The quest for absolute abundance: the use of internal standards for DNA-barcoding in microbial ecology

Joshua G. Harrison[*1,2], W. John Calder[*1], Bryan Shuman[1], and C. Alex Buerkle[1]

[1]University of Wyoming,
Laramie, WY 82071, USA

[*] These authors contributed equally to this manuscript.

[2]*Corresponding author*: Joshua G. Harrison
1000 E. University Ave.
Department of Botany, 3165
University of Wyoming
Laramie, WY 82071, USA
joshua.harrison@uwyo.edu
Fax: 307-766-2851

Keywords: microbiome, internal standard, spike-in, compositional data, absolute abundances, relative abundances, microbial ecology

Running title: Internal standards

Abstract

To characterize microbiomes, microbial ecologists routinely sequence and compare short loci that differ among focal taxa. Counts of these sequences convey information regarding the occurrence and relative abundances of taxa in an assemblage, but provide no direct measure of their absolute abundances, due to the limitations of the sequencing process. The relative abundances in compositional data are inherently constrained and difficult to interpret. The incorporation of internal standards (ISDs; colloquially referred to as "spike-ins") into DNA pools for sequencing can ameliorate the problems posed by relative abundance data and allow absolute abundances to be approximated. Unfortunately, many laboratory and sampling biases cause ISDs to underperform or fail. Here, we discuss how careful deployment of ISDs can avoid these complications and be an integral component of well-designed, amplicon-based studies of microbial ecology.

Introduction

Microbial assemblages are routinely characterized by DNA sequencing of marker loci, which are typically short and are chosen because they vary among focal taxa (Caporaso et al. 2012; Carini 2019; Goodrich et al. 2014)—portions of the ribosomal RNA operon are particularly popular markers. Characterizing assemblages in this way is referred to as metabarcoding (Schmidt et al. 2013; Taberlet et al. 2012). Qualitative differences in the sequences obtained from a metabarcoding study can be used to generate hypotheses regarding the types of organisms present in an assemblage, but understanding the abundances of each of these organisms from sequence data alone has proven extremely challenging. This is because sequencing methods yield a finite number of sequences per operational period, which are then parsed among samples and molecules within each sample. Thus, DNA sequencing can only provide direct knowledge of the relative abundances of organisms, not their absolute abundances.

Analyzing relative abundances is challenging for several reasons. First, biological insights often depend on knowledge of absolute abundances. For instance, in a study of the faecal microbiome of patients with Crohn's disease, absolute abundance data (obtained through flow cytometry) revealed that bacterial load was associated with disease phenotype—an unobtainable result when using relative abundance data. More generally, dramatically different results were obtained from analyses of absolute versus relative abundance data. For example, the use of absolute abundance data led to detection of 76 covarying microbial genera, compared to detection of only 10 covarying genera when using relative abundance information. Relative abundance data were misleading about microbial richness, rank abundances, and associations of specific taxa with disease phenotype (Vandeputte et al. 2017)—thus demonstrating that relative abundance data are unsuitable for addressing many biological questions.

The second, and more insidious, problem with relative abundances is that they are compositional (Aitchison 1982), that is, as one taxon increases within a sample, it does so relative to some other taxon (or taxa) that must decrease. For over a hundred years, mathematicians have been aware of the numerous problems associated with the analysis of compositional data (Pearson 1897) and several sub-fields of ecology have developed rich literatures about these complications (Jackson 1997). In some cases, disciplinary names for the challenges of compositionality are used, such as the 'Fagerlind effect' (i.e. a term used in paleoecology to refer to the problems inherent to the analysis of compositional pollen data), which complicate cross-disciplinary transfer of relevant information (Davis 1963; Fagerlind 1952; Prentice and Webb 1986). Nevertheless, acknowledgement of the constraints imposed by compositional data is becoming more commonplace among microbial ecologists (Gloor and Reid 2016; Weiss et al. 2017). However many studies still do not adequately confront the problem of compositionality and are hampered by the limitations of relative abundance data.

A variety of statistical transformations involving log ratios have been suggested to address the problems of compositionality, the most common being the centered log ratio transformation (Aitchison 1982; Egozcue et al. 2003; Fernandes et al. 2014; Gloor et al. 2017). However, the benefits of these transformations are limited for high-dimensional, sparse data (data with many zeros, such as those describing assemblages with numerous rare taxa, which may not be observed at all in many samples), such as those characterizing microbial biodiversity (for more see Tsilimigras and Fodor 2016). Furthermore, the transformations alone do not allow for the conversion of relative abundance estimates to absolute abundances.

A promising solution to these problems is the incorporation of an internal standard (ISD) into the DNA sequencing process (Chen et al. 2016; Hossain et al. 2020; Jiang et al. 2011; Smets et al. 2016; Tourlousse et al. 2017; Zemb et al. 2020). Similar approaches to spiking samples with an ISD have been applied in other disciplines seeking absolute abundances (e.g. paleoecology; Benninghoff 1962; Davis 1966; Davis and Deevey 1964), although complications have caused some communities to abandon methods of calculating absolute counts (Giesecke and Fontana 2008). In microbial analyses, the relevant ISD is a unique molecule (or cell, see below) that is added to all samples in a known absolute abundance (i.e., as measured in cells or moles). Through comparison to the ISD, the relative abundances of other sequenced features can be converted to units of absolute abundance. ISDs are powerful tools that are rapidly gaining attention, but they are still not routinely used by microbial ecologists (Chen and Li 2013; Fernandes et al. 2014; Gloor et al. 2017). As ISDs become regarded as critical components of a well-designed sequencing study (Chen et al. 2016; Jones et al. 2015), there is a need for clear understanding of the many commonly-encountered sampling scenarios and the laboratory biases that can undercut the efficacy of the standards. Here, we describe these considerations and suggest best practices for the design and use of ISDs.

How does an internal standard work?

The potential benefit of ISDs is that they allow the conversion of relative abundances into absolute abundances. To see why this is desirable and why relative abundances in compositional data are problematic, consider a hypothetical comparison of two microbiome samples. The first sample contains two equally-abundant microbial taxa and the second sample contains the same two taxa, but their relative abundances have shifted such that one is more abundant than the other. We could represent sequence data for these samples as vectors of proportions, with the first sample consisting of two equally abundant elements with proportions that sum to one $\vec{p}_1 = [0.5, 0.5]$. Whereas, the second sample has unequal elements, but the proportions also sum to one, e.g.: $\vec{p}_2 = [0.7, 0.3]$. The fact that both vectors must share the same sum (1 in this case) is referred to as the "constant sum constraint" of compositional data (Gloor et al. 2017) and is why neither of these vectors, nor the underlying sequence data, contain direct information regarding the absolute abundances of the microbial taxa being examined. For instance, it is impossible to know why in sample two the first microbe is greater in relative abundance compared to sample one. The difference could be due to the first taxon truly having a higher absolute abundance in sample two than in sample one. But it could also be due to a *decrease* in the second microbial taxon, or some combination of both possibilities, because the constant sum constraint of relative abundance data must be satisfied.

This conundrum can potentially be resolved if a known quantity of a third microbial taxon is added to each sample as an ISD. Continuing with the previous example, we could include an ISD as the third element of each sample. After adding the *same* number of cells of the ISD to both microbial samples and repeating the sequencing process, one might obtain a proportion vector for sample one of: $\vec{p}_1 = [0.45, 0.45, 0.1]$, and for sample two of $\vec{p}_2 = [0.7, 0.25, 0.05]$ (the proportion taken by the ISD, the third number, could take any non-zero value). Because the same cell count of ISD was added to each sample, calculating the ratio of microbial relative abundances to the relative abundance of the ISD transforms the relative abundances making them proportional to absolute abundances, with units of the ISD (Fig. 1). In the example, on the scale of the ISD, the absolute abundances in sample one are $[4.5, 4.5, 1]$ and in sample two are $[14, 5, 1]$. We found that for every unit of ISD we observed 14 of the first microbial taxon in sample two, but only 4.5 in sample one, indicating that the first microbial taxon is present at higher absolute abundance in sample two. The second microbial taxon also increased in abundance in sample two compared to sample one, but did not do so as much as the first taxon. Absolute abundances in units of the ISD can be scaled appropriately to other units by knowing the amount of standard that was added (the number of cells, or the number of moles of a DNA molecule).

What type of internal standard should be used?

Two main approaches exist for using ISDs in sequencing studies. The first involves adding a foreign molecule (or cell) to samples to be sequenced; we will refer to this method as a “spike-in” ISD. Alternatively, invariant features already present within samples can be used; we will refer to this type of ISD as an “inherent” ISD.

Researchers studying gene expression have long relied on inherent ISDs to facilitate comparison of transcription levels across samples (reviewed by Eisenberg and Levanon 2013; Thellin et al. 1999). Inherent ISDs are chosen from among those genes that contribute to the basic functioning of the cell (“housekeeping” genes) and are thus expected to be constitutively expressed. The idea is that these genes constantly produce the same number of transcripts, thus reads from them can be used as a baseline when comparing the expression levels of other genes among samples. Identifying housekeeping genes that are suitable for use as inherent ISDs is challenging and highly system-dependent because constitutively expressed genes differ among organisms and tissues. Moreover, the assumption that housekeeping genes do not vary in expression among focal tissues is often violated (Eisenberg and Levanon 2013; Jonge et al. 2007; Lun et al. 2017; Thellin et al. 1999; Tricarico et al. 2002). These drawbacks have led many geneticists away from inherent ISDs and toward spike-in standards (Chen and Li 2013; Jiang et al. 2011). For the same reason, inherent ISDs are inappropriate for molecular community ecology—no taxon is expected to exist at identical abundances among habitats.

Developing and using a spike-in ISD is not without its own challenges, however, because a successful ISD must satisfy the following assumptions: 1.) the ISD must behave similarly to template nucleic acids during laboratory practices, a characteristic referred to as "commutability" (Hardwick et al. 2017; Risso et al. 2014); and, 2.) there can be no chance that the ISD can be mistaken for a feature naturally occurring in samples.

The development of molecular spike-in ISDs was pioneered by functional geneticists interested in gene expression (e.g., Jiang et al. 2011) and microbial ecologists can learn much from their work. Pools of RNA represent particularly complex chemical mixtures because transcripts can differ dramatically in length, nucleotide composition (e.g., GC content, repeat density), and concentration (Lynch 2007; Oshlack and Wakefield 2009; Risso et al. 2011). Moreover, alternative splicing of transcripts leads to multiple isoforms. Given this complexity, no ISD will mirror the behavior of all transcripts present within even a single cell during laboratory preparation. Accordingly, the External RNA Controls Consortium (ERCC) developed an ISD mixture comprising 92 RNA sequences that vary in length from 250–2000 nucleotides, differ dramatically in GC content, and that span a concentration range of 2^{20} (Jiang et al. 2011; also see Hardwick et al. 2016 and Hardwick et al. 2018). Even such a thorough approach has its limitations—Risso et al. 2014 reported unsatisfactorily high technical variation upon sequencing the ISD mixture (also see Qing et al. 2013). Accordingly, Risso et al. 2014 suggested a statistical modeling approach to estimate and remove unwanted technical variation as informed by ISD read counts (see below for more regarding the benefits of such modeling).

The challenges facing microbial ecologists are somewhat less daunting than those with which functional geneticists must contend—this is because among-amplicon variation for commonly used microbial marker loci is typically much lower than what would be expected within a pool of RNA, given that transcripts can vary by over ten thousand nucleotides (nt) in length (Oshlack and Wakefield 2009). By comparison, for many bacterial taxa the 16s marker gene is approximately 1,500 nt long (Bibby et al. 2010; Case et al. 2007; Clarridge

2004), and often a smaller subunit is amplified for sequencing. The ITS operon, which is the typical marker for fungal ecology, is more complex—among taxa it can vary in length by several orders of magnitude (Schoch et al. 2012; Stewart and Cavanaugh 2007). But an ITS amplicon pool will still contain less among-sequence variation than an RNA pool (Lynch 2007). Consequently, ISD solutions tailored for molecular community ecology can be relatively simple and typically consist of adding a known DNA sequence or cells from a specific microbial taxon to samples.

One of the first studies to demonstrate the benefits of ISDs for microbial ecology was Stämmler et al. 2016. These researchers suggested using cells of several halophilic bacterial taxa and one bacterial taxon that occurs in the plant rhizosphere as ISDs for studies of the mammalian faecal microbiome (also see Piwosz et al. 2018). This approach has the important benefit of measuring potential variation in extraction performance among samples, which is likely to dramatically improve ISD commutability for many substrates (see below). The downsides to cellular ISDs are two-fold: first, choosing a cellular ISD can be challenging because it must have similar traits to focal organisms, be easily cultured (or available commercially), and cannot occur in the biological samples. Second, a mixed culture of a cellular ISD could possess copy number variation (CNV) in marker loci that must be measured and accounted for, else the ISD will not provide consistent and accurate absolute abundance estimates (Kembel et al. 2012). For well-known taxa, estimates of CNV for marker loci could be obtained from published genomic resources (Stoddard et al. 2015) or, for less studied taxa, quantitative PCR (qPCR) could be used to estimate copy number per cell. Likewise, clonal propagation of cellular ISDs could minimize CNV for marker loci.

An alternative approach to cellular ISDs is the use of DNA molecules. Many microbial ecologists have advocated DNA ISDs, either in the form of extracted genomic DNA from organisms not likely to be present in samples or as synthetically designed molecules (Hardwick et al. 2018; Lin et al. 2019; Smets et al. 2016; Tkacz et al. 2018; Tourlousse et al. 2017; Venkataraman et al. 2018; Yang et al. 2018; Zemb et al. 2020). We suggest that synthetic sequences are superior to biologically-derived DNA for several reasons. First, and most obviously, there is no chance a synthetic sequence will occur naturally in samples, regardless of sample type. Second, reference DNA for a standard that is isolated from the genome could correspond to a variable number of genomic loci (CNV; as would actual cells; see above) and accounting for this potential variation among different isolates of a standard would require additional laboratory work, such as qPCR. Third, the nucleotide composition of an extracted DNA sequence is fixed and will likely only be commutable to a subset of focal taxa. By comparison, a synthetic ISD's DNA sequence can be specified such that it is comparable to the nucleotide composition of any organism (e.g., in length, GC content, repeat density, etc.) and thus could be tailored to fit the specific needs of a study.

The design of a synthetic ISD is fairly simple. The primary requirement is that the sequence cannot match any known organisms and is long enough that it will not be removed during PCR clean up (e.g., when using size selection to remove excess primer molecules). If a generic ISD is desired, then the sequence should minimize homopolymers and internal complementarity, have balanced GC content, and be approximately the same length as the focal barcoding locus (see Tourlousse et al. 2017, for guidance). After designing the ISD sequence it must be bracketed by the preferred primer pair, with the complement of the forward primer at the beginning of the read and the uncomplemented reverse primer appended to the read (assuming single stranded synthesis). A variety of ISD designs are present in the literature and can be inexpensively synthesized by various commercial suppliers (Palmer et al. 2018; Tkacz et al. 2018; Tourlousse et al. 2017; Zemb et al. 2020). Hardwick et al. 2018 describe an elegant approach to ensure ISDs emulate focal taxa during laboratory preparation through preserving sequence composition characteristics (e.g., GC content, etc.). These researchers suggest simply reversing the portion of the genome of the focal taxon under consideration (e.g., the portion of the rRNA operon commonly used for molecular barcoding).

As we have described, tradeoffs exist with any ISD such that a general statement regarding the superiority of any single approach would be misleading. However, we do suggest that actual microbial cells should be used as ISDs for studies involving samples that are likely to vary in nucleic extraction yield. On the other hand, if a study uses samples that are not likely to vary systematically in extraction performance (e.g., leaves from the same plant taxon; aliquots of similar soils) then a synthetic ISD, such as those described by Tourlousse et al. 2017, should suffice and could be simpler to employ than a cellular spike-in ISD.

Regardless of whether a study design dictates the use of either cellular or synthetic spike-in ISDs, researchers should consider the benefits of using a mixture of multiple ISDs as opposed to a single sequence or taxon. By adding a known amount of multiple ISDs to each sample, the failure of an ISD to act as a true standard can be detected (Ji et al. 2020). For instance, if three ISDs were added to each sample in equal abundance and the relative abundance of the ISDs in the sequences were 1:2:1, then it is clear that the second ISD was over-represented and should be omitted from consideration for that sample. Identification of a single malfunctioning standard is possible when using three (or more) standards, whereas if only two standards were used it would not be possible to determine which of the two ISDs had failed.

Another benefit of a mixture of ISDs is that it may lead to increased ability to estimate technical variation. For instance, Tourlousse et al. 2017 created 12 synthetic ISDs and reported that each responded slightly differently to laboratory practices. Accordingly, they reported an improvement in the accuracy of absolute abundance calculations when summing read counts across ISDs. The same result was reported by Stämmler et al. 2016, who used several cellular ISDs.

A final benefit of an ISD mixture is that sequences (or cells) emulating a variety of taxa can be included; thus, providing insight into the effects of laboratory practices across taxa akin to using a mock community as a positive control (Goodrich et al. 2014; Nguyen et al. 2014). Clearly, as ISD mixtures become more complex, they demand more sequencing depth—saying nothing of the time spent on their design. Until a sufficient breadth of ISD mixtures becomes commercially available, we suggest that researchers strike a balance between commutability and logistical cost by choosing a handful of sequences (or cells) that emulate those of focal taxa.

Prior to designing an ISD suitable for a particular study design, it is worth considering to what extent an ISD is needed at all. For instance, if the sample can be homogenized to allow counting of target cells within an aliquot then an ISD will provide little additional benefit—though it could still act as a positive control and provide insight into technical variation. Counting cells may be possible for studies with few samples and can be accomplished through fluorescence microscopy (Amann and Fuchs 2008; Daims et al. 2001) or flow cytometry (Props et al. 2017a,b). For example, Vandeputte et al. (2017) used flow cytometry to count cells within a series of faecal samples and used these counts to transform 16s data from relative to actual abundances (also see Frossard et al. 2016). Such approaches hold great merit because many of the concerns with ISD efficacy that we describe below would be obviated by having a cell count in hand. Unfortunately, optimizing flow cytometry protocols for focal substrates may be impractical for many researchers, particularly those studying microbial assemblages living inside tissues of a host organism (Doležel et al. 2007). Moreover, flow cytometry requires specialized equipment and skill, and can increase the logistical burden of a study more than the use of a spike-in ISD.

Quantitative PCR can also be used to estimate total copies of a genomic feature in a sample (e.g., copies of 16s), which can then be used to convert relative abundance estimates for each taxon to absolute abundances (Bonk et al. 2018; Dannemiller et al. 2014; Higuchi et al. 1993; Jian et al. 2020; Lou et al. 2018; Zhang et al. 2017). Droplet digital PCR (ddPCR; Hindson et al. 2011), is a promising tool for this approach because it provides heightened accuracy and throughput compared to conventional real-time qPCR; most importantly, it estimates abundances directly and does not rely on comparison to a quantitative standard (Baker 2012; Hindson et al. 2011; Kim et al. 2015; Morella et al. 2018). At the time of writing, ddPCR is currently more expensive than qPCR and also operates over a smaller dynamic range. The use of qPCR, via ddPCR or traditional techniques, is a simple, elegant approach to estimate absolute microbial abundances, however many of the pitfalls affecting ISDs can also affect this technique (e.g., primer bias, PCR inhibitors; Bonk et al. 2018). Moreover, while qPCR is relatively inexpensive, costs can mount when analyzing many thousands of samples and, therefore, the use of an ISD may save time and money for large-scale sequencing studies. The benefits and drawbacks of qPCR versus ISDs are poorly characterized, however, Stämmler et al. 2016 suggested that cellular ISDs outperformed qPCR for conversion of relative abundances to absolute abundances. These authors were studying the faecal microbiome and it is unclear if their findings translate to other substrates.

Considerations when deploying an internal standard

The primary reason ISDs can fail to act as a standard is when the ratio of focal cells (or sequences) to the ISD shifts among samples in unexpected and unmeasured ways (Fig. 1, 2). A simple way this can happen is if there is unmeasured and unaccounted for variation among samples in input mass. To see why this is problematic, consider the situation in which two samples have identical microbial assemblages, but one sample has half the input mass of the other sample and therefore contains half as much DNA (Fig. 1c). If the same amount of ISD were added to each sample and normalization calculations performed as described above, then it would appear as if microbial abundance was twice as high for one of the samples. While the two samples truly differ in microbial abundance, the difference is driven by differences in input mass among samples, not by differences in the microbial abundance in the source material. Consequently, laboratory methods typically involve standardization of the input mass of samples. However, imprecision in mass measurements made prior to nucleic acid extraction is rarely accounted for during data analysis and can add misleading variation to absolute abundance estimates. Problematic confounding could arise if sample mass were to differ systematically by substrate, experimental treatment, or among other batches. Fortunately, if input mass or volume varied among samples but was recorded, researchers can transform absolute abundances to absolute densities, on a scale of units of the ISD per unit of input mass (or volume).

A more insidious problem is when samples possess similar total masses but differ in the amount of target substrate present. For instance, if samples differ in hydration, then variation in the amount of water present could obscure differences in extractable mass among samples. Therefore, samples should be well dried prior to weighing and ISD incorporation. Variation in the amount of inorganic substrate present is particularly challenging for soil samples, which often differ in mineral composition, and hence density. In such cases, researchers should consider if volume is a more appropriate unit by which to standardize samples. The problem becomes amplified by comparisons across different substrates with fundamentally different characteristics and varying mixtures of potential microbial 'habitats' (e.g., comparisons across water, soil, and plants, or even different soils containing assemblages derived from communities within pore water, organic, and inorganic matter pools). Two soils could have identical soil water masses and microbial communities, but varying soil matrices and associated microbial masses that could alter the final homogenized samples if normalized by total volume or mass. Time represented by the sample may also be important (e.g., duration of water filtration or soil accumulation and dormant microbial burial).

ISD efficacy can also be undercut by variation in nucleic extraction performance among samples (Fig. 1c). For instance, if samples differ in physical toughness, such as what could be expected among tissue types of plants (i.e., stems versus leaves), more DNA will be obtained from samples with cells that are easier to lyse and the ratio of ISD to template DNA obtained will shift among samples, leading to inaccurate absolute abundance calculations. The same problem could occur if samples differ in the presence of compounds that inhibit extraction effectiveness (e.g., phenols in plants; Wilson 1997).

Variation in extraction yield is particularly difficult to measure for researchers interested in endosymbiotic microbial assemblages. This is because the recalcitrance of samples is defined by the traits of the host cells within and among which focal microbes reside (e.g., cell wall thickness can vary among plant taxa and tissue type) and a microbial cellular ISD will not emulate these traits. A possible solution for this problem is suggested through recent work by Karasov et al. (2019) who show that host-derived DNA can function as an inherent ISD when examining microbial symbiont assemblages. These researchers suggest estimation of microbial load as the ratio of host to bacterial reads obtained from shotgun metagenomic sequencing (also see Karasov et al. 2018, 2019; Regalado et al. 2019). A possible benefit of this approach, as stated by the authors, is that metagenomic sequencing is a less biased way to estimate total host and bacterial load than amplicon sequencing.

Unfortunately, nucleic acid extraction methodology is not the only laboratory technique that can influence the effectiveness of an ISD. Compounds that can inhibit or facilitate PCR (Rossen et al. 1992; Wilson and Carroll 1997) may also cause problems by imposing biases upon amplicon mixtures. Consider the case when variation in amplification has occurred across samples that differ only in the presence of inhibiting or facilitating compounds (reviewed by Schrader et al. 2012). Such a scenario would give the erroneous impression that shifts in actual abundance had taken place. Commonly encountered inhibitors include humic and fulvic acids in soil (Opel et al. 2010; Yeates et al. 1998) and phenols and polysaccharides in plants (Schrader et al. 2012; Wilson 1997). It is reasonable to assume that inhibitory compounds commonly vary in their concentrations among environmental samples (e.g., among soil types or plant taxa). Quantifying and accounting for variation in these compounds is onerous, thus the use of nucleic acid extraction protocols that consistently remove problematic compounds at the outset will minimize this source of variation—a stated benefit of many commercially available extraction kits (e.g., the Qiagen PowerSoil kit removes humic acid; Mahmoudi et al. 2011).

Given the many ways an ISD can fail as standards, we suggest researchers incorporate several control measures into sequencing studies to ensure ISDs perform as expected. At the minimum, ISDs should be added to technical replicates of samples representative of the biological variation present. Upon sequencing, the ISD should capture approximately the same proportion of reads in each of these replicates. Secondly, as mentioned above, we advocate for using a mixture composed of at least three ISDs. Finally, when using a new ISD, or using an ISD in a new substrate, it is ideal to test for quantitative behavior through sequencing a dilution series; reads should increase proportionally to ISD concentration.

At what laboratory step should an ISD be added?

One critical consideration when using a spike-in ISD is determining an appropriate time to add the ISD to samples. Most authors advocate adding the ISD before nucleic acid extraction (Jones et al. 2015; Smets et al. 2016; Tourlousse et al. 2017; Venkataraman et al. 2018; Zemb et al. 2020). This allows an ISD to capture variation in extraction performance (as mentioned above; Fig. 1d). If samples come from the same substrate and are thus not expected to behave differently during nucleic acid extraction, then an ISD could be added after extraction but prior to normalizing DNA concentrations for PCR (Fig. 1a). If the ISD is added after equimolar normalization of input DNA, then the IDS functions as a constant, positive control for PCR and sequencing (Fig. 1b) of each sample, but does not provide a standard for calculating absolute abundances in the original samples (prior to normalization).

Given that the efficacy of nucleic acid extraction is likely to vary among samples and sampling groups for many study designs, we suggest that incorporating an ISD into samples prior to extraction as the ideal. We note that measuring variation in extraction performance requires a cellular ISD (see above), however adding a nucleic acid ISD into samples prior to DNA extraction can be beneficial (Zemb et al. 2020). The benefit arises because the abundance of the ISD in the sample would track the expected and potentially variable loss of some DNA in extraction, such as would be caused by incomplete processing of all sample mass, variance during movement of supernatant and sample mass through the extraction protocol, or variable elution of nucleic acids from the solid-phase of columns used to isolate those acids.

Additional considerations when basing inference on microbial abundances

Comparison of absolute abundances among taxa is potentially misleading

ISDs can account for among-sample variation when comparing the effects of treatment or ecological covariates on abundances (both relative and absolute) of a particular microbial taxon. They cannot however address all the concerns that complicate the comparison of abundances of *different* taxa among and within samples. This is because every step of the library preparation process has the potential to impose idiosyncratic, selective biases for and against the DNA sequences associated with different taxa in a sample (Fig. 2). For example, PCR primers do not match their target sequences equally well in all taxa, leading to preferential amplification of some taxa, and substantial differences in selectivity among different primers (Fouhy et al. 2016; Hong et al. 2009). Thus, if a primer pair is biased against a particular sequence, then the abundance within the sample will be underestimated and an ISD cannot remedy this error. Aside from primer pair, the type of polymerase, PCR cycle count, PCR reagents used (Nilsson et al. 2018; Pollock et al. 2018; Schori et al. 2013), GC content (Laursen et al. 2017; Risso et al. 2011), length of the amplicon (Oshlack and Wakefield 2009), and even sequencing platform (D'Amore et al. 2016), can all impose further biases that influence resulting sequence data. Thus, these procedural biases can cause false negatives in inferences about external determinants of assemblage composition and simply make it difficult to know true abundances.

Estimates of abundances of taxa are further complicated by error that arises due to high copy number variation (CNV) among taxa in marker loci. For example, Lofgren et al. (2019) reported that fungal taxa can differ in ITS copy number by an order of magnitude or more. Even within a single fungal taxon, *Suillus brevipes*, ITS copy number ranged from 72–156. While not quite as extreme as for fungi, CNV is also widespread among bacteria for the commonly used 16s marker (Kembel et al. 2012; Lee et al. 2009; Perisin et al. 2016;

Stoddard et al. 2015; Větrovský and Baldrian 2013). Of course, variation in ploidy-level (Pecoraro et al. 2011), or the number of nuclei in a cell (which can vary for fungi; Gladieux et al. 2014), can also influence copy number variation. A possible mitigation solution for bacteria and archaea is bioinformatic correction of CNV of focal taxa via comparison to the popular rrnDB database (Stoddard et al. 2015).

When taken together, these biases suggest extreme caution is in order when interpreting sequence data with the intention of inter-taxa comparisons of abundance (Fig. 2), such as when analyses focus on description of overall shifts in community composition as defined by changes in rank order abundances among taxa. Unfortunately, many microbial ecology studies rely on a common suite of such analyses, including description of patterns in diversity entropies, ordination techniques, and PERMANOVA. If taxon-specific analyses are used instead, or in conjunction with these techniques, many of the biases we describe here become much less problematic. This is because most biases will affect a taxon in the same way across samples and, therefore, biases will not be confounded with experimental treatment(s) or ecological covariates of interest. Moreover, many ecological questions are better answered by quantifying the effect of treatment on specific taxa, rather than documenting shifts in overall assemblage composition.

To learn about the biological causes of differences in taxon abundances among samples, it is helpful to partition variation that arises from replicated laboratory processes and biological variation among samples. As is the case for many experimental designs, statistical models for community composition can explicitly attribute variation to experimental and biological sources. In particular, hierarchical models for variation parameterize the mean frequency of taxa and variation among replicates, and mean frequency of taxa for each treatment (or sampling group) and variation among treatments. For instance, a hierarchical model for relative abundances of taxa in replicates and treatments can be specified with the multinomial and Dirichlet distributions (Coblentz et al. 2017; Fordyce et al. 2011; Harrison et al. 2020), with the additional benefit of providing robust estimates of familiar community ecology statistics (*sensu* Harrison et al. 2020; Marion et al. 2018). One or more ISDs can be used to partition technical from biological variation. Assuming ISDs behave as do focal taxa (i.e., they are commutable), technical variation among replicates can be estimated for the ISDs and subtracted from estimates of variation for individual taxa to yield an estimate of biological variation. Bayesian hierarchical models make this partitioning of variation possible, in part because they fully use and formally describe the counts of DNA sequences (and differences in information among samples). This is in contrast to rarefaction methods, which discard observed data and information about technical and biological variation among samples (McMurdie and Holmes 2014).

Conclusion

Sequencing is a powerful tool to measure abundance of organisms that are difficult to observe and count directly. We are growing increasingly aware of the challenges of using sequence data to measure abundances and the benefits provided by internal standards, but, as we have shown, their efficacy is dependent upon careful accounting during laboratory practices and potentially unrealistic assumptions of biological simplicity (e.g., in CNV). Nevertheless, ISDs liberate researchers from the constraints imposed by relative abundance data and we suggest that their use become a standard component of sequence-based microbial ecology studies (Jones et al. 2015; Stämmler et al. 2016; Tourlousse et al. 2017).

References

Aitchison, J. (1982). *The statistical analysis of compositional data.* New York, NY: Chapman and Hall.

Amann, R. and B. M. Fuchs (2008). “Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques”. *Nature Reviews Microbiology* 6.5, pp. 339–348.

Baker, M. (2012). “Digital PCR hits its stride”. *Nature Methods* 9.6, pp. 541–544.

Benninghoff, W. S. (1962). “Calculation of pollen and spore density in sediments by addition of exotic pollen in known quantities”. *Pollen et Spores* 4, pp. 332–333.

Bibby, K., E. Viau, and J. Peccia (2010). “Pyrosequencing of the 16S rRNA gene to reveal bacterial pathogen diversity in biosolids”. *Water Research* 44.14, pp. 4252–4260.

Bonk, F. et al. (2018). “PCR-based quantification of taxa-specific abundances in microbial communities: Quantifying and avoiding common pitfalls”. *Journal of Microbiological Methods* 153, pp. 139–147.

Caporaso, J. G. et al. (2012). “Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms”. *The ISME Journal* 6.8, pp. 1621–1624.

Carini, P. (2019). “A “cultural” renaissance: genomics breathes new life into an old craft”. *mSystems* 4.3, e00092–19.

Case, R. J. et al. (2007). “Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies”. *Applied and Environmental Microbiology* 73.1, pp. 278–288.

Chen, J. and H. Li (2013). “Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis”. *The Annals of Applied Statistics* 7.1, pp. 418–442.

Chen, K. et al. (2016). “The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses”. *Molecular and Cellular Biology* 36.5, pp. 662–667.

Clarridge, J. E. (2004). “Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases”. *Clinical Microbiology Reviews* 17.4, pp. 840–862.

Coblentz, K. E., A. E. Rosenblatt, and M. Novak (2017). “The application of Bayesian hierarchical models to quantify individual diet specialization”. *Ecology* 98.6, pp. 1535–1547.

D’Amore, R. et al. (2016). “A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling”. *BMC Genomics* 17.

Daims, H. et al. (2001). “Cultivation-independent, semiautomatic determination of absolute bacterial cell numbers in environmental samples by fluorescence in situ hybridization”. *Applied and Environmental Microbiology* 67.12, pp. 5810–5818.

Dannemiller, K. C. et al. (2014). “Combining real-time PCR and next-generation DNA sequencing to provide quantitative comparisons of fungal aerosol populations”. *Atmospheric Environment* 84, pp. 113–121.

Davis, M. B. (1963). “On the theory of pollen analysis”. *American Journal of Science* 261.10, pp. 897–912.

— (1966). “Determination of absolute pollen frequency”. *Ecology* 47.2, pp. 310–311.

Davis, M. B. and E. S. Deevey (1964). “Pollen accumulation rates: estimates from late-glacial sediment of Rogers Lake”. *Science* 145.3638, pp. 1293–1295.

Doležel, J., J. Greilhuber, and J. Suda (2007). “Estimation of nuclear DNA content in plants using flow cytometry”. *Nature Protocols* 2.9, pp. 2233–2244.

Egozcue, J. J. et al. (2003). “Isometric logratio transformations for compositional data analysis”. *Mathematical Geology* 35.3, pp. 279–300.

Eisenberg, E. and E. Y. Levanon (2013). “Human housekeeping genes, revisited”. *Trends in Genetics*. Human Genetics 29.10, pp. 569–574.

Fagerlind, F. (1952). “The real signification of pollen diagrams”. *Botaniska Notiser* 105, pp. 185–224.

Fernandes, A. D. et al. (2014). “Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis”. *Microbiome* 2, p. 15.

Fordyce, J. A. et al. (2011). “A hierarchical Bayesian approach to ecological count data: a flexible tool for ecologists”. *PLOS ONE* 6.11, e26785.

Fouhy, F. et al. (2016). “16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform”. *BMC Microbiology* 16.1, p. 123.

Frossard, A., F. Hammes, and M. O. Gessner (2016). “Flow cytometric assessment of bacterial abundance in soils, sediments and sludge”. *Frontiers in Microbiology* 7.

Giesecke, T. and S. L. Fontana (2008). “Revisiting pollen accumulation rates from Swedish lake sediments”. *The Holocene* 18.2, pp. 293–305.

Gladieux, P. et al. (2014). “Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes”. *Molecular Ecology* 23.4, pp. 753–773.

Gloor, G. B. and G. Reid (2016). “Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data”. *Canadian Journal of Microbiology* 62.8, pp. 692–703.

Gloor, G. B. et al. (2017). “Microbiome datasets are compositional: and this is not optional”. *Frontiers in Microbiology* 8.

Goodrich, J. K. et al. (2014). “Conducting a microbiome study”. *Cell* 158.2, pp. 250–262.

Hardwick, S. A., I. W. Deveson, and T. R. Mercer (2017). “Reference standards for next-generation sequencing”. *Nature Reviews Genetics* 18.8, pp. 473–484.

Hardwick, S. A. et al. (2016). “Spliced synthetic genes as internal controls in RNA sequencing experiments”. *Nature Methods* 13.9, pp. 792–798.

Hardwick, S. A. et al. (2018). “Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis”. *Nature Communications* 9.1, pp. 1–10.

Harrison, J. G. et al. (2020). “Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data”. *Molecular Ecology Resources* 20.2, pp. 481–497.

Higuchi, R. et al. (1993). “Kinetic PCR Analysis: Real-time Monitoring of DNA Amplification Reactions”. *Bio/Technology* 11.9, pp. 1026–1030.

Hindson, B. J. et al. (2011). “High-throughput droplet digital PCR system for absolute quantitation of DNA copy number”. *Analytical Chemistry* 83.22, pp. 8604–8610.

Hong, S. et al. (2009). “Polymerase chain reaction primers miss half of rRNA microbial diversity”. *The ISME Journal* 3.12, pp. 1365–1373.

Hossain, A. et al. (2020). *A massively parallel COVID-19 diagnostic assay for simultaneous testing of 19200 patient samples*. Google Docs. URL: https://docs.google.com/document/d/1kP2w_uTMSep2UxTCOnUhh1TMCjWvHEYOsUUpkJHPYV4/preview?sle=true&usp=embed_facebook (visited on 2020).

Jackson, D. A. (1997). "Compositional data in community ecology: the paradigm or peril of proportions?" *Ecology* 78.3, pp. 929–940.

Ji, Y. et al. (2020). "SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes". *Molecular Ecology Resources* 20.1, pp. 256–267.

Jian, C. et al. (2020). "Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling". *PLOS ONE* 15.1, e0227285.

Jiang, L. et al. (2011). "Synthetic spike-in standards for RNA-seq experiments". *Genome Research*.

Jones, M. B. et al. (2015). "Library preparation methodology can influence genomic and functional predictions in human microbiome research". *Proceedings of the National Academy of Sciences* 112.45, pp. 14024–14029.

Jonge, H. J. M. de et al. (2007). "Evidence based selection of housekeeping genes". *PLoS ONE* 2.9.

Karasov, T. L. et al. (2018). "*Arabidopsis thaliana* and *Pseudomonas* pathogens exhibit stable associations over evolutionary timescales". *Cell Host & Microbe* 24.1, 168–179.e4.

Karasov, T. L. et al. (2019). "The relationship between microbial biomass and disease in the *Arabidopsis thaliana* phyllosphere". *bioRxiv*, p. 828814.

Kembel, S. W. et al. (2012). "Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance". *PLOS Computational Biology* 8.10, e1002743.

Kim, T. G., S.-Y. Jeong, and K.-S. Cho (2015). "Development of droplet digital PCR assays for methanogenic taxa and examination of methanogen communities in full-scale anaerobic digesters". *Applied Microbiology and Biotechnology* 99.1, pp. 445–458.

Laursen, M. F., M. D. Dalgaard, and M. I. Bahl (2017). "Genomic GC-Content Affects the Accuracy of 16S rRNA Gene Sequencing Based Microbial Profiling due to PCR Bias". *Frontiers in Microbiology* 8.

Lee, Z. M.-P., C. Bussema, and T. M. Schmidt (2009). "rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea". *Nucleic Acids Research* 37 (suppl_1), pp. D489–D493.

Lin, Y. et al. (2019). "Towards quantitative microbiome community profiling using internal standards". *Applied and Environmental Microbiology* 85.5.

Lofgren, L. A. et al. (2019). "Genome-based estimates of fungal rDNA copy number variation across phylogenetic scales and ecological lifestyles". *Molecular Ecology* 28.4, pp. 721–730.

Lou, J. et al. (2018). "Assessing soil bacterial community and dynamics by integrated high-throughput absolute abundance quantification". *PeerJ* 6, e4514.

Lun, A. T. L. et al. (2017). "Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data". *Genome Research* 27.11, pp. 1795–1806.

Lynch, M. (2007). *The origins of genome architecture.* Vol. 98. Sunderland, MA, USA: Sinauer Associates.

Mahmoudi, N., G. F. Slater, and R. R. Fulthorpe (2011). “Comparison of commercial DNA extraction kits for isolation and purification of bacterial and eukaryotic DNA from PAH-contaminated soils”. *Canadian Journal of Microbiology* 57.8, pp. 623–628.

Marion, Z. H., J. A. Fordyce, and B. M. Fitzpatrick (2018). “A hierarchical Bayesian model to incorporate uncertainty into methods for diversity partitioning”. *Ecology* 99.4, pp. 947–956.

McMurdie, P. J. and S. Holmes (2014). “Waste not, want not: why rarefying microbiome data Is inadmissible”. *PLOS Comput Biol* 10.4, e1003531.

Morella, N. M. et al. (2018). “Rapid quantification of bacteriophages and their bacterial hosts in vitro and in vivo using droplet digital PCR”. *Journal of Virological Methods* 259, pp. 18–24.

Nguyen, N. H. et al. (2014). “Parsing ecological signal from noise in next generation amplicon sequencing”. *New Phytologist* 205.4, pp. 1389–1393.

Nilsson, R. H. et al. (2018). “Mycobiome diversity: high-throughput sequencing and identification of fungi”. *Nature Reviews Microbiology*, p. 1.

Opel, K. L., D. Chung, and B. R. McCord (2010). “A study of PCR inhibition mechanisms using real time PCR”. *Journal of Forensic Sciences* 55.1, pp. 25–33.

Oshlack, A. and M. J. Wakefield (2009). “Transcript length bias in RNA-seq data confounds systems biology”. *Biology Direct* 4.1, p. 14.

Palmer, J. M. et al. (2018). “Non-biological synthetic spike-in controls and the AMPtk software pipeline improve mycobiome data”. *PeerJ* 6, e4925.

Pearson, K. (1897). “Mathematical contributions to the theory of evolution—on a form of spurious correlation which may arise when indices are used in the measurement of organs”. *Proceedings of the Royal Society of London* 60.359, pp. 489–498.

Pecoraro, V. et al. (2011). “Quantification of ploidy in Proteobacteria revealed the existence of monoploid, (mero-)oligoploid and polyploid species”. *PLoS ONE* 6.1.

Perisin, M. et al. (2016). “16Stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies”. *The ISME Journal* 10.4, pp. 1020–1024.

Piwosz, K. et al. (2018). “Determining lineage-specific bacterial growth curves with a novel approach based on amplicon reads normalization using internal standard (ARNIS)”. *The ISME Journal* 12.11, pp. 2640–2654.

Pollock, J. et al. (2018). “The madness of microbiome: attempting to find consensus “best practice” for 16S microbiome studies”. *Applied and Environmental Microbiology* 84.7, e02627–17.

Prentice, I. C. and T. Webb (1986). “Pollen percentages, tree abundances and the Fagerlind effect”. *Journal of Quaternary Science* 1.1, pp. 35–43.

Props, R. et al. (2017a). “Absolute quantification of microbial taxon abundances”. *The ISME journal* 11.2, pp. 584–587.

Props, R. et al. (2017b). “Measuring the biodiversity of microbial communities by flow cytometry”. *Methods in Ecology and Evolution*, pp. 1376–1385.

Qing, T. et al. (2013). “mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies”. *Science China Life Sciences* 56.2, pp. 134–142.

Regalado, J. et al. (2019). “Combining whole genome shotgun sequencing and rDNA amplicon analyses to improve detection of microbe-microbe interaction networks in plant leaves”. *bioRxiv*, p. 823492.

Risso, D. et al. (2011). “GC-content normalization for RNA-Seq data”. *BMC Bioinformatics* 12.1, p. 480.

Risso, D. et al. (2014). “Normalization of RNA-seq data using factor analysis of control genes or samples”. *Nature Biotechnology* 32.9, pp. 896–902.

Rossen, L. et al. (1992). “Inhibition of PCR by components of food samples, microbial diagnostic assays and DNA-extraction solutions”. *International Journal of Food Microbiology* 17.1, pp. 37–45.

Schmidt, P.-A. et al. (2013). “Illumina metabarcoding of a soil fungal community”. *Soil Biology and Biochemistry* 65, pp. 128–132.

Schoch, C. L. et al. (2012). “Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi”. *Proceedings of the National Academy of Sciences* 109.16, pp. 6241–6246.

Schori, M. et al. (2013). “Engineered DNA polymerase improves PCR results for plastid DNA”. *Applications in Plant Sciences* 1.2, p. 1200519.

Schrader, C. et al. (2012). “PCR inhibitors – occurrence, properties and removal”. *Journal of Applied Microbiology* 113.5, pp. 1014–1026.

Smets, W. et al. (2016). “A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing”. *Soil Biology and Biochemistry* 96, pp. 145–151.

Stämmler, F. et al. (2016). “Adjusting microbiome profiles for differences in microbial load by spike-in bacteria”. *Microbiome* 4.1, p. 28.

Stewart, F. J. and C. M. Cavanaugh (2007). “Intragenomic variation and evolution of the internal transcribed spacer of the rRNA operon in bacteria”. *Journal of Molecular Evolution* 65.1, pp. 44–67.

Stoddard, S. F. et al. (2015). “rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development”. *Nucleic Acids Research* 43 (D1), pp. D593–D598.

Taberlet, P. et al. (2012). “Towards next-generation biodiversity assessment using DNA metabarcoding”. *Molecular Ecology* 21.8, pp. 2045–2050.

Thellin, O. et al. (1999). “Housekeeping genes as internal standards: use and limits”. *Journal of Biotechnology* 75.2, pp. 291–295.

Tkacz, A., M. Hortala, and P. S. Poole (2018). “Absolute quantitation of microbiota abundance in environmental samples”. *Microbiome* 6.1, p. 110.

Tourlousse, D. M. et al. (2017). “Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing”. *Nucleic Acids Research* 45.4, e23–e23.

Tricarico, C. et al. (2002). “Quantitative real-time reverse transcription polymerase chain reaction: normalization to rRNA or single housekeeping genes is inappropriate for human tissue biopsies”. *Analytical Biochemistry* 309.2, pp. 293–300.

Tsilimigras, M. C. B. and A. A. Fodor (2016). “Compositional data analysis of the microbiome: fundamentals, tools, and challenges”. *Annals of Epidemiology.* The Microbiome and Epidemiology 26.5, pp. 330–335.

Vandeputte, D. et al. (2017). “Quantitative microbiome profiling links gut community variation to microbial load”. *Nature* 551.7681, pp. 507–511.

Venkataraman, A. et al. (2018). “Spike-in genomic DNA for validating performance of metagenomics workflows”. *BioTechniques* 65.6, pp. 315–321.

Větrovský, T. and P. Baldrian (2013). “The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses”. *PLOS ONE* 8.2, e57923.

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

Wilson, D. and G. C. Carroll (1997). “Avoidance of high-endophyte space by gall-forming insects”. *Ecology* 78.7, pp. 2153–2163.

Wilson, I. G. (1997). “Inhibition and facilitation of nucleic acid amplification.” *Applied and Environmental Microbiology* 63.10, pp. 3741–3751.

Yang, L. et al. (2018). “Use of an improved high-throughput absolute abundance quantification method to characterize soil bacterial community and dynamics”. *Science of The Total Environment* 633, pp. 360–371.

Yeates, C. et al. (1998). “Methods for microbial DNA extraction from soil for PCR amplification”. *Biological Procedures Online* 1.1, pp. 40–47.

Zemb, O. et al. (2020). “Absolute quantitation of microbes using 16S rRNA gene metabarcoding: A rapid normalization of relative abundances by quantitative PCR targeting a 16S rRNA gene spike-in standard”. *MicrobiologyOpen*, e977.

Zhang, Z. et al. (2017). “Soil bacterial quantification approaches coupling with relative abundances reflecting the changes of taxa”. *Scientific Reports* 7.1, pp. 1–11.

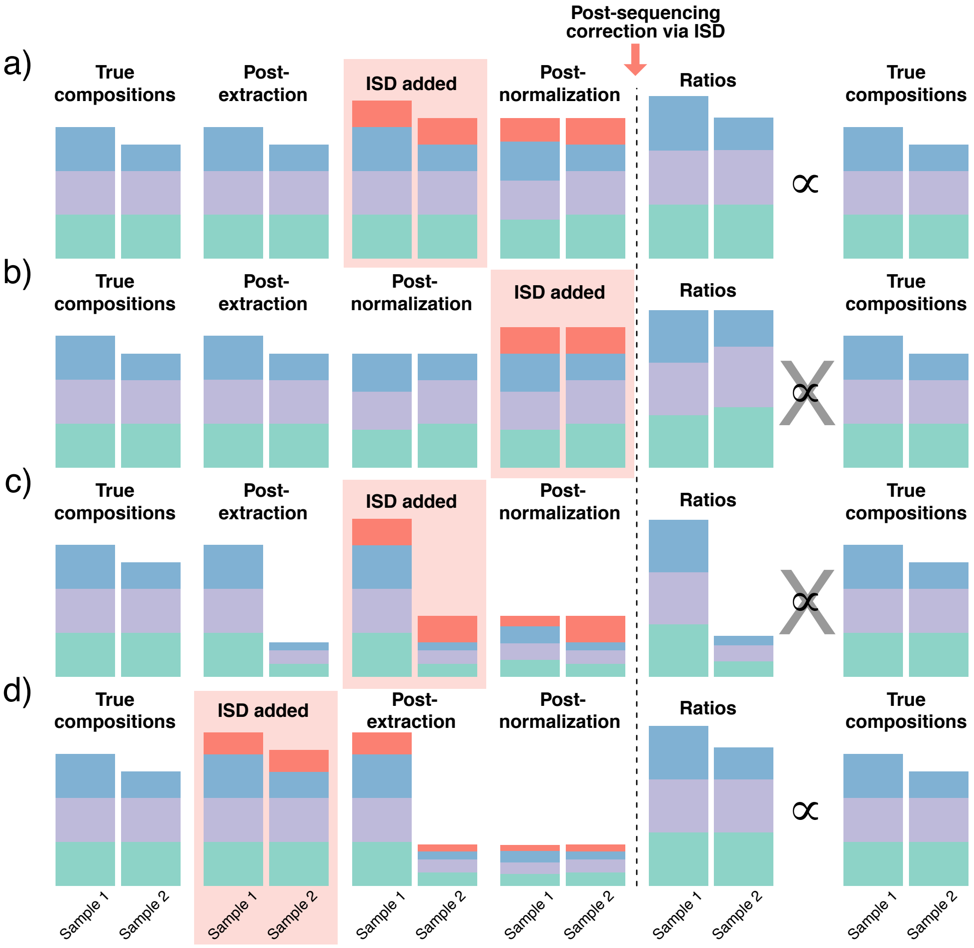

Figure 1: The addition of an internal standard (ISD) to samples can correct for the problems posed by the compositional nature of sequencing data. This is because the ISD can ensure the relative abundances of reads obtained from sequencing are proportional to those in the original composition, thus allowing calculation of absolute abundances for each sequenced feature. Here, we present data representative of four laboratory scenarios that affect ISD efficacy. For each scenario, we present relative abundance data for two samples, each of which contains three features that are shown in different colors. The ISD is shown in orange and, for each scenario, a light orange box denotes the step at which the ISD is added. a) Here, ISD is added prior to equimolar pooling of nucleic acids for PCR and there is no variation in sample mass, or yield from nucleic acid extraction, or other biases induced by laboratory-practice. In this case the ISD performs as desired. b) If, however, the ISD is added after equimolar pooling of samples then it is no longer effective. c) Similarly, if samples differ in yield from nucleic acid extraction per unit of mass and the ISD does not reflect those differences, then the ISD is no longer effective. d) If the ISD is added prior to nucleic acid extraction and reflects variation in extraction yield among samples (i.e., as would be expected for a cellular ISD), then the ISD can be used to back-calculate absolute abundances.

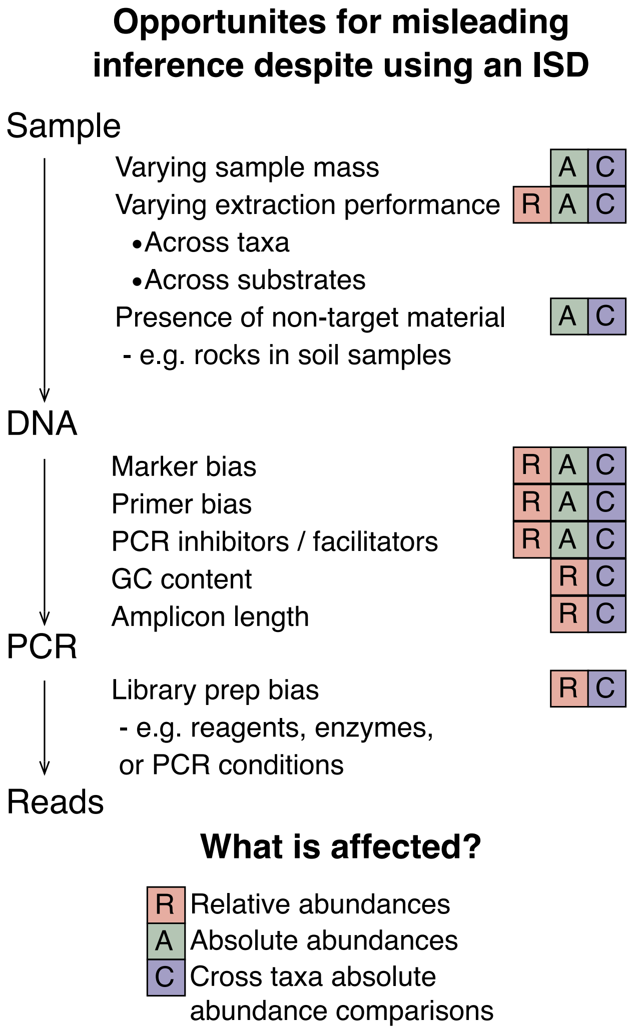

Figure 2: Biases can be introduced throughout the process of obtaining DNA sequence data from samples and will interfere with estimating abundances, despite the use of an internal standard (ISD). These biases are organized chronologically following the data generation process—from sampling to sequencing. Colored boxes next to each source of bias denote whether it can affect relative abundances or absolute abundances. All sources of bias interfere with comparisons across taxa. This catalogue of biases does not mean amplicon-based sequencing with internal standards is doomed to fail, only that biases must carefully considered when planning an experiment so that the most meaning can be extracted from the resulting data.