05.025

Quinn, T. P., Erb, I., Gloor, G., Notredame, C., Richardson, M. F., & Crowley, T. M. (2019). A field guide for the compositional analysis of any-omics data. *GigaScience*, *8*(9), 1–14. https://doi.org/10.1093/gigascience/giz107

Rambold, G., Yilmaz, P., Harjes, J., Klaster, S., Sanz, V., Link, A., Glöckner, F. O., & Triebel, D. (2019). Meta-omics data and collection objects (MOD-CO): A conceptual schema and data model for processing sample data in meta-omics research. *Database*. *2019*, 1–13. https://doi.org/10.1093/database/baz002

Ramirez, K. S., Leff, J. W., Barberán, A., Bates, S. T., Betley, J., Crowther, T. W., Kelly, E. F., Oldfield, E. E., Shaw, E. A., Steenbock, C., Bradford, M. A., Wall, D. H., & Fierer, N. (2014). Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. *Proceedings of the Royal Society B*, *281*(1795). Article No: 20141988. https://doi.org/10.1098/rspb.2014.1988

Robeson, M. S., King, A. J., Freeman, K. R., Birky, C. W., Martin, A. P., & Schmidt, S. K. (2011). Soil rotifer communities are extremely diverse globally but spatially autocorrelated locally. *Proceedings of the National Academy of Sciences*, *108*(11), 4406–4410. https://doi.org/10.1073/pnas.1012678108

Rodriguez-R, L. M., Castro, J. C., Kyrpides, N. C., Cole, J. R., Tiedje, J. M., & Konstantinidis, K. T. (2018). How much do rRNA gene surveys underestimate extant bacterial diversity? *Applied and Environmental Microbiology*, *84*(6), e00014–18. https://doi.org/10.1128/AEM.00014-18

Royle, J. A., & Link, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, *87*(4), 835–841.

Ruppert, K. M., Kline, R. J., & Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, *17*. Article No: e00547. https://doi.org/10.1016/j.gecco.2019.e00547

Santos, S. S., Nunes, I., Nielsen, T. K., Jacquiod, S., Hansen, L. H., & Winding, A. (2017). Soil DNA extraction procedure influences protist 18S rRNA gene community profiling outcome. *Annals of Anatomy*, *168*(3), 283–293. https://doi.org/10.1016/j.protis.2017.03.002

Schenk, J., Geisen, S., Kleinboelting, N., & Traunspurger, W. (2019). Metabarcoding data allow for reliable biomass estimates in the most abundant animals on earth. *Metabarcoding and Metagenomics*, *3*, 117–126. https://doi.org/10.3897/mbmg.3.46704

Shade, A., Dunn, R. R., Blowes, S. A., Keil, P., Bohannan, B. J. M., Herrmann, M., Küsel, K., Lennon, J. T., Sanders, N. J., Storch, D.,

& Chase, J. (2018). Macroecology to unite all life, large and small. *Trends in Ecology and Evolution*, *33*(10), 731–744. https://doi.org/10.1016/j.tree.2018.08.005

Steen, A. D., Crits-Christoph, A., Carini, P., DeAngelis, K. M., Fierer, N., Lloyd, K. G., & Cameron Thrash, J. (2019). High proportions of bacteria and archaea across most biomes remain uncultured. *The ISME Journal*, *13*(12), 3126–3130. https://doi.org/10.1038/s41396-019-0484-y

Swift, M. J., Heal, O. W., Anderson, J. M., & Anderson, J. M. (1979). *Decomposition in terrestrial ecosystems*, Vol. *5*. University of California Press.

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: For biodiversity research and monitoring*. Oxford University Press.

Taberlet, P., Prud'homme, S. M., Campione, E., Roy, J., Miquel, C., Shehzad, W., Gielly, L., Rioux, D., Choler, P., Clément, J.-C., Melodelima, C., Pompanon, F., & Coissac, E. (2012). Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. *Molecular Ecology*, *21*(8), 1816–1820. https://doi.org/10.1111/j.1365-294X.2011.05317.x

Tang, C. Q., Leasi, F., Obertegger, U., Kieneke, A., Barraclough, T. G., & Fontaneto, D. (2012). The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Science*, *109*(40), 16208–16212. https://doi.org/10.1073/pnas.1209160109

Tedersoo, L., Bahram, M., Cajthaml, T., Põlme, S., Hiiesalu, I., Anslan, S., Harend, H., Buegger, F., Pritsch, K., Koricheva, J., & Abarenkov, K. (2016). Tree diversity and species identity effects on soil fungi, protists and animals are context dependent. *ISME Journal*, *10*(2), 346–362. https://doi.org/10.1038/ismej.2015.116

Tedersoo, L., Bahram, M., Põlme, S., Kõljalg, U., Yorou, N. S., Wijesundera, R., Ruiz, L. V., Vasco-Palacios, A. M., Thu, P. Q., Suija, A., Smith, M. E., Sharp, C., Saluveer, E., Saitta, A., Rosas, M., Riit, T., Ratkowsky, D., Pritsch, K., Põldmaa, K., … Abarenkov, K. (2014). Global diversity and geography of soil fungi. *Science*, *346*(6213), 1256688. https://doi.org/10.1126/science.1256688

Thakur, M. P., Phillips, H. R. P., Brose, U., De Vries, F. T., Lavelle, P., Loreau, M., Mathieu, J., Mulder, C., Van der Putten, W. H., Rillig, M. C., Wardle, D. A., Bach, E. M., Bartz, M. L. C., Bennett, J. M., Briones, M. J. I., Brown, G., Decaëns, T., Eisenhauer, N., Ferlian, O., … Cameron, E. K. (2019). Towards an integrative understanding of soil biodiversity. *Biological Reviews*, *95*, 350–364. https://doi.org/10.1111/brv.12567

Thomas, A. C., Deagle, B. E., Eveson, J. P., Harsch, C. H., & Trites, A. W. (2016). Quantitative DNA metabarcoding: Improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources*, *16*, 714–726. https://doi.org/10.1111/1755-0998.12490

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., Prill, R. J., Tripathi, A., Gibbons, S. M., Ackermann, G., Navas-Molina, J. A., Janssen, S., Kopylova, E., Vázquez-Baeza, Y., González, A., Morton, J. T., Mirarab, S., Zech Xu, Z., Jiang, L., … Consortium, T. E. M. P. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, *551*, 457–463. https://doi.org/10.1038/nature24621

Thornhill, D. J., Lajeunesse, T. C., & Santos, S. R. (2007). Measuring rDNA diversity in eukaryotic microbial systems: How intragenomic variation, pseudogenes, and PCR artifacts confound biodiversity estimates. *Molecular Ecology*, *16*(24), 5326–5340. https://doi.org/10.1111/j.1365-294X.2007.03576.x

Tilman, D., Isbell, F., & Cowles, J. M. (2014). Biodiversity and ecosystem functioning. *Annual Review of Ecology, Evolution, and Systematics*, *45*, 471–493. https://doi.org/10.1146/annurev-ecolsys-120213-091917

Veltri, K. L., Espiritu, M., & Singh, G. (1990). Distinct genomic copy number in mitochondria of different mammalian organs. *Journal of Cellular Physiology*, *143*(1), 160–164. https://doi.org/10.1002/jcp.1041430122

Větrovský, T., Kohout, P., Kopecký, M., Machac, A., Man, M., Bahnmann, B. D., Brabcová, V., Choi, J., Meszárošová, L., Human, Z. R., Lepinay, C., Lladó, S., López-Mondéjar, R., Martinović, T., Mašínová, T., Morais, D., Navrátilová, D., Odriozola, I., Štursová, M., … Baldrian, P. (2019). A meta-analysis of global fungal distribution reveals climate-driven patterns. *Nature Communications*, *10*(1), 1–9. https://doi.org/10.1038/s41467-019-13164-8

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., & Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1), 27. https://doi.org/10.1186/s40168-017-0237-y

Wenger, S. J., & Freeman, M. C. (2008). Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Ecology*, *89*(10), 2953–2959. https://doi.org/10.1890/07-1127.1

Wiesel, L., Daniell, T. J., King, D., & Neilson, R. (2014). Determination of the optimal soil sample size to accurately characterise nematode communities in soil. *Soil Biology and Biochemistry*, *80*, 89–91. https://doi.org/10.1016/j.soilbio.2014.09.026

Willoughby, J. R., Wijayawardena, B. K., Sundaram, M., Swihart, R. K., & DeWoody, J. A. (2016). The importance of including imperfect detection models in eDNA experimental design. *Molecular Ecology Resources*, *16*(4), 837–844. https://doi.org/10.1111/1755-0998.12531

Woodcock, S., Curtis, T. P., Head, I. M., Lunn, M., & Sloan, W. T. (2006). Taxa-area relationships for microbes: The unsampled and the unseen. *Ecology Letters*, *9*(7), 805–812. https://doi.org/10.1111/j.1461-0248.2006.00929.x

Wu, T., Ayres, E., Bardgett, R. D., Wall, D. H., & Garey, J. R. (2011). Molecular study of worldwide distribution and diversity of soil animals. *Proceedings of the National Academy of Science*, *108*(43), 17720–17725. https://doi.org/10.1073/pnas.1103824108

Wu, Z.-W., Wang, Q.-M., Liu, X.-Z., & Bai, F.-Y. (2016). Intragenomic polymorphism and intergenomic recombination in the ribosomal RNA genes of strains belonging to a yeast species Pichia membranifaciens. *Mycology*, *7*(3), 102–111. https://doi.org/10.1080/21501203.2016.1204369

Yergeau, E., Bokhorst, S., Huiskes, A. H. L., Boschker, H. T. S., Aerts, R., & Kowalchuk, G. A. (2007). Size and structure of bacterial, fungal and nematode communities along an Antarctic environmental gradient. *FEMS Microbiology Ecology*, *59*(2), 436–451. https://doi.org/10.1111/j.1574-6941.2006.00200.x

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., Gilbert, J. A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., … Glöckner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, *29*(5), 415–420. https://doi.org/10.1038/nbt.1823

Yoccoz, N. G., Bråthen, K. A., Gielly, L., Haile, J., Edwards, M. E., Goslar, T., Von Stedingk, H., Brysting, A. K., Coissac, E., Pompanon, F., Sønstebø, J. H., Miquel, C., Valentini, A., De Bello, F., Chave, J., Thuiller, W., Wincker, P., Cruaud, C., Gavory, F., … Taberlet, P. (2012). DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology*, *21*(15), 3647–3655. https://doi.org/10.1111/j.1365-294X.2012.05545.x

Zinger, L., Chave, J., Coissac, E., Iribar, A., Louisanna, E., Manzi, S., Schilling, V., Schimann, H., Sommeria-Klein, G., & Taberlet, P. (2016). Extracellular DNA extraction is a fast, cheap and reliable alternative for multi-taxa surveys based on soil DNA. *Soil Biology*

and Biochemistry, *96*, 16–19. https://doi.org/10.1016/j.soilbio.2016.01.008

Zinger, L., Taberlet, P., Schimann, H., Bonin, A., Boyer, F., De Barba, M., Gaucher, P., Gielly, L., Giguet-Covex, C., Iribar, A., Réjou-Méchain, M., Rayé, G., Rioux, D., Schilling, V., Tymen, B., Viers, J., Zouiten, C., Thuiller, W., Coissac, E., & Chave, J. (2019). Body size determines soil community assembly in a tropical forest. *Molecular Ecology*, *28*(3), 528–543. https://doi.org/10.1111/mec.14919

How to cite this article: 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;21:1759–1771. https://doi.org/10.1111/1755-0998.13401