AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

METHODS AND PROTOCOLS

Characterizing Microbiomes via Sequencing of Marker Loci: Techniques To Improve Throughput, Account for Cross-Contamination, and Reduce Cost

Joshua G. Harrison,[a] Gregory D. Randolph,[a] C. Alex Buerkle[a]

[a]University of Wyoming, Laramie, Wyoming, USA

ABSTRACT New approaches to characterizing microbiomes via high-throughput sequencing provide impressive gains in efficiency and cost reduction compared to approaches that were standard just a few years ago. However, the speed of method development has been such that staying abreast of the latest technological advances is challenging. Moreover, shifting laboratory protocols to include new methods can be expensive and time consuming. To facilitate adoption of new techniques, we provide a guide and review of recent advances that are relevant for single-locus sequence-based study of microbiomes—from extraction to library preparation—including a primer regarding the use of liquid-handling automation in small-scale academic settings. Additionally, we describe several amendments to published techniques to improve throughput, track contamination, and reduce cost. Notably, we suggest adding synthetic DNA molecules to each sample during nucleic acid extraction, thus providing a method of documenting incidences of cross-contamination. We also describe a dual-indexing scheme for Illumina sequencers that allows multiplexing of many thousands of samples with minimal PhiX input. Collectively, the techniques that we describe demonstrate that laboratory technology need not impose strict limitations on the scale of molecular microbial ecology studies.

IMPORTANCE New methods to characterize microbiomes reduce technology-imposed limitations to study design, but many new approaches have not been widely adopted. Here, we present techniques to increase throughput and reduce contamination alongside a thorough review of current best practices.

KEYWORDS microbiome, high throughput, next-generation sequencing, spike in, internal standard, library preparation, PCR, automation, multiplexing, metabarcoding

Microbiomes have been at the forefront of biological discovery over the past few decades, largely because of ongoing improvements to nucleic acid sequencing technology. Indeed, new sequencing tools have facilitated the expansion of the microbial portion of the tree of life (1), led to widespread acknowledgment of the importance of microbial symbionts (2, 3), and spurred the development of industries to harness microbiomes (4, 5). However, for most laboratories, adopting the latest sequencing approaches is challenging because best practices are constantly evolving, and shifting to new techniques is time consuming. Thus, many biologists resort to established protocols that can be costly and low throughput and can limit the inferences made possible by sequence data.

For example, we used Google Scholar to search papers published since 2019 for the two terms "microbiome" and "16S" (the latter is a common barcoding locus for bacteria). To gauge current typical practices, we examined the first 50 papers returned from this query that used sequencing tools to characterize microbiomes. We also took a

Citation Harrison JG, Randolph GD, Buerkle CA. 2021. Characterizing microbiomes via sequencing of marker loci: techniques to improve throughput, account for cross-contamination, and reduce cost. mSystems 6: e00294-21. https://doi.org/10.1128/mSystems.00294-21.

Editor Peter J. Turnbaugh, University of California, San Francisco

Ad Hoc Peer Reviewer Dieter Tourlousse, National Institute of Advanced Industrial Science and Technology (AIST)

Address correspondence to Joshua G. Harrison, joshua.grant.harrison@gmail.com.

Review of current best practices to reduce costs associated with microbiome metabarcoding studies. New methods to reduce and account for contamination and improve throughput are also presented and discussed.

Received 10 March 2021
Accepted 7 June 2021
Published 13 July 2021

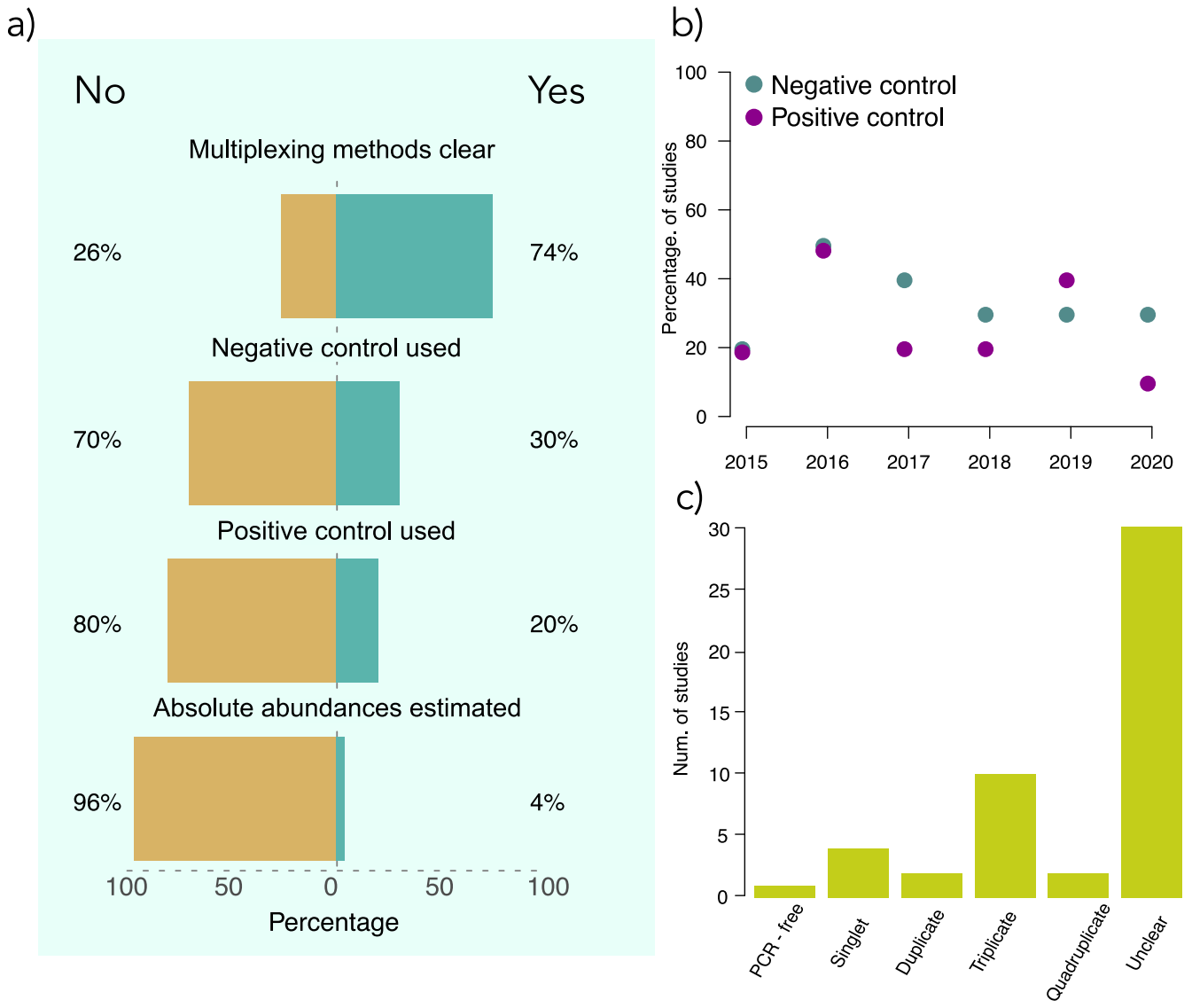

FIG 1 Snapshot of methodological practices employed from 50 microbiome publications from 2019 to 2020 (see main text). Our goal with this survey was to demonstrate the state of the field. We do not wish to disparage existing methodologies but rather point out that improvements to quality control and throughput are readily possible. (a) Proportion of publications that had clear multiplexing methods and used proper controls. No publication employed an approach that could account for subtle cross-contamination. Additionally, none of the publications reported the use of automation tools, and 41 of the studies used the Illumina MiSeq, which has been superseded by machines with vastly higher output capabilities. (b) Results from a cursory survey of control use in microbiome papers from 2015 to 2020 (see main text for details). (c) PCR replication in the 50 papers surveyed for panel a. More importantly, this panel illustrates that many publications had somewhat unclear methods sections.

cursory look back to 2015 (10 papers per year) to determine if the use of controls has changed over time.

We found that few studies adopted best practices for quality control (Fig. 1). For instance, we found that only 15 of 50 studies used a negative control to account for laboratory reagent contamination (6), and only 10 studies mentioned a positive control of some sort (also see reference 7). There was no obvious trend toward improved inclusion of proper controls with time (Fig. 1b). Fewer studies still, only four, included an internal standard or used quantitative PCR to place compositional relative abundance data from the sequencer on a standard scale to facilitate analysis (see below). Additionally, we found that most studies relied on expensive, but proven, techniques that support relatively limited throughput. For example, Illumina offers several new machines with extreme output (e.g., the NovaSeq), yet 42 of 50 studies used the older MiSeq instrument. Perhaps the most concerning trend we

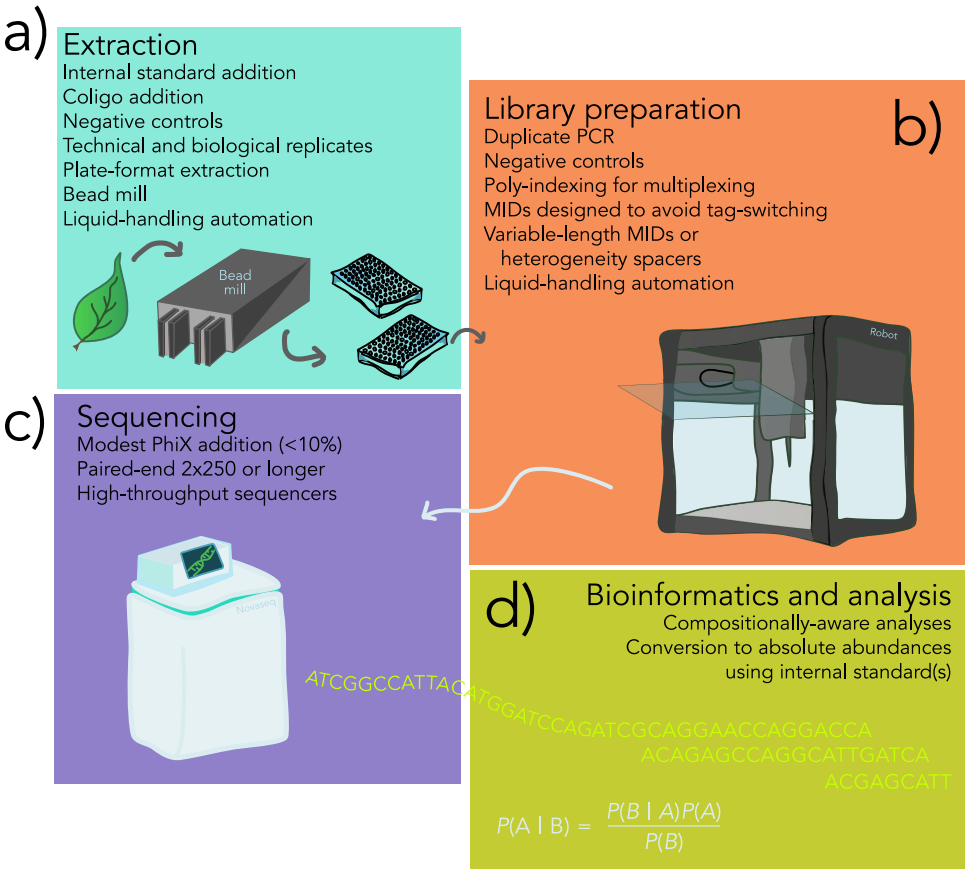

FIG 2 Overview of best practices to improve throughput and lower the costs associated with amplicon-based characterization of microbiomes. (a) Sample preparation (e.g., weighing tissue) and DNA extraction are time intensive because they are difficult to automate. Bead mills and liquid-handling systems can help improve extraction yield and reduce contamination. Extraction is the ideal time to add internal standards and cross-contamination-checking oligonucleotides (coligos). Due to ubiquitous contamination of extraction reagents with microbes, negative controls are essential. (b) Library preparation is very amenable to automation. New polyindexing strategies, such as the dual-indexing approach we describe here, allow multiplexing of many thousands of samples. Sequencer yield can be improved through ensuring adequate sequence variation at the start of reads. This can be accomplished through the use of variable-length molecular identifiers (MIDs) or heterogeneity spacers. (c) Modern sequencing instruments generate sufficient volume of data for extreme multiplexing, bringing the cost of sequencing down to several US dollars or less per sample. (d) Bioinformatics and analytical techniques are critical to the success of any sequencing project, and the challenge of properly analyzing and storing data should not be overlooked. Users must be aware of the limitations posed by compositional data and curtail inferences as required. Internal standards are of great benefit because they allow relative abundance data to be placed on a scale proportional to absolute abundances.

noticed was that most papers lacked clarity regarding laboratory procedures, thus hampering reproducibility.

While cursory, this survey mirrored our expectations regarding the limitations of typical microbiome laboratory practices that we have observed in reading the current literature. Our goal here is not to disparage the state of the field but rather to draw attention to the opportunity that exists to dramatically reduce costs and improve research outcomes through adoption of new tools and techniques.

Consequently, here, we critically examine every step of single-locus, sequence-based microbiome characterization (often referred to as metabarcoding) (Fig. 2). Our goal is to describe the advantages and disadvantages of new methods while paying specific attention to time and cost-saving techniques (such as automation). Alongside our critical review, we present several improvements to existing protocols for library preparation. When taken together, the techniques we discuss greatly reduce technologically imposed limitations on study design. Indeed, we have found that the primary logistical consideration when planning research is no longer the costs associated with

laboratory procedures but instead those associated with sample collection and handling.

Importantly, in this review, we do not discuss experimental design (including sample collection and storage [8, 9]), laboratory inventory management systems (LIMS) (10), or bioinformatic approaches (e.g., see references 11–14). We also do not compare sequencing instruments (including new long-read machines), though the multiplexing advances we describe require the use of the latest generation of short-read sequencers (e.g., the Illumina NovaSeq). While bioinformatics and statistical analysis is beyond the purview of this review, we wish to be explicit that the computational burden incurred by higher-throughput sequencing can be significant and should be considered during the early phases of study design. Notably, as new sequencing platforms are brought to market, existing bioinformatics methods are challenged and can fail; thus, researchers should expect to continually modify their bioinformatic pipeline.

RESULTS AND DISCUSSION

Best practices for the characterization of microbiomes—an overview. (i) Robots in the lab: a word regarding liquid-handling automation. Much of the laboratory work we discuss involves moving small amounts of liquid from one place to another via pipetting. This process is fraught because cross-contamination is a constant threat and variation in pipetting technique among practitioners can influence results. A variety of benchtop robots exist that can automate liquid-handling tasks, including models by Eppendorf, Integra, Opentrons, and others, that cost less than $20,000 USD new (with some simple models costing a fraction of this amount). These instruments consist of a programmable pipette on a movable gantry. Despite their simplicity, they can be extremely useful during nucleic acid extraction and library preparation. More complex robots have large multiposition "decks" that can hold a variety of consumables and additional tools, including, for example, heating or cooling blocks, shakers, centrifuges, or vacuum manifolds (used in place of a centrifuge to pull solutions through columns or filters). These added capabilities come with increased list price—many of these robots are more than $100,000 new.

Notably, the refurbished and used market is large for automation systems. For simple robots, a used machine could suffice, as troubleshooting can be quite straightforward. However, for more complex systems, the benefits of warrantied repair and technical assistance could justify a new purchase, because, in our experience, there will be considerable programming and other technical challenges to surpass. For the motivated and cost-conscious scientist, open-source plans for conversion of three-dimensional (3D) printers to perform liquid handling are available (15, 16) and represent an inexpensive way to experiment with automation (e.g., see reference 17).

When choosing a robot, it is imperative to consider the programming required to accomplish a task. Some machines rely on easy-to-use graphical user interfaces, while others employ proprietary programming languages that are time consuming to learn. Another consideration is error handling, as not all automation systems provide sensible approaches for detecting and reporting errors. Ideally, users will be notified of an error and asked how to proceed. If an instrument does not provide such functionality, its benefits will be undercut, because it will require chaperoning. In the worst case, the instrument will proceed with no documentation of the error and much time will be lost sorting out the mistake. Speaking generally, we have found that robots often require maintenance and troubleshooting, and this should be expected as a probable time cost before purchasing an automation system.

The consumables required by robots are another purchasing consideration. Many liquid-handling systems use proprietary pipette tips (and other plastics) that can add costs. We recommend choosing a robot that can handle both skirted and unskirted 96-well plates as well as 384-well plates. While most protocols used by academic labs rely on 96-well plates, we anticipate a shift to 384-well plates as more researchers seek increased sample throughput (18).

Ideally, robot functionality should allow for a variety of flourishes that can reduce contamination. For example, dispensation speed and height can be reduced to prevent splashing, and pipette tips can be touched to the sides of wells to wick off droplets prior to movement of the pipettor to some other location on the deck. Some robots boast drip detection technology that can warn users of possible contamination. Attention to such details is important, else automation will worsen the threat of cross-contamination.

(ii) Diagnosing contamination—a ubiquitous problem for microbiome sequencing. There are two primary types of contamination to consider when performing sequence-based surveys of microbiomes: contamination of samples by foreign microbes and cross-contamination among samples (19, 20). It has become clear that regardless of the care taken by practitioners during laboratory work, contamination is always a threat. This is because microbes are known to occur in many reagents and solvents (6) and are thus unavoidable.

While statistical removal of contaminants has been suggested (e.g., see reference 21) and can be a valuable tool, such approaches should not be substituted for incorporation of negative controls into the design of a study. Moreover, bioinformatic procedures may depend upon data from negative controls (*sensu* the decontam software; Davis et al. [21]). Unfortunately, proper use of negative controls is still surprisingly uncommon (see the introduction above), and the likely prevalence of cross-contamination is particularly concerning (22).

The negative controls required will be determined by study design; however, at minimum, aliquots of all reagents and solvents should be used as template for sequencing (including aliquots from each extraction kit used). Aliquots of reagents should be taken at the end of a laboratory process to maximize the chances of diagnosing contamination that occurred during work. Contamination of negative controls is often tested via PCR; however, we suggest that controls be sequenced, because PCR lacks the sensitivity to characterize instances of minor contamination. Moreover, sequencing allows contaminants to be identified and potentially omitted from downstream analyses.

It is plausible that common laboratory contaminants are present in natural systems; thus, it is potentially inappropriate to remove all taxa that appear in negative controls from a data set. Instead, the abundance of possible contaminants in negative controls versus that in biological samples should be considered. If a taxon occurs with high relative abundance in biological samples but is at low relative abundance in the negative control, then it is likely that the taxon is not solely present due to laboratory contamination. Determining appropriate treatment of contaminants is a topic of ongoing research (19, 21, 23). While bioinformatic and statistical guidance for dealing with contamination is nascent, at a minimum, users can flag possible contaminants and qualify inferences regarding those taxa. For study of low-biomass samples, contamination is a pressing concern (6, 20); however, for those studying systems with high microbial biomass, mild contamination is much less likely to affect inferences.

Most practitioners are now aware of the threat posed by contaminant microbial taxa; however, far less attention has been paid to the specter of cross-contamination. Cross-contamination is potentially more troubling than contamination by nuisance microbes, since the latter type of contamination should occur haphazardly among samples, whereas cross-contamination could be confounded with treatment group (e.g., among samples on a 96-well plate). Therefore, it is important to design laboratory protocols such that cross-contamination can be detected and addressed. While sequencing of negative controls can alert practitioners to catastrophic cross-contamination, such practice does little to indicate the existence of minor bouts of contamination, for instance, when a droplet from a well of a PCR plate migrates to a neighboring well (22).

Tourlousse et al. (24) recently suggested a clever approach for tracking cross-contamination through the use of synthetic oligonucleotides (often referred to colloquially as "oligos"). These authors synthesized 12 unique sequences that were approximately 1,500 nucleotides (nt) long and emulated full-length 16S rRNA genes but that had

negligible similarity to published 16S sequences (for oligonucleotide design, see reference 25). By combining three of these oligonucleotides, 220 unique mixtures can be created. Aliquots of these mixtures can be added to PCR or extraction plates, and the constituent sequences can be used to alert the user to instances of cross-contamination. Tourlousse et al. (24) suggested a way to array mixtures within plates such that neighboring wells are filled with as distinctive mixtures as possible. The downside to this approach is that it only allows approximately 60% of instances of cross-contamination between two samples to be unambiguously detected. For cross-contamination involving three or more samples, detection ability is reduced to ~0.7%.

Accordingly, we have modified the technique described by Tourlousse et al. (24) by designing more and shorter oligonucleotides. We refer to these oligonucleotides as cross-contamination checking oligonucleotides, or "coligos" for short. These coligos consist of a sequence that is complementary to that of forward and reverse primers that bookend a unique sequence taken from the report by Hawkins et al. (26) (for sequences, see Text S1 in the supplemental material). The sequences described by Hawkins et al. (26) allow for detection and correction of insertion, deletion, and substitution events while avoiding extensive internal-complementary-minimizing homopolymers and aiming for reasonably balanced GC content. For most uses, we suggest that 96 coligos is sufficient, as contamination among different 96-well plates is less likely than well-to-well contamination within a plate. However, Hawkins et al. (26) describe 10^{15} suitable sequences, if more coligos are desired.

We synthesized 96 coligos that included the popular 515/806 primer pair for 16S and also for the ITS1f/ITS2 primer pair for the internal transcribed spacer (ITS) (192 total). Any primer pair could be substituted for those we chose, including the primers used during genotyping-by-sequencing studies (e.g., see reference 27). Similarly, for transcriptomic studies, coligo sequences could be added to cDNA pools prior to adapter ligation. Coligos can be added at any point during sample processing to good effect; however, we suggest addition prior to DNA extraction (according to reference 24), thus allowing contamination during extraction to be detected. Aside from tracking sample provenance, coligos can also be used to determine if a plate has inadvertently been rotated.

To test their effectiveness, we sequenced a library containing our 96 coligos. Each coligo was added to a single well of a 96-well plate (for additional library preparation details, see Text S1), and three plates were prepared using the two-step PCR approach we describe below. Using the Illumina iSeq instrument, we obtained 2,541,033 reads. We removed the 13 bases immediately following the primer of each forward read (recall that coligos are 13 nt long) and matched these reads to our coligo sequences. We observed negligible contamination among wells—less than 0.01% of all coligo reads were out of place (see Table S1). Cross-contamination, while very minor when it occurred, did occur fairly often; 22% of wells showed some evidence of contamination. When summing read counts across replicates, variation among coligos was approximately normally distributed (see Fig. S2), though substantial variation in read count among technical replicates was observed (Fig. S2) (this variation suggests coligos should not be used as internal standards [ISDs]; see below).

Tourlousse et al. (24) reported variation in read counts among their sample provenance-tracking oligonucleotides as well and suggested that this variation is due to the necessary differences among oligonucleotides in nucleotide composition, including GC content, which have systematic effects on their abundances in sequence reads. This adds to mounting evidence that technically derived variation is a significant source of noise in sequencing data. Indeed, in a recent study of the human gut microbiome, Ji et al. (28) suggested that an abundance threshold exists below which technical variation drives among-sample differences in microbial abundances.

To determine the effectiveness of coligos for typical, likely more complex, libraries from empirical studies, we added them to an Illumina NovaSeq library (2 by 250 created using our two-step protocol; see below) containing over 10,000 replicates

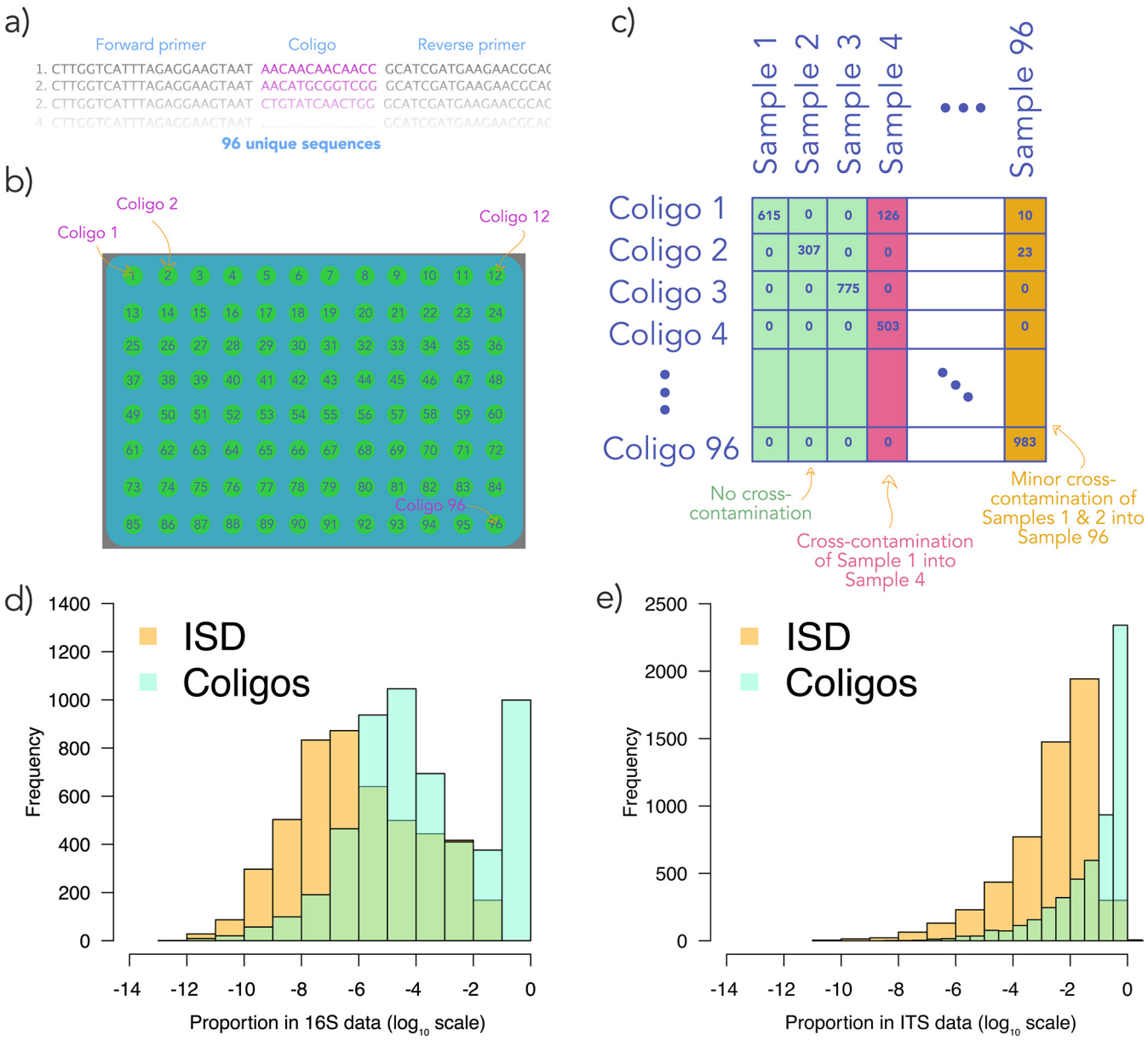

FIG 3 Cross-contamination occurs when samples are inadvertently mixed during laboratory procedures. We have developed oligonucleotides (which we refer to as coligos) to track incidences of cross-contamination. (a) To make coligos, we placed an identifiable sequence between priming sequences for the marker loci used. (b) Coligos are added to each well of a 96-well plate, preferably prior to DNA extraction. After sequencing, incidence of cross-contamination can be determined. (c) Toy data mimicking an operational taxonomic unit (OTU) table, where samples are columns and OTUs are rows. For these data, samples and coligos are matched such that sample one contains coligo one, sample two contains coligo two, and so on. In this example, samples one through three contain the single expected coligo and thus are uncontaminated. However, sample four contains many reads of coligo one, even though coligo four was the only expected sequence; thus, this sample has been seriously contaminated. Finally, sample 96 has relatively minor contamination by two foreign samples. To demonstrate the use of coligos, we injected them into a complex library composed of both 16S and ITS sequences from a variety of samples (e.g., soil, plant, and water) that differed in DNA concentration and quality. We also included a synthetic DNA internal standard in each library (see main text). The proportion of reads per sample that were associated with coligos and the ISD are shown as frequency histograms; 16S (d) and ITS (e) data are shown separately.

(including molecular identifier [MID]-labeled PCR duplicates from >5,000 samples) collected from various substrates and prepared by different scientists. These samples necessarily varied wildly in the amount and quality of DNA present. We added 0.0016 pg/μl of coligos to extracted template DNA, prior to normalization of the template to a standard concentration. We observed coligos in ~99.8% of samples. The percentage of reads within each sample ascribed to coligos was generally quite low (on the order of 10^{-8}% to 1%) (Fig. 3). This result suggests that preferential amplification of coligos, due to their short length, is unlikely to cause undesirable bandwidth capture except,

possibly, for libraries including samples with very little template DNA. In the latter case, coligo and internal standard DNA should be reduced so that the final library is not dominated by nontarget amplicons.

To determine the degree of cross-contamination present, we calculated the median percentage of coligo reads that were present within a sample but that should not have been. Contamination was very minor, with only 0.2% of coligo reads in unexpected wells (see Fig. S3). On the other hand, instances of very minor cross-contamination were common, as 79% of replicates had detectable cross-contamination. We also observed that cross-contamination by multiple samples was common; we observed more than a single foreign coligo in 44% of samples. Contamination was most likely from adjacent wells and decreased with Euclidean distance between wells (see Fig. S4), but contamination from remote wells happened often (as was also found for reference 22). It is plausible that some cross-contamination occurred during synthesis of the coligos.

Given the prevalence of minor cross-contamination that we discovered, we suggest that the use of coligos, or some analogous approach such as that described by Tourlousse et al. (24), become standard for sequence-based studies. While minor cross-contamination will not affect most statistical analyses, given that contaminants are likely only represented by a few reads, it is important to diagnose instances of severe cross-contamination and either remove those samples or otherwise mitigate the influence contaminants could have on inferences. The degree of cross-contamination that is detectable will depend upon the ratio of coligo DNA to template DNA. As more coligos are added, more minor instances of cross-contamination will be detectable. We urge researchers that are interested in using coligos to consider the concentration that we used in our libraries as a starting point and modify that concentration as required.

We note that, ideally, coligos should be added prior to DNA extraction, which we did not do in our simple experiment (because many independent researchers performed extractions using a variety of methods). Determining the appropriate amount of coligo to add to each sample can be challenging, as the amount of DNA in samples, extraction success, and sequencing depth affect the ability to recover reads from coligos. Therefore, the concentration of coligos that we added to each sample could be regarded as a starting point and optimized for a particular study system. If desired, additional coligos could be designed and added to samples during library preparation; thus, contamination during extraction could be distinguished from contamination during library preparation. Finally, we designed our coligos to be very short to reduce monetary cost; however, short sequences do not merge well because of so called "staggering" of reads; thus, we have used forward reads only to determine the incidence of cross-contamination (see the Text S1).

The apparent prevalence of cross-contamination (22) adds further weight to the notion that qualitative analyses (i.e., presence/absence-based analyses and richness) should be undertaken with care when reliant upon metabarcoding data (14, 29).

(iii) Proper usage of "spike-in" sequences as internal standards. Data output by current sequencing instruments only provide relative abundance information regarding template molecules. These data are referred to as compositional (30–32) and are a severe limitation of high-throughput sequencing, because biological insight can depend upon accurate measurement of absolute abundances (e.g., see reference 33).

Compositional data arise because sequencers have a finite output; only so many reads are generated per operative period, and those reads are parsed among samples and taxa within a sample. More reads are assigned to taxa that have higher relative abundance. Problematically, when the relative abundance of a taxon increases, it is impossible to know if this was caused by an increase in actual abundance of the focal taxon, a decrease in other taxa, or both (34). This undercuts correlational analyses and, if not corrected, can lead to erroneous inferences (32, 35). Several methods have been suggested to convert relative abundance sequence data into estimates of absolute

microbial abundances (reviewed in reference 34). One such approach is the inclusion of internal standards (ISDs), or "spike-ins," into sequencing efforts.

To use an ISD, a known quantity of a molecule is added to each replicate to be sequenced. The same ISD molecule (or mix of molecules) must be added to each replicate, as any variation within a DNA sequence can affect how well the molecule can be amplified through PCR and sequenced. Since the same amount of ISD has been added to each replicate prior to sequencing, the relative abundances of each taxon (*i*) in a replicate (*x*) can be divided by the relative abundance of the ISD in that replicate using $\frac{x_i}{x_{ISD}}$, thus converting data to a consistent scale. This normalization is effective whether read counts or proportions are used. By placing all taxa on the scale of the ISD, one can determine to what extent a change in the relative abundance of a focal taxon is due to the effect of sampling group (e.g., treatment) rather than a statistical artifact imposed by compositionality. Moreover, multiplication of ratios by the absolute abundance of the ISD (i.e., in cells, moles, or some other unit of abundance) allows the absolute abundance of co-occurring taxa in samples to be estimated (36, 37). A final benefit of an ISD is that it allows explicit statistical modeling of technical variation, which can then be subtracted from among-replicate variation for focal taxa to improve estimates of biological variation (34, 38). The latter benefit is not available when estimating absolute microbial abundance via quantitative PCR (qPCR).

To be effective, an ISD must mirror the behavior of focal microbial taxa during laboratory procedures. This is challenging because every component of a microbial ecology study has the potential to impose taxon-specific bias (e.g., see references 13, 39–42, and 43), and no single ISD can accurately account for these biases for all taxa. Three primary approaches for ISD use exist in the microbiome literature: cellular ISDs, synthetic DNA ISDs, and microbial genomic DNA. Previously, we argued that cellular ISDs or synthetic DNA ISDs should be preferred over microbial genomic DNA (34). Cellular ISDs have the advantage that they could respond to DNA extraction similarly to focal organisms; however, choosing an appropriate culturable organism can be challenging. While synthetic DNA cannot measure variation in extraction success, such ISDs can be highly flexible and cost effective (*sensu* reference 25) and may be more practical than cellular ISDs for many study designs. This is because the synthetic sequence can be optimized to model the focal organism during PCR. Ideally, a simple mixture of ISDs should be included in each sample, with constituents of the mixture designed to mimic focal taxa.

Aside from laboratory biases, various biological and sampling contingencies can undercut the performance of ISDs (reviewed in reference 34), including unaccounted-for differences in sample mass and density, presence of PCR and extraction inhibitors, variation in the lysability of microbial cells (and host cells, if examining endosymbiont assemblages), and copy number variation (Box 1). Thus, careful planning is required to ensure ISDs perform effectively. While adding an ISD at any step prior to DNA normalization provides some benefits, it is best to add ISDs to samples prior to DNA extraction, thus allowing accounting of biases imposed during this step.

As an example of a cost-effective ISD, we shortened one of the synthetic sequences described by Tourlousse et al. (25) from ~1,500 nucleotides to ~170 nucleotides. Short oligonucleotides can be synthesized at lower cost than long oligonucleotides; indeed, for the short ISD we created, enough molar mass for many thousands of samples can be purchased for less than several hundred US dollars. We sequenced aliquots of the ISD that spanned differences in concentration of 5 orders of magnitude and that were mixed with a fixed concentration of DNA from a mock community composed of 10 microbial taxa (for details of library preparation, see Text S1). We observed quantitative behavior of the ISD; that is, as more ISD was added to samples, the proportion of reads assigned to the ISD increased concomitantly (see Fig. S1). We found that two exact sequence variants (ESVs; also referred to as amplicon sequence variants [ASVs]) were associated with the ISD. Tourlousse et al. (25) also reported a proliferation of ESVs for each of their ISDs, depending upon filtering, trimming, and other bioinformatic steps employed. To better determine ISD relative abundance, all ESVs that aligned to

BOX 1: THE PROBLEM WITH COPY NUMBER VARIATION

The rRNA gene is the standard locus used to characterize microbiomes. Unfortunately, it occurs multiple times throughout the genomes of many microbes and their hosts (44–46), a phenomenon referred to as copy number variation (CNV). CNV is problematic because it means that PCR of the same molar mass of DNA from different organisms will not result in the same amounts of amplicons for each organism. This distorts the relative abundances of sequence counts away from the true relative abundances of the organisms. Several databases (e.g., see reference 47) and software tools exist that provide some insight into CNV for common taxa, in many cases using phylogenetic reconstruction to predict copy number as a character state (e.g., see references 48, 49, and 50). However, copy number can vary both within and among microbial taxa, which undercuts the utility of currently available CNV prediction methods. For instance, Lofgren et al. (46) reported 72 to 156 copies of the ITS among 12 isolates of the fungus *Suillus brevipes*. Since very little is known regarding CNV in the natural world, predictions derived from phylogenetic reconstruction should be regarded as hypotheses and are likely inaccurate. Indeed, Louca et al. (51) demonstrated that most CNV correction tools performed very poorly for the majority of taxa.

Because among-taxa CNV of standard metabarcoding loci is commonplace, it is not generally possible to compare the absolute abundances of different microbial taxa using single-locus sequencing data. However, it is possible to determine shifts in the relative and absolute abundances of a single taxon among sampling groups, assuming that CNV of that taxon is not confounded with the sampling group. For every metabarcoding study, thought should be given to how CNV could be affecting results and biasing inferences.

the ISD were summed for each replicate, and this sum was used in proportion calculations. We recommend others use a similar approach.

To demonstrate the use of our ISD in a complex library, we added 0.0005 pg/μl of ISD to each sample within the aforementioned Illumina NovaSeq library (the same library used to test coligo effectiveness). The ISD captured a low proportion of reads (Fig. 3) within each sample but was present in ~91% of samples.

While our approach to ISD design is simplistic and, for many studies, an ISD mixture would be superior, we suggest that the approach described here represents an easy-to-adopt baseline way to circumvent the problems of compositionality within sequencing data (for more, see reference 34).

Nucleic acid extraction and other preliminaries to library preparation. Prior to library preparation, nucleic acids must be extracted from cells, PCR inhibitors removed, extraction success quantified in terms of DNA yield, and internal standards added (if desired). Combined, these steps typically require much more time and expense than library preparation and sequencing. Indeed, sample weighing and grinding are often the most time-consuming laboratory steps, as they are difficult to automate (52). Additionally, these steps typically require considerable expenditure of single-use plastic consumables (i.e., pipette tips and microcentrifuge tubes). We are aware of pipette tip-washing tools (e.g., those made by Grenova, Richmond, VA), but these tools are currently unsuitable for pipette tips with filters. Moreover, many pipette tips, because of their filtering inserts, which are typically a different plastic then the body of the tip, are not recyclable, thus contributing to the large amount of laboratory waste generated worldwide (53).

After samples are weighed, DNA extraction can begin. For some sample types, notably plant tissue, mechanical lysis is the first step of the extraction process and, in our

experience, the most critical for obtaining good DNA yield. We suggest that mechanical lysis be considered regardless of substrate (also see references 54 and 55). Grinding can be expedited through the use of a bead mill (e.g., the Qiagen TissueLyser), which is a simple device that shakes tubes containing samples and 2- to 5-mm metal beads (3 mm is our preferred size). Beads can be acid washed and reused. Tungsten carbide beads and stainless-steel beads both work well; however, tungsten carbide beads are more expensive. Steel ball bearings can be purchased extremely cheaply but may not last as long as tungsten carbide beads.

DNA extractions are typically performed either in single-tube or 96-well plate formats; recently, however, various manufacturers have begun offering 384-well plate format kits as well (e.g., the TaKaRa NucleoMag 384 plant kit). Prepackaged DNA extraction kits are available for numerous substrates (e.g., animal and plant tissue, soil, and various culture media). Many of these kits rely on solid-phase extraction (SPE) technology, where nucleic acids are suspended in a high-salt solution and passed through a column, where they are retained within the stationary phase. Once the nucleic acids are bound to the stationary phase, solvents are used to separate and remove unwanted proteins and cellular detritus prior to elution of nucleic acids. This technique can provide high-concentration, pure nucleic acids, but it is time consuming and costly. Notably, these kits can remove PCR inhibitors that are common to certain substrates, such as humic acid in soil and phenols in plants (56). These inhibitors not only undercut PCR but also can reduce the efficacy of internal standards, because they may cause inconsistency among samples in extraction yield (34). A do-it-yourself approach to 96-well plate extraction reliant on Whatman filter paper as a solid phase has been suggested (57, 58). Since a filter paper SPE approach can reduce costs and chemical exposure, it is particularly suited to laboratories interested in substituting monetary cost for time cost and for training laboratories that must minimize exposure to toxins.

Costly SPE kits have been widely used for microbiome studies; however, magnetic bead-based extraction represents an appealing alternative, because it uses fewer consumables and can greatly reduce expense. Briefly, magnetic beads are hybridized to a molecule with affinity for DNA, such as short single-stranded oligonucleotides (59, 60). The beads are then suspended in solution with DNA, which the beads bind to. A magnetic field is used to pull the bound DNA out of solution, allowing contaminants to be washed away. The beads are superparamagnetic, which means they are magnetic only when in the presence of an external magnetic field. A variety of commercially available kits are available. Alternatively, Oberacker et al. (60) provide instructions for the synthesis of magnetic beads consisting of ferrite nanoparticles encased in silica or methacrylic acid. They also provide templates for 3D printed magnet racks that are much less expensive than commercial varieties. Through synthesizing beads, per sample nucleic acid extraction costs can reportedly be reduced to $0.32 (not including plastic consumables).

The bulk of the time involved in DNA extraction, using either SPE or magnetic beads, is spent isolating DNA from PCR-inhibiting compounds. DNA purification has traditionally been required to ensure PCR success; however, improvements have been made to DNA polymerases that allow them to bind to DNA in the presence of inhibitors (e.g., the Thermo-Scientific Phire and Phusion polymerases). So-called "direct PCR" technology uses these improved polymerases to amplify template sequences according to a simple tissue-lysing step (e.g., bead beating or incubation in hot water with degrading enzymes). Direct PCR has been shown to generate similar results to those if traditional extraction techniques (61, 62) but saves a great deal of time and consumables. Recently, Kai et al. (54) reported that mechanical disruption is a necessary preliminary step to direct PCR because of differences in cell wall morphology among microbial taxa, which affect cell lysability and thus extraction yield. Direct PCR requires very little sample mass, which can be a benefit of the approach. Because direct PCR involves little more than a shift to using different polymerases, the adoption of the

technique should not require extensive changes to laboratory protocols and thus deserves more attention.

(i) Logistical considerations for extraction of microbial DNA. Each DNA extraction method has its own bias that discriminates against the DNA of certain microbial taxa (e.g., see references 63, 64, and 41). Therefore, to the extent possible, identical extraction methods should be used for all samples and care taken to avoid confounding the extraction technique with the sampling group. Generally, to avoid batch effects, samples should be randomized prior to extraction. We also advocate, if possible, the inclusion of technical replicates during extraction, where the same sample is subdivided and extracted multiple times. Data from these samples can be used to estimate the amount of intrasample biological variation that is present (28).

If possible, positive controls containing cells from taxa of known interest should be subjected to extraction protocols. For exploratory work, a cellular mock community could be included as a positive control. Such mock communities could be made from available culture stocks or purchased (e.g., ZymoBIOMICS provides such an offering). A mock community can also confirm the results of bioinformatics (i.e., the number of ESVs obtained from bioinformatics matches or does not match expectations). Finally, positive controls can be titrated, such that the limit of detection for a particular number of cells or number of molecules can be approximately quantified (65).

DNA extraction is tedious and time consuming. Unfortunately, automation options for DNA extraction that are suitable for most labs are either relatively low throughput, expensive, or require chaperoning (but see reference 66). This is because most extraction techniques require centrifugation to pass solutions through a solid phase or otherwise separate chemical mixtures. The loading and unloading of centrifuges is expensive to automate; consequently, some manufacturers have developed a vacuum manifold system that can pull solutions through a filter. The benefits of such a system include little required oversight after sample loading, assuming the vacuum is strong enough to pull DNA through the solid phase (clogging is a concern when processing soil and other challenging substrates). Alternatively, a simple pipette-on-a-gantry-style system can be used to automate much of the extraction process, with centrifugation steps facilitated by hand. Finally, magnetic bead-based extraction kits may be easier to automate, because magnetic plates are available for many automation systems (or could be custom fabricated).

While we advocate the use of 96-well plates during extraction, we acknowledge that the potential for cross-contamination is high during plate loading. Accordingly, to minimize contamination potential, we suggest suspending dried ground material in lysis buffer prior to transfer to plates via pipetting (this step can be automated). Recently, Custer and Dibner (52) suggested a clever method to avoid contamination during plate loading and eliminate the possibility of double-filling wells through using perforated plate seals and microcentrifuge tubes as funnels to transfer samples.

(ii) DNA normalization. It is sometimes desirable to standardize the concentration of extracted DNA prior to PCR and sequencing. Otherwise samples with more DNA are expected to generate more amplicons and more sequence reads. We note that if DNAs are normalized to a standardized concentration, then an ISD should be used if estimates of absolute abundances of microbes are desired (see above). Normalization can be time consuming when the concentration of each replicate is assayed independently (e.g., with a NanoDrop spectrophotometer) and diluted or concentrated manually. Numerous microplate readers are available that can measure each well of a 96-well plate, and some models can measure up to 384 samples at a time. Additionally, multimode microplate readers are available that can measure both absorbance and fluorescence. Fluorescence-based assays are thought to provide more accurate measurements of double-stranded DNA (dsDNA) concentrations than the measurement of absorbance (67), particularly for low-concentration samples. Additionally, spectrophotometric assays are more influenced

by nontarget nucleic acids, such as RNA and single-stranded DNA, than fluorescence methods. This is because fluorescence-based methods rely on dyes with high affinity for dsDNA. A drawback of fluorescence-based assays is that they require additional reagents that add to project costs. When choosing a microplate reader, the minimum volume required for accurate quantification should be considered, as not all devices have the same requirements and repeated quantification can deplete stocks of precious template DNA. To be clear, spectrophotometry and fluorometry measure total nucleic acids present; if measurements of amplicon concentration are desired, then qPCR is a more appropriate approach (68), and none of these tools can accurately estimate cell densities, due to a lack of resolution, CNV (Box 1), and differences in genome size among taxa. Additionally, normalization of samples that contain various amounts of eukaryotic DNA can be challenging, as eukaryotic DNA often will not amplify during PCR.

Library preparation. DNA extracted from samples becomes the template from which targeted loci are typically amplified via PCR for detection by sequencing machines. The process of modifying template DNA for sequencing is referred to as "library preparation." Primer choice is a critical consideration when preparing a library and should be in accordance with recommendations for the characterization of focal taxa (Box 2).

BOX 2: CHOOSING PRIMERS FOR PCR

The majority of single-locus assays of microbial biodiversity rely on sequencing some portion of the rRNA. Often, the V4 region of the 16S locus is chosen to characterize bacteria, and the internal transcribed spacer (ITS) is chosen to characterize fungi. These loci are commonly used because their sequences evolve rapidly enough to distinguish organisms recognized as belonging to different species or variants within species. Sequences of these "barcoding" loci from known organisms are stored in various taxon-specific databases (e.g., SILVA, Greengenes, and UNITE databases [69–71]). Comparison of sequences to these databases allows for detection of known organisms in a sample. Moreover, a taxonomic hypothesis can be generated for a sequence that is not present in these databases by assuming that sequence similarity is predictive of taxonomy. For example, a sequence that is not present in the database but is similar to various known *Actinobacteria* sequences that are in the database can be hypothesized as being from an actinobacter.

While certain primer pairs have become favored among microbial ecologists (e.g., 515/806 for 16S and ITS1f/ITS2 for ITS), every primer pair necessarily imposes some taxonomic bias (40, 72–74). For instance, the first part of the ITS tandem repeat (ITS1) can recover slightly fewer fungal taxa than its counterpart (ITS2), leading Nilsson et al. (13) to suggest ITS2 be adopted as the preferred fungal barcoding locus. Because of the inevitable biases characteristic of a given primer pair, using multiple primer pairs for a study can be beneficial, because it expands the taxonomic breadth that can be surveyed. However, using multiple primers can increase time and consumable costs, and so a balance must be struck between obtaining adequate taxonomic breadth to address the question being asked and study cost.

We also note that the choice of primer and marker locus determines the ideal read length desired from a sequencer. At the time of writing, NovaSeq machines can provide read lengths of up to 250 bases. Even when using paired reads, this may not be sufficient length to recover the whole marker locus from all organisms. In our work, we have noticed that this is a particular problem with using the ITS1 locus for fungi. We often cannot merge paired reads and must resort to concatenating reads or analyzing forward reads only.

Several simple modifications can be made to existing operating procedures to reduce library preparation costs. For example, we advocate performing PCR in duplicate instead of in triplicate reactions (75, 76) and using minimal total reaction volumes. Marotz et al. (75) suggested that singlet PCR is sufficient; however, we have found duplicate PCR to be worth the additional cost, because it provides more assurance that a sample will not be neglected due to minor errors during pipetting. Moreover, if PCR replicates are assigned unique MIDs, then performing PCR in duplicates can provide ample quantification of technical variation induced by library preparation. Typical reaction volumes for PCR range from 20 μl to 25 μl. Decreasing the reaction volumes can reduce costs and does not appear to negatively affect results. Indeed, we have achieved satisfactory amplification using a reaction volume of 15 μl, and Minich et al. (18) have reported good results using as little as a 5-μl total reaction volume.

When designing library preparation protocols, it can be beneficial to minimize PCR cycles to reduce potential accumulation of technical error due to polymerase infidelity. While the optimal number of PCR cycles will vary with primer choice, template quality, reagents, and so on, Sze and Schloss (77) suggested that cycle count be kept below 35 if possible. High cycle count can also lead to undesirable amplification of very rare template molecules, such as those derived from technical error. Regardless of the methodology chosen, we suggest that greater attention be paid to accurately communicating library preparation methods, as they greatly affect the reproducibility of a study. In our brief survey of literature norms (see the introduction), we often were unable to determine the exact library preparation methods used, particularly when preparation was outsourced to a service provider.

Multiplexing during library construction. Typically, researchers wish to sequence many samples simultaneously—a process referred to as multiplexing. Such an approach is made possible by attaching unique, laboratory-synthesized DNA sequences (referred to as oligonucleotides or colloquially as "oligos") to template molecules during library preparation. These oligonucleotides are referred to as molecular "barcodes" (note that this term is also used for amplicons in other contexts, so we prefer the following term) or "molecular identifiers" (MIDs), and because they are sequenced along with template DNA, they allow attribution of sequences to samples (Fig. 4; an additional term for MIDs in the literature is "index"). Here, we discuss several multiplexing approaches and provide a brief overview of their benefits and challenges. We then present a novel method that improves upon existing techniques.

Multiplexing can be accomplished through appending a MID to one end of a template molecule, referred to as "single indexing," or by appending MIDs to both ends of a template molecule. This process is referred to as "dual indexing" and can drastically reduce oligonucleotide costs, because fewer oligonucleotides are required to achieve the desired level of multiplexing (78, 79). For example, two unique MIDs can be arranged at either end of a template molecule in four combinations. Thus, dual indexing allows for n^2 combinations of n MIDs. Triple- and even quad-indexing techniques have been described and can allow extreme multiplexing using few MIDs (e.g., see reference 80). Such approaches can be very efficient in terms of oligonucleotide purchase; however, they may require more complex library preparation and bioinformatics. Thus, we suggest that it is simpler to increase the number of dual-indexing MIDs to achieve the desired level of multiplexing.

A primary difference among multiplexing strategies for Illumina instruments is the placement of the MID in the template molecule in relation to the adapter sequence. Adapters are sequences that include regions that bind to the flow cell and thus allow the template molecule to adhere to the flow cell and be sequenced. Illumina adapters also include MIDs and primers that allow the MIDs to be sequenced (see below). More generally, MIDs can be placed within adapters, according to Illumina's design, or between adapters that solely bind to the flow cell and locus-specific primers (Fig. 4).

The Illumina style of dual indexing relies on MIDs that are placed within the adapter sequence so the entire template to be sequenced is of the form (5′ end): flow cell binding

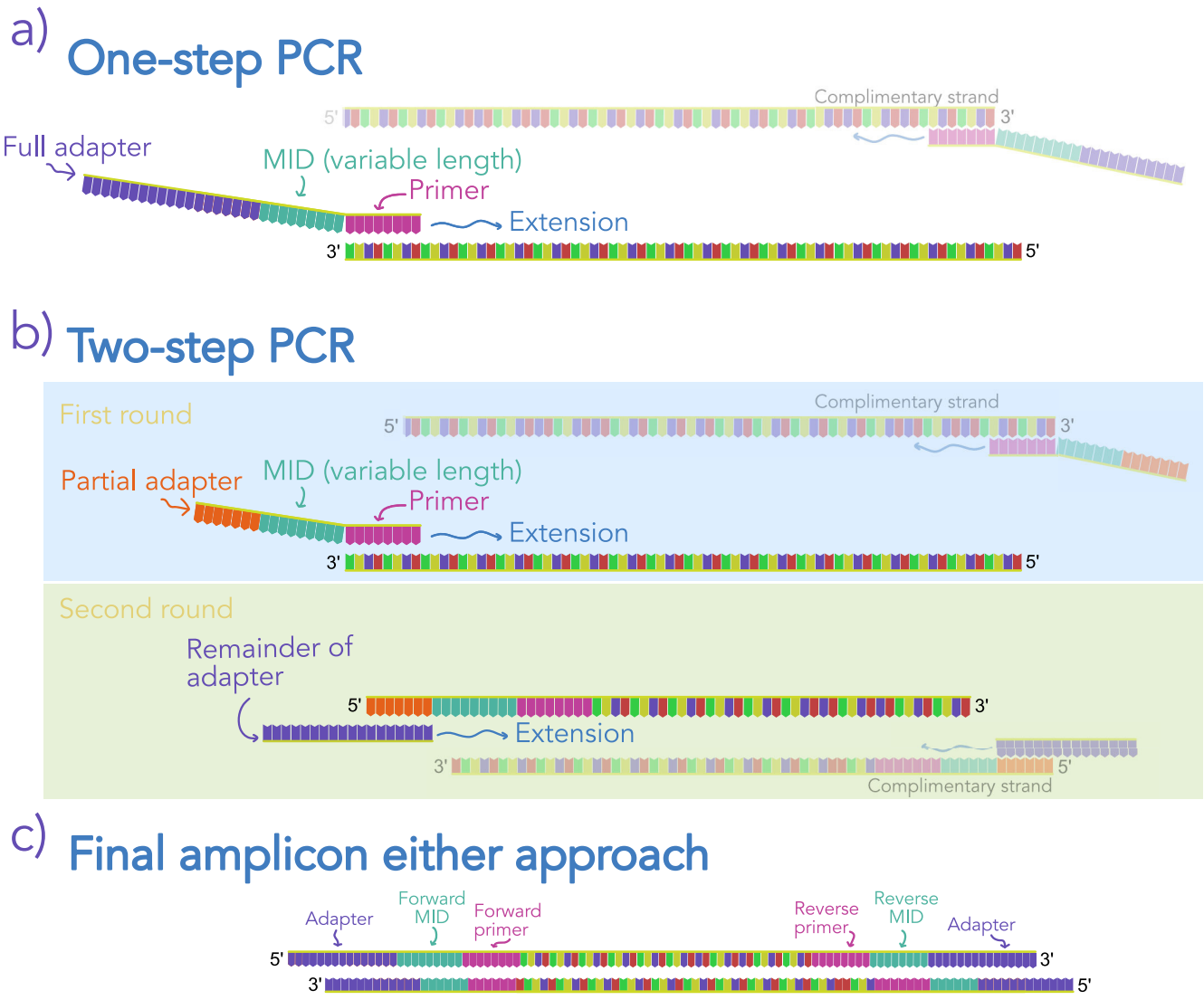

FIG 4 Visual description of library preparation, which relies on either a one-step (a) or two-step (b) PCR-based procedure. The two techniques differ in the length of oligonucleotides that are required. For most uses, a one-step approach is more cost effective but requires longer oligonucleotides, which only recently became inexpensive and widely available. In the one-step procedure, an Illumina flow cell adapter and molecular identifier (MID; for multiplexing) are added simultaneously when amplifying the template, whereas in the two-step procedure, the template is amplified and a MID sequence is added upstream of the priming region, during an initial round of PCR. A portion of the Illumina flow cell adapter is also added, which serves as the anchor for the second round of PCR, where the remainder of the adapter is added to the amplicon. (c) Both the one- and two-step procedures result in the same amplicon and both rely on dual indexing for multiplexing. We suggest using MIDs that vary in length and that sufficiently differ from one another to allow for unambiguous sample assignment in the event of technically derived sequence variation. Variable-length MIDs inject sequence heterogeneity into the beginning of reads, which improves sequencing performance through prevention of cluster loss on Illumina instruments (see main text).

region–MID (i7)–primer for i7 index–locus-specific primer–template–reverse locus-specific primer–primer for i5 index–MID (i5)–flow cell binding region. During sequencing, the instrument starts at the beginning of the locus-specific primer sequence and performs sequencing by synthesis. Then, during the next "cycle" of the instrument, sequencing begins at the primer for the MID and sequences it; thus, the MIDs are not sequenced simultaneously with the template. This is a general description of Illumina sequencing, and different sequencers and chemistries may vary slightly.

An issue with Illumina's approach is that free adapters within the library can bind to the template along with their associated MIDs, because those free adapters have complementary sequences for the index primers. This can lead to a phenomenon known as "tag switching" or "index hopping," where a MID is assigned to the wrong sample (81–83). This in turn causes sample cross-contamination (also referred to as "cross

talk"). Illumina currently recommends using unique pairings of MIDs at either end of template molecules so that tag switching can be identified and suspect sequences removed from the data (84). While this approach adequately addresses tag switching, it greatly reduces multiplexing capacity.

Alternatively, MIDs can be placed at the 3′ end of the adapter and precede the locus-specific primer (Fig. 4); thus, the MID is internal to the adapter and sequenced simultaneously with the template. Such an approach was devised to improve multiplexing compared to early single-indexing techniques (78, 79, 85). The benefit of this approach is that it is immune to tag switching during sequencing and allows for greater multiplexing with fewer oligonucleotides. The technique's downside is that some portion of the read must be assigned to the MID sequence, because the MID is sequenced along with the template.

Multiplexing techniques also differ as to how they inject sequence heterogeneity into the library. Sequence heterogeneity is beneficial, because a common reason Illumina runs fail to provide the expected output is insufficient variation at the beginning of sequences. If not enough variation is present, the sequencer's detection system cannot distinguish between clusters of molecules being sequenced; thus, those clusters are lost and output is diminished. To prevent this, Illumina recommends a portion of sequencing libraries be composed of random portions of the PhiX bacteriophage genome, which scatters highly variable sequences of balanced purine and pyrimidine content around the flow cell. The recommended amount of PhiX depends upon the complexity of the library to be sequenced and the sequencing instrument, but in some extreme cases, Illumina recommends as high as 50% of the library be composed of PhiX (86). Unfortunately, this means that a concomitant proportion of the sequencer's output will be PhiX sequences that are of no scientific interest; consequently, PhiX addition is a costly way to increase heterogeneity within a library. To reduce the amount of PhiX required, Fadrosh et al. (78) suggested that 0- to 7-nt-long variable sequences ("heterogeneity spacers") be added directly after MIDs, but before locus-specific primers, during oligonucleotide synthesis. Fadrosh et al. (78) report that their approach allowed them to reduce the PhiX component of the library to 10% (also see reference 87).

We suggest an alternative approach that relies on variable-length MIDs to interject sequence variation into libraries. Variable-length MIDs (*sensu* reference 88) obviate heterogeneity spacers, thus reducing the portion of the sequence dedicated to nonbiological data while allowing PhiX input to be reduced. By way of demonstration, we present 96 forward and reverse MIDs for both the 16S and ITS loci (192 unique sequences) (see supplemental material). Our MIDs are a subset of those reported by Gompert et al. (89) and Parada et al. (27) and vary in length from 8 to 10 nt. Sequences were chosen to minimize internal complementarity and homopolymers and allow differentiation using edit distances (all MIDs of the same length are at least a Levenshtein edit distance of two apart from one another [90]). The MIDs are directly followed by sequence that corresponds to the primer region and preceded by the Illumina adapter. We use all 96 by 96 unique combinations of forward and reverse oligonucleotides to support multiplexing of 9,216 samples. If additional multiplexing is desired, then additional MIDs can be designed easily (the extension of MIDs by a few bases leads to many more sequences meeting the aforementioned criteria for distinguishable MIDs [26, 90]). Incorporating amplicons of multiple loci (e.g., ITS and 16S) can also increase library heterogeneity.

We have incorporated our MIDs into oligonucleotides that also include Illumina flow cell-binding regions and primers for target loci. This allows us to make a library using a single round of PCR, which reduces consumable costs (this is similar to the method used by the Earth Microbiome project [91]). We also present a variant of this protocol that uses shorter oligonucleotides and two rounds of PCR (Fig. 4). The difference between the one-step and two-step protocols is that flow cell adapters are added in a second round of PCR during the two-step approach (Fig. 4). The potential benefit of a two-step protocol is that shorter oligonucleotides can be used than those required by a one-step protocol. Shorter oligonucleotides cost less than longer oligonucleotides, and so it may

be possible to reduce costs somewhat for small batches of samples through the use of a two-step approach (though for large batches of samples, there likely will be no cost savings; see below). Additionally, the primers used to add flow cell adapters during the two-step procedure can be paired with any locus-specific primer. Thus, the two-step procedure is quite flexible and could be useful for research groups that wish to sequence multiple loci or use multiple primers and that wish to avoid the large initial cost of longer oligonucleotides.

We tested our two-step library preparation strategy through sequencing a 16S library containing only the ZymoBIOMICS mock community (Zymo Research, Irvine, CA), which includes eight bacterial taxa and two fungal taxa. We used the Illumina iSeq (paired-end 2 by 150) for sequencing. We recovered all expected taxa from those samples. Subsequently, we sequenced 4,608 samples (each PCR replicate had a unique MID pairing, for a total of 9,216 replicates), consisting of both 16S and ITS amplicons isolated from a variety of substrates (e.g., soil, water, and plant tissue), on the Illumina NovaSeq instrument (this is the same library used to demonstrate coligo and ISD use, as mentioned above). These samples were collected by a number of researchers and extracted using a variety of solid-phase extraction kits. Thus, this library represents the heterogeneity that could be encountered by an active sequencing laboratory. The library was prepared as described in Text S1. We obtained 630,568,959 paired reads that mapped to samples, with an average of 136,842 reads per sample. We obtained more reads for ITS than 16S, but median read count across samples for both loci was high (median read count for 16S, 26,377; median read count for ITS, 123,441). While read count per MID pair necessarily varied depending upon template quality, which varied among substrates and projects, we recovered reads from all samples. We have since obtained similar results from three more libraries of similar complexity and sample count that were sequenced on the NovaSeq instrument.

Given the success of our two-step protocol (which we developed first), we expected the one-step version to provide good results as well. To test this, we sequenced a one-step library containing both 16S and ITS sequences on the Illumina iSeq (paired-end 2 by 150). The library contained DNA extracted from snow, coligos, and ISD. We obtained 3,840,079 reads, which included 800,471 ITS reads and 3,039,608 16S reads (mean of 5,000 sample^{-1} locus^{-1}). These results confirm the utility of a one-step PCR approach. By our calculations, adopting a one-step procedure can lead to cost recovery of the initial oligonucleotide expenditure after just a few library preparations (assuming multiplexing of several thousand samples per library).

Library clean up. Prior to sequencing, libraries will typically require some form of "cleanup," where unused primers, deoxynucleoside triphosphates (dNTPs), and adapter sequences are removed. Cleanup is particularly important when using Illumina-style dual indexing (see above) to minimize tag switching. Additionally, cleanup can be an effective way to maximize PCR yield when performing two-step PCR, because it prevents unspent primers from the first step of PCR from being amplified during the second round. Unfortunately, library cleanup can be quite costly; so, if the protocol allows, cleanup should be performed only once on pooled MID-labeled DNAs.

Cleanup can be accomplished through chromatography or ethanol precipitation (92, 93), but these methods are cumbersome and time consuming and have fallen out of favor due to their low throughput. A modernized analogous approach relies on the BluePippin instrument (Sage Science, Beverley, MA), which uses a combination of pulsed-field electrophoresis and spectrophotometry to automatically separate and output sequences of a certain size range. Automation of size selection in this way can more precisely select a desired sequence length, is less prone to contamination, and saves time compared to manual gel electrophoresis (94), but it does require specialized equipment and consumables.

Magnetic beads can also be used to perform PCR cleanup (e.g., the popular Axygen kits; Corning, Corning, NY, USA). These beads are nanoscale magnetic particles that bind to DNA. Size selection is achieved via the ratio of beads to DNA; larger DNA molecules preferentially bind to the beads, and as more beads are added, smaller DNA molecules are also bound.

Application of a magnetic field allows the bound DNA to be pulled out of solution. Magnetic bead cleanups do not require specialized equipment (a strong magnetic strip is all that is required); however, the beads are costly. As with bead-based nucleic acid extraction, beads can be synthesized oneself for substantial cost savings. For instance, Oberacker et al. (60) reported that synthesis of beads can reduce costs to approximately $0.05 USD per sample.

Enzymatic PCR cleanup represents an alternative to magnetic bead-based protocols. Cleanup is accomplished through enzymatic degradation of single-stranded DNA and dephosphorylation of surplus dNTPs (95) through the combined action of exonuclease I and shrimp alkaline phosphatase. These enzymes can be purchased pure, and several manufacturers package them as kits (e.g., ExoSAP-IT [Affymetrix, Santa Clara, CA]). When performing cleanup using these enzymes, no loss of template is reported for amplicons of various sizes, including short amplicons approximately 100 nucleotides long (96).

Sequencing symbionts: how to deal with unwanted host DNA. The amplification and sequencing of nontarget DNA is a particular challenge for microbial ecologists studying host-associated symbionts. For instance, botanists interested in the bacteria within plants must contend with chloroplast DNA (cpDNA), which is amplified by many commonly used rRNA primers. Indeed, in some cases, 90% or more of reads recovered from plant tissues are cpDNA (97–99). Similar challenges face researchers interested in symbionts within animals, given the abundance of host nuclear rRNA and mitochondrial DNA (mtDNA). Accordingly, researchers have explored three main avenues to reduce nontarget DNA during microbiome sequencing: the use of more selective primers, targeted removal of unwanted sequences from extracted nucleic acid pools, and separation of nontarget cells prior to extraction. For many use cases, such as the characterization of certain taxa, selective primers may be the ideal tool, as they are inexpensive and easily obtainable. However, for exploratory research seeking to broadly characterize microbial assemblages, the taxonomic biases imposed by restrictive primers are undesirable.

Perhaps the most widely used technique by microbial ecologists to avoid amplification of nontarget sequences is peptide nucleic acid (PNA) clamping (100–105). PNAs are oligonucleotides that are complementary to the sequence to be suppressed. The bases in a PNA are attached to a neutral backbone of *N*-(2-aminoethyl)-glycine instead of charged phosphate groups (106, 107). This neutrality allows for a stronger bond between PNAs and single-stranded DNA than what would be experienced during DNA-to-DNA bonding. This allows the link between PNA and its target to persist through PCR, thus blocking the action of DNA polymerase.

Lundberg et al. (101) demonstrated the use of PNAs to suppress plastid DNA in a study of the microbiome of *Arabidopsis thaliana*. Fitzpatrick et al. (108) explored the limits of this approach through sequencing of 32 plant taxa from across the angiosperm phylogeny. In this study, PNA addition was reported to suppress cpDNA for each host taxon, in some cases by up to 65%. However, even single-base mismatches between target sequences and the PNA caused a reduction in performance. A concern with PNA use is that they may impede amplification of target taxa, thus imposing taxonomic biases on sequencing results (e.g., see reference 98).

PNAs can be made to mimic any sequence but are typically used to block either cpDNA or mtDNA. In our own work on plant microbiomes, we have noticed that when PNAs are added to block cpDNA, this can lead to a higher proportion of remaining reads being allocated to mtDNA (unpublished data). Thus, we suggest that researchers consider the use of multiple PNAs to suppress both cpDNA and mtDNA simultaneously.

Several additional methods to selectively reduce the relative abundance of nontarget reads have been proposed but have gained less traction than PNAs among microbial ecologists. For instance, Green and Minz (109) suggest restriction endonuclease digestion of double-stranded DNA created using primers specific to nontarget sequences; the remaining DNA can then be amplified using more general primers. Similarly, Dolinšek et al. (110) use enzymatic oligonucleotides to selectively degrade target RNA. Magnetic bead pulldown methods have also proven effective at reducing nontarget sequences. For

example, Feehery et al. (111) used magnetic beads with methyl-binding domains that specifically bind to a portion of mammalian and fish DNA to reduce the presence of these nontarget molecules (also see reference 112) (commercially available kits include the NEBNext microbiome DNA enrichment kit). Yigit et al. (113) extended this approach to plant tissues and selectively shifted the ratio of nuclear to organellar DNA obtained from five model angiosperm taxa (also see reference 114).

A very different approach to minimizing the concentration of nontarget DNA in libraries is to separate or degrade nontarget cells prior to DNA extraction. For example, the MolYsis kit (Molzym, Bremen, Germany) uses chaotropic solutions to selectively degrade host cells that have weaker cell walls than many microbes (this approach will not work for plants, given the rigid cell walls present in plant tissue). Thoendel et al. (115) suggested that the MolYsis kit outperformed the NEBNext microbiome DNA enrichment kit for removal of nontarget DNA taken from infections of prosthetic joints (also see references 116, 117, and 118). A similar but much more cost-effective approach was demonstrated by Marotz et al. (119) that relies on selective lysis of mammalian cells followed by propidium monoazide treatment (this publication suggests a $0.15 USD cost per sample for this method).

Alternatively, flow cytometry and microfluidics can be used to remove nontarget cells from samples. For example, Wu et al. (120) separated bacteria from human blood cells via microfluidics. Flow cytometry is often used to sort cells by phenotype (121) and can be used to distinguish cells in terms of DNA content (122–124). Such an approach could plausibly be extended to separate eukaryotic cells from microbial cells based on DNA content; however, we are unaware of any studies using this technique. While labor intensive, centrifugation can also be used to separate cells; for example, Utturkar et al. (125) used a combination of centrifugation and cytometry to separate microbial cells from plant cells and allow single-cell genomics of the microbial fraction.

Finally, methods employed by functional genomicists to normalize cDNA libraries could be useful for microbial ecologists. In this context, normalizing refers to manipulating the relative abundances of molecules in libraries to reduce the variation in those abundances. A variety of strategies have been proposed that rely on the activity of different enzymes to selectively degrade the most abundant nucleic acids within a solution (briefly reviewed in reference 126). For example, Zhulidov et al. (127) use a duplex-specific nuclease isolated from Kamchatka crabs that attacks double-stranded DNA. Since complementary sequences of abundant single-stranded DNAs (ssDNAs) are more likely to encounter one another during reassociation, the enzyme can be used to attack these duplexes during a short incubation period and then be deactivated. The resulting library has a higher proportion of low-frequency sequences. The same rationale is applied by Ramond et al. (126) to suppress amplicons of abundant microbial taxa through the activity of S1 nuclease. To our knowledge, the benefits of normalizing DNA amplicon libraries have not been studied thoroughly by microbial ecologists. We suggest these techniques have strong potential to improve qualitative studies of the rare biosphere (128), such as when assaying the presence of rare pathogens (e.g., in wastewater) or dormant taxa.

We note that any method to reduce nontarget sequences will likely impose undesirable taxonomic biases on the resulting library. Consequently, researchers should ensure that suitably complex mock communities are included in libraries to quantify and document these biases.

Conclusion. Here, we have provided a methodological "state-of-the-field" assessment for single-locus sequence-based characterization of microbiomes and presented several improvements to existing procedures. Of the techniques discussed, we particularly advocate the rapid adoption of internal standards and methods to account for cross-contamination (such as the "coligo" approach we present here). These technologies are simple and inexpensive to incorporate but can drastically improve experimental outcomes. We also found significant cost and time savings can be obtained through switching to a one-step PCR using our variable-length MIDs.

We acknowledge the challenge of staying up to date with ever-changing sequencing technology. Methods are published weekly, and determining best practices is difficult, particularly for research groups new to sequence-based characterization of microbiomes. Our goal here is to build awareness for methodological advances in hopes that adoption of these techniques will cut costs and improve research outcomes. Better methods can allow for better science, but we urge practitioners to carefully consider the needs of their study. All the techniques we describe here have both advantages and drawbacks, including the not-insignificant time cost of learning and deploying a new protocol. Thus, the best practice of all is to critically appraise a method and determine its suitability for a particular experimental design given logistical constraints.

MATERIALS AND METHODS

Data availability. Oligonucleotide sequences for coligos, a custom demultiplexing script, and the library preparation procedures we discuss can be found at https://github.com/JHarrisonEcoEvo/Genome_Technologies_lab_of_Univ_Wyoming.git.

We provide example sequencing data created using our one-step PCR library preparation procedure at https://mountainscholar.org/handle/20.500.11919/7186. These data were generated by the Illumina iSeq instrument and contain 16S and ITS sequences from snow collected by Abigail Hoffman. At this same URL, we also have provided iSeq data of a library containing only coligos, mock community, and internal standard (at various concentrations). This library was created using our two-step PCR procedure.

Example NovaSeq data generated using our two-step library preparation procedure can be found at https://hdl.handle.net/20.500.11919/7166.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TEXT S1, PDF file, 0.1 MB.

TEXT S2, PDF file, 0.1 MB.

FIG S1, PDF file, 0.1 MB.

FIG S2, PDF file, 0.1 MB.

FIG S3, PDF file, 0.1 MB.

FIG S4, PDF file, 0.1 MB.

TABLE S1, PDF file, 0.1 MB.

ACKNOWLEDGMENTS

We thank the many scientists whose work we referenced and built upon and two anonymous reviewers for comments on an earlier draft of the manuscript. We also would like to thank the members of the Microbial Ecology Collaborative at the University of Wyoming for letting us use their samples and sequencing data. Computing was performed using the Teton Computing Environment at the Advanced Research Computing Center, University of Wyoming, Laramie (https://doi.org/10.15786/M2FY47). Thanks to Shannon Harris and Muhammad Saqib for assistance with portions of the laboratory work.

This research was supported by National Science Foundation award EPS-1655726.

REFERENCES

1. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nat Microbiol 1:16048. https://doi.org/10.1038/nmicrobiol.2016.48.
2. McFall-Ngai M, Hadfield MG, Bosch TCG, Carey HV, Domazet-Lošo T, Douglas AE, Dubilier N, Eberl G, Fukami T, Gilbert SF, Hentschel U, King N, Kjelleberg S, Knoll AH, Kremer N, Mazmanian SK, Metcalf JL, Nealson K, Pierce NE, Rawls JF, Reid A, Ruby EG, Rumpho M, Sanders JG, Tautz D, Wernegreen JJ. 2013. Animals in a bacterial world, a new imperative for the life sciences. Proc Natl Acad Sci U S A 110:3229–3236. https://doi.org/10.1073/pnas.1218525110.
3. Vacher C, Hampe A, Porté AJ, Sauer U, Compant S, Morris CE. 2016. The phyllosphere: microbial jungle at the plant–climate interface. Annu Rev Ecol Evol Syst 47:1–24. https://doi.org/10.1146/annurev-ecolsys-121415-032238.
4. Ryan RP, Germaine K, Franks A, Ryan DJ, Dowling DN. 2008. Bacterial endophytes: recent developments and applications. FEMS Microbiol Lett 278:1–9. https://doi.org/10.1111/j.1574-6968.2007.00918.x.
5. Stanton C, Gardiner G, Meehan H, Collins K, Fitzgerald G, Lynch PB, Ross RP. 2001. Market potential for probiotics. Am J Clin Nutr 73:476s–483s. https://doi.org/10.1093/ajcn/73.2.476s.
6. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12:87. https://doi.org/10.1186/s12915-014-0087-z.
7. Hornung BVH, Zwittink RD, Kuijper EJ. 2019. Issues and current standards of controls in microbiome research. FEMS Microbiol Ecol 95:fix045. https://doi.org/10.1093/femsec/fiz045.
8. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall L-I, McDonald D, Melnik AV, Morton JT, Navas J, Quinn RA, Sanders JG, Swafford AD, Thompson LR, Tripathi A, Xu ZZ, Zaneveld JR, Zhu Q, Caporaso JG, Dorrestein PC. 2018. Best

practices for analysing microbiomes. Nat Rev Microbiol 16:410–422. https://doi.org/10.1038/s41579-018-0029-9.
9. U'Ren JM, Riddle JM, Monacell JT, Carbone I, Miadlikowska J, Arnold AE. 2014. Tissue storage and primer selection influence pyrosequencing-based inferences of diversity and community composition of endolichenic and endophytic fungi. Mol Ecol Resour 14:1032–1048. https://doi.org/10.1111/1755-0998.12252.
10. Vejnar CE, Giraldez AJ. 2020. LabxDB: versatile databases for genomic sequencing and lab management. Bioinformatics 36:4530–4531. https://doi.org/10.1093/bioinformatics/btaa557.
11. Bálint M, Schmidt P-A, Sharma R, Thines M, Schmitt I. 2014. An Illumina metabarcoding pipeline for fungi. Ecol Evol 4:2642–2653. https://doi.org/10.1002/ece3.1107.
12. Nguyen NH, Smith D, Peay K, Kennedy P. 2015. Parsing ecological signal from noise in next generation amplicon sequencing. New Phytol 205:1389–1393. https://doi.org/10.1111/nph.12923.
13. Nilsson RH, Anslan S, Bahram M, Wurzbacher C, Baldrian P, Tedersoo L. 2019. Mycobiome diversity: high-throughput sequencing and identification of fungi. Nat Rev Microbiol 17:95–109. https://doi.org/10.1038/s41579-018-0116-y.
14. Pauvert C, Buée M, Laval V, Edel-Hermann V, Fauchery L, Gautier A, Lesur I, Vallance J, Vacher C. 2019. Bioinformatics matters: the accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. Fungal Ecol 41:23–33. https://doi.org/10.1016/j.funeco.2019.03.005.
15. Barthels F, Barthels U, Schwickert M, Schirmeister T. 2020. FINDUS: an open-source 3D printable liquid-handling workstation for laboratory automation in life sciences. SLAS Technol 25:190–199. https://doi.org/10.1177/2472630319877374.
16. Gome G, Waksberg J, Grishko A, Walk IY, Zuckerman O. 2019. OpenLH: open liquid-handling system for creative experimentation with biology, p 55–64. Proceedings of the thirteenth international conference on tangible, embedded, and embodied interaction. TEI '19. Association for Computing Machinery, Tempe, AZ.
17. Opentrons. 2021. Lab automation has never been easier. https://opentrons.com/ot-2/. Accessed 21 June 2021.
18. Minich JJ, Humphrey G, Benitez RAS, Sanders J, Swafford A, Allen EE, Knight R. 2018. High-throughput miniaturized 16S rRNA amplicon library preparation reduces costs while preserving microbiome integrity. mSystems 3:e00166-18. https://doi.org/10.1128/mSystems.00166-18.
19. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. 2019. Contamination in low microbial biomass microbiome studies: issues and recommendations. Trends Microbiol 27:105–117. https://doi.org/10.1016/j.tim.2018.11.003.
20. de Goffau MC, Lager S, Salter SJ, Wagner J, Kronbichler A, Charnock-Jones DS, Peacock SJ, Smith GCS, Parkhill J. 2018. Recognizing the reagent microbiome. Nat Microbiol 3:851–853. https://doi.org/10.1038/s41564-018-0202-y.
21. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. 2017. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 6:226. https://doi.org/10.1186/s40168-018-0605-2.
22. Minich JJ, Sanders JG, Amir A, Humphrey G, Gilbert JA, Knight R. 2019. Quantifying and understanding well-to-well contamination in microbiome research. mSystems 4:e00186-19. https://doi.org/10.1128/mSystems.00186-19.
23. Weyrich LS, Farrer AG, Eisenhofer R, Arriola LA, Young J, Selway CA, Handsley-Davis M, Adler CJ, Breen J, Cooper A. 2019. Laboratory contamination over time during low-biomass sample analysis. Mol Ecol Resour 19:982–996. https://doi.org/10.1111/1755-0998.13011.
24. Tourlousse DM, Ohashi A, Sekiguchi Y. 2018. Sample tracking in microbiome community profiling assays using synthetic 16S rRNA gene spike-in controls. Sci Rep 8:9095. https://doi.org/10.1038/s41598-018-27314-3.
25. Tourlousse DM, Yoshiike S, Ohashi A, Matsukura S, Noda N, Sekiguchi Y. 2017. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. Nucleic Acids Res 45:e23. https://doi.org/10.1093/nar/gkw984.
26. Hawkins JA, Jones SK, Finkelstein IJ, Press WH. 2018. Indel-correcting DNA barcodes for high-throughput sequencing. Proc Natl Acad Sci U S A 115:E6217–E6226. https://doi.org/10.1073/pnas.1802640115.
27. Parada AE, Needham DM, Fuhrman JA. 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ Microbiol 18:1403–1414. https://doi.org/10.1111/1462-2920.13023.
28. Ji BW, Sheth RU, Dixit PD, Huang Y, Kaufman A, Wang HH, Vitkup D. 2019. Quantifying spatiotemporal variability and noise in absolute microbiota abundances using replicate sampling. Nat Methods 16:731–736. https://doi.org/10.1038/s41592-019-0467-y.
29. Cline LC, Song Z, Al-Ghalith GA, Knights D, Kennedy PG. 2017. Moving beyond *de novo* clustering in fungal community ecology. New Phytol 216:629–634. https://doi.org/10.1111/nph.14752.
30. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. 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 2:15. https://doi.org/10.1186/2049-2618-2-15.
31. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017. Microbiome datasets are compositional: and this is not optional. Front Microbiol 8:2224. https://doi.org/10.3389/fmicb.2017.02224.
32. Tsilimigras MCB, Fodor AA. 2016. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Ann Epidemiol 26:330–335. https://doi.org/10.1016/j.annepidem.2016.03.002.
33. Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, Wang J, Tito RY, De Commer L, Darzi Y, Vermeire S, Falony G, Raes J. 2017. Quantitative microbiome profiling links gut community variation to microbial load. Nature 551:507–511. https://doi.org/10.1038/nature24460.
34. Harrison JG, Calder WJ, Shuman B, Buerkle CA. 2021. The quest for absolute abundance: the use of internal standards for DNA-based community ecology. Mol Ecol Resour 21:30–43. https://doi.org/10.1111/1755-0998.13247.
35. 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. Proc R Soc Lond 60:359–367.
36. Smets W, Leff JW, Bradford MA, McCulley RL, Lebeer S, Fierer N. 2016. A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. Soil Biol Biochem 96:145–151. https://doi.org/10.1016/j.soilbio.2016.02.003.
37. Stämmler F, Gläsner J, Hiergeist A, Holler E, Weber D, Oefner PJ, Gessner A, Spang R. 2016. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. Microbiome 4:28. https://doi.org/10.1186/s40168-016-0175-0.
38. Risso D, Schwartz K, Sherlock G, Dudoit S. 2011. GC-content normalization for RNA-Seq data. BMC Bioinformatics 12:480. https://doi.org/10.1186/1471-2105-12-480.
39. Laursen MF, Dalgaard MD, and, Bahl MI. 2017. Genomic GC-content affects the accuracy of 16S rRNA gene sequencing based microbial profiling due to PCR bias. Front Microbiol 8:1934. https://doi.org/10.3389/fmicb.2017.01934.
40. Oshlack A, Wakefield MJ. 2009. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct 4:14. https://doi.org/10.1186/1745-6150-4-14.
41. Pollock J, Glendinning L, Wisedchanwet T, Watson M. 2018. The madness of microbiome: attempting to find consensus "best practice" for 16S microbiome studies. Appl Environ Microbiol 84:e02627-17. https://doi.org/10.1128/AEM.02627-17.
42. Risso D, Ngai J, Speed TP, Dudoit S. 2014. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol 32:896–902. https://doi.org/10.1038/nbt.2931.
43. Schori M, Appel M, Kitko A, Showalter AM. 2013. Engineered DNA polymerase improves PCR results for plastid DNA. Appl Plant Sci 1:1200519. https://doi.org/10.3732/apps.1200519.
44. Gong W, Marchetti A. 2019. Estimation of 18S gene copy number in marine eukaryotic plankton using a next-generation sequencing approach. Front Mar Sci 6:219. https://doi.org/10.3389/fmars.2019.00219.
45. Kembel SW, Wu M, Eisen JA, Green JL. 2012. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. PLoS Comput Biol 8:e1002743. https://doi.org/10.1371/journal.pcbi.1002743.
46. Lofgren LA, Uehling JK, Branco S, Bruns TD, Martin F, Kennedy PG. 2019. Genome-based estimates of fungal rDNA copy number variation across phylogenetic scales and ecological lifestyles. Mol Ecol 28:721–730. https://doi.org/10.1111/mec.14995.
47. Lee ZM-P, Bussema C, Schmidt TM. 2009. rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. Nucleic Acids Res 37:D489–D493. https://doi.org/10.1093/nar/gkn689.
48. Angly FE, Dennis PG, Skarshewski A, Vanwonterghem I, Hugenholtz P, Tyson GW. 2014. CopyRighter: a rapid tool for improving the accuracy of

microbial community profiles through lineage-specific gene copy number correction. Microbiome 2:11. https://doi.org/10.1186/2049-2618-2-11.
49. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31:814–821. https://doi.org/10.1038/nbt.2676.
50. Perisin M, Vetter M, Gilbert JA, Bergelson J. 2016. 16Stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies. ISME J 10:1020–1024. https://doi.org/10.1038/ismej.2015.161.
51. Louca S, Doebeli M, Parfrey LW. 2018. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome 6:41. https://doi.org/10.1186/s40168-018-0420-9.
52. Custer GF, Dibner RR. 2020. Modified methods for loading of high-throughput DNA extraction plates reduce potential for contamination. J Vis Exp 2020:e61405. https://doi.org/10.3791/61405.
53. Urbina MA, Watts AJR, Reardon EE. 2015. Labs should cut plastic waste too. Nature 528:479–479. https://doi.org/10.1038/528479c.
54. Kai S, Matsuo Y, Nakagawa S, Kryukov K, Matsukawa S, Tanaka H, Iwai T, Imanishi T, Hirota K. 2019. Rapid bacterial identification by direct PCR amplification of 16S rRNA genes using the MinION nanopore sequencer. FEBS Open Bio 9:548–557. https://doi.org/10.1002/2211-5463.12590.
55. Small CM, Currey M, Beck EA, Bassham S, Cresko WA. 2019. Highly reproducible 16S sequencing facilitates measurement of host genetic influences on the stickleback gut microbiome. mSystems 4:e00331-19. https://doi.org/10.1128/mSystems.00331-19.
56. Schrader C, Schielke A, Ellerbroek L, Johne R. 2012. PCR inhibitors – occurrence, properties and removal. J Appl Microbiol 113:1014–1026. https://doi.org/10.1111/j.1365-2672.2012.05384.x.
57. Gan W, Zhuang B, Zhang P, Han J, Li C-X, Liu P. 2014. A filter paper-based microdevice for low-cost, rapid, and automated DNA extraction and amplification from diverse sample types. Lab Chip 14:3719–3728. https://doi.org/10.1039/c4lc00686k.
58. Shi R, Panthee DR. 2017. A novel plant DNA extraction method using filter paper-based 96-well spin plate. Planta 246:579–584. https://doi.org/10.1007/s00425-017-2743-3.
59. Archer MJ, Lin B, Wang Z, Stenger DA. 2006. Magnetic bead-based solid phase for selective extraction of genomic DNA. Anal Biochem 355:285–297. https://doi.org/10.1016/j.ab.2006.05.005.
60. Oberacker P, Stepper P, Bond DM, Höhn S, Focken J, Meyer V, Schelle L, Sugrue VJ, Jeunen G-J, Moser T, Hore SR, von Meyenn F, Hipp K, Hore TA, Jurkowski TP. 2019. Bio-On-Magnetic-Beads (BOMB): open platform for high-throughput nucleic acid extraction and manipulation. PLoS Biol 17:e3000107. https://doi.org/10.1371/journal.pbio.3000107.
61. Flores GE, Henley JB, Fierer N. 2012. A direct PCR approach to accelerate analyses of human-associated microbial communities. PLoS One 7:e44563. https://doi.org/10.1371/journal.pone.0044563.
62. Videvall E, Strandh M, Engelbrecht A, Cloete S, Cornwallis CK. 2017. Direct PCR offers a fast and reliable alternative to conventional DNA isolation methods for gut microbiomes. mSystems 2:e00132-17. https://doi.org/10.1128/mSystems.00132-17.
63. Henderson G, Cox F, Kittelmann S, Miri VH, Zethof M, Noel SJ, Waghorn GC, Janssen PH. 2013. Effect of DNA extraction methods and sampling techniques on the apparent structure of cow and sheep rumen microbial communities. PLoS One 8:e74787. https://doi.org/10.1371/journal.pone.0074787.
64. Kennedy NA, Walker AW, Berry SH, Duncan SH, Farquarson FM, Louis P, Thomson JM, UK IBD Genetics Consortium. 2014. The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. PLoS One 9:e88982. https://doi.org/10.1371/journal.pone.0088982.
65. Minich JJ, Zhu Q, Janssen S, Hendrickson R, Amir A, Vetter R, Hyde J, Doty MM, Stillwell K, Benardini J, Kim JH, Allen EE, Venkateswaran K, Knight R. 2018. KatharoSeq enables high-throughput microbiome analysis from low-biomass samples. mSystems 3:e002189-17. https://doi.org/10.1128/mSystems.00218-17.
66. Marotz C, Amir A, Humphrey G, Gaffney J, Gogul G, Knight R. 2017. DNA extraction for streamlined metagenomics of diverse environmental samples. Biotechniques 62:290–293. https://doi.org/10.2144/000114559.
67. Singer VL, Jones LJ, Yue ST, Haugland RP. 1997. Characterization of PicoGreen reagent and development of a fluorescence-based solution assay for double-stranded DNA quantitation. Anal Biochem 249:228–238. https://doi.org/10.1006/abio.1997.2177.
68. Nakayama Y, Yamaguchi H, Einaga N, Esumi M. 2016. Pitfalls of DNA quantification using DNA-binding fluorescent dyes and suggested solutions. PLoS One 11:e0150528. https://doi.org/10.1371/journal.pone.0150528.
69. Balvočiūtė M, and, Huson DH. 2017. SILVA, RDP, Greengenes, NCBI and OTT – how do these taxonomies compare? BMC Genomics 18:114. https://doi.org/10.1186/s12864-017-3501-4.
70. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–5072. https://doi.org/10.1128/AEM.03006-05.
71. Nilsson RH, Larsson K-H, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L, Saar I, Kõljalg U, Abarenkov K. 2019. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Res 47:D259–D264. https://doi.org/10.1093/nar/gky1022.
72. Apprill A, McNally S, Parsons R, Weber L. 2015. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. Aquat Microb Ecol 75:129–137. https://doi.org/10.3354/ame01753.
73. Beckers B, Op De Beeck M, Thijs S, Truyens S, Weyens N, Boerjan W, Vangronsveld J. 2016. Performance of 16s rDNA primer pairs in the study of rhizosphere and endosphere bacterial microbiomes in metabarcoding studies. Front Microbiol 7:650. https://doi.org/10.3389/fmicb.2016.00650.
74. Walters W, Hyde ER, Berg-Lyons D, Ackermann G, Humphrey G, Parada A, Gilbert JA, Jansson JK, Caporaso JG, Fuhrman JA, Apprill A, Knight R. 2016. Improved bacterial 16S rRNA gene (V4 and V4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. mSystems 1:e00009-15. https://doi.org/10.1128/mSystems.00009-15.
75. Marotz C, Sharma A, Humphrey G, Gottel N, Daum C, Gilbert JA, Eloe-Fadrosh E, Knight R. 2019. Triplicate PCR reactions for 16S rRNA gene amplicon sequencing are unnecessary. Biotechniques 67:29–32. https://doi.org/10.2144/btn-2018-0192.
76. Smith DP, Peay KG. 2014. Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. PLoS One 9:e90234. https://doi.org/10.1371/journal.pone.0090234.
77. Sze MA, Schloss PD. 2019. The impact of DNA polymerase and number of rounds of amplification in PCR on 16S rRNA gene sequence data. mSphere 4:e00163-19. https://doi.org/10.1128/mSphere.00163-19.
78. Fadrosh DW, Ma B, Gajer P, Sengamalay N, Ott S, Brotman RM, Ravel J. 2014. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. Microbiome 2:6. https://doi.org/10.1186/2049-2618-2-6.
79. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 79:5112–5120. https://doi.org/10.1128/AEM.01043-13.
80. de Muinck EJ, Trosvik P, Gilfillan GD, Hov JR, Sundaram AYM. 2017. A novel ultra high-throughput 16S rRNA gene amplicon sequencing library preparation method for the Illumina HiSeq platform. Microbiome 5:68. https://doi.org/10.1186/s40168-017-0279-1.
81. Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, Chan CKF, Nabhan AN, Su T, Morganti RM, Conley SD, Chaib H, Red-Horse K, Longaker MT, Snyder MP, Krasnow MA, Weissman IL. 9 April 2017. Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. bioRxiv https://doi.org/10.1101/125724.
82. van der Valk T, Vezzi F, Ormestad M, Dalén L, Guschanski K. 2020. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. Mol Ecol Resour 20:1171–1181. https://doi.org/10.1111/1755-0998.13009.
83. Wright ES, and, Vetsigian KH. 2016. Quality filtering of Illumina index reads mitigates sample cross-talk. BMC Genomics 17:876. https://doi.org/10.1186/s12864-016-3217-x.
84. Illumina. 2021. Minimize index hopping in multiplexed runs. https://www.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing/index-hopping.html. Accessed 21 June 2021.
85. Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res 40:e3. https://doi.org/10.1093/nar/gkr771.
86. Illumina. 2021. How much PhiX spike-in is recommended when sequencing low diversity libraries on Illumina platforms? https://support.illumina.com/bulletins/2017/02/how-much-phix-spike-in-is-recommended-when-sequencing-low-divers.html. Accessed 21 June 2021.

87. Holm JB, Humphrys MS, Robinson CK, Settles ML, Ott S, Fu L, Yang H, Gajer P, He X, McComb E, Gravitt PE, Ghanem KG, Brotman RM, Ravel J. 2019. Ultrahigh-throughput multiplexing and sequencing of >500-base-pair amplicon regions on the Illumina HiSeq 2500 platform. mSystems 4:e00029-19. https://doi.org/10.1128/mSystems.00029-19.
88. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379. https://doi.org/10.1371/journal.pone.0019379.
89. Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, Buerkle CA. 2012. Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. Evolution 66:2167–2181. https://doi.org/10.1111/j.1558-5646.2012.01587.x.
90. Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc 2010:pdb.prot5448. https://doi.org/10.1101/pdb.prot5448.
91. Caporaso JG, Ackermann G, Apprill A, Bauer M, Berg-Lyons D, Betley J, Fierer N, Fraser L, Fuhrman JA, Gilbert JA, Gormley N, Humphrey G, Huntley J, Jansson JK, Knight R, Lauber CL, Lozupone CA, McNally S, Needham DM, Owens SM, Parada AE, Parsons R, Smith G, Thompson LR, Thompson L, Turnbaugh PJ, Walters WA, Weber L. 2018. EMP 16S Illumina amplicon protocol. https://www.protocols.io/view/emp-16s-illumina-amplicon-protocol-nuudeww.
92. Katz ED, Haff LA, Eksteen R. 1990. Rapid separation, quantitation, and purification of products of polymerase chain reaction by liquid chromatography. J Chromatogr 512:433–444. https://doi.org/10.1016/S0021-9673(01)89509-1.
93. LaMontagne MG, Michel FC, Holden PA, Reddy CA. 2002. Evaluation of extraction and purification methods for obtaining PCR-amplifiable DNA from compost for microbial community analysis. J Microbiol Methods 49:255–264. https://doi.org/10.1016/S0167-7012(01)00377-3.
94. Quail MA, Gu Y, Swerdlow H, Mayho M. 2012. Evaluation and optimisation of preparative semi-automated electrophoresis systems for Illumina library preparation. Electrophoresis 33:3521–3528. https://doi.org/10.1002/elps.201200128.
95. Dugan KA, Lawrence HS, Hares DR, Fisher CL, Budowle B. 2002. An improved method for post-PCR purification for mtDNA sequence analysis. J Forensic Sci 47:811–818.
96. Bell JR. 2008. A simple way to treat PCR products prior to sequencing using ExoSAP-IT. Biotechniques 44:834. https://doi.org/10.2144/000112890.
97. Hanshew AS, Mason CJ, Raffa KF, Currie CR. 2013. Minimization of chloroplast contamination in 16S rRNA gene pyrosequencing of insect herbivore bacterial communities. J Microbiol Methods 95:149–155. https://doi.org/10.1016/j.mimet.2013.08.007.
98. Jackrel SL, Owens SM, Gilbert JA, Pfister CA. 2017. Identifying the plant-associated microbiome across aquatic and terrestrial environments: the effects of amplification method on taxa discovery. Mol Ecol Resour 17:931–942. https://doi.org/10.1111/1755-0998.12645.
99. Karasov TL, Neumann M, Duque-Jaramillo A, Kersten S, Bezrukov I, Schröppel B, Symeonidi E, Lundberg DS, Regalado J, Shirsekar G, Bergelson J, Weigel D. 8 April 2020. The relationship between microbial biomass and disease in the *Arabidopsis thaliana* phyllosphere. bioRxiv https://doi.org/10.1101/828814.
100. Hyrup B, Nielsen PE. 1996. Peptide nucleic acids (PNA): synthesis, properties and potential applications. Bioorg Med Chem 4:5–23. https://doi.org/10.1016/0968-0896(95)00171-9.
101. Lundberg DS, Yourstone S, Mieczkowski P, Jones CD, Dangl JL. 2013. Practical innovations for high-throughput amplicon sequencing. Nat Methods 10:999–1002. https://doi.org/10.1038/nmeth.2634.
102. Moccia K, Papoulis S, Willems A, Marion Z, Fordyce JA, Lebeis SL. 2020. Using the Microbiome Amplification Preference Tool (MAPT) to reveal Medicago sativa-associated eukaryotic microbes. Phytobiomes J 4:340–350. https://doi.org/10.1094/PBIOMES-02-20-0022-R.
103. Nielsen PE. 2001. Peptide nucleic acid: a versatile tool in genetic diagnostics and molecular biology. Curr Opin Biotechnol 12:16–20. https://doi.org/10.1016/S0958-1669(00)00170-1.
104. Stender H, Fiandaca M, Hyldig-Nielsen JJ, Coull J. 2002. PNA for rapid microbiology. J Microbiol Methods 48:1–17. https://doi.org/10.1016/S0167-7012(01)00340-2.
105. von Wintzingerode F, Landt O, Ehrlich A, Gobel UB. 2000. Peptide nucleic acid-mediated PCR clamping as a useful supplement in the determination of microbial diversity. Appl Environ Microbiol 66:549–557. https://doi.org/10.1128/AEM.66.2.549-557.2000.
106. Chandler DP, Stults JR, Cebula S, Schuck BL, Weaver DW, Anderson KK, Egholm M, Brockman FJ. 2000. Affinity purification of DNA and RNA from environmental samples with peptide nucleic acid clamps. Appl Environ Microbiol 66:3438–3445. https://doi.org/10.1128/AEM.66.8.3438-3445.2000.
107. Sakai M, Ikenaga M. 2013. Application of peptide nucleic acid (PNA)-PCR clamping technique to investigate the community structures of rhizobacteria associated with plant roots. J Microbiol Methods 92:281–288. https://doi.org/10.1016/j.mimet.2012.09.036.
108. Fitzpatrick CR, Lu-Irving P, Copeland J, Guttman DS, Wang PW, Baltrus DA, Dlugosch KM, Johnson MTJ. 2018. Chloroplast sequence variation and the efficacy of peptide nucleic acids for blocking host amplification in plant microbiome studies. Microbiome 6:144. https://doi.org/10.1186/s40168-018-0534-0.
109. Green SJ, Minz D. 2005. Suicide polymerase endonuclease restriction, a novel technique for enhancing PCR amplification of minor DNA templates. Appl Environ Microbiol 71:4721–4727. https://doi.org/10.1128/AEM.71.8.4721-4727.2005.
110. Dolinšek J, Dorninger C, Lagkouvardos I, Wagner M, Daims H. 2013. Depletion of unwanted nucleic acid templates by selective cleavage: LNAzymes, catalytically active oligonucleotides containing locked nucleic acids, open a new window for detecting rare microbial community members. Appl Environ Microbiol 79:1534–1544. https://doi.org/10.1128/AEM.03392-12.
111. Feehery GR, Stewart F, Mcfarland J, Pradhan S. March 2015. Methods and compositions for enriching either target polynucleotides or non-target polynucleotides from a mixture of target and non-target polynucleotides. Patent US8980553B2.
112. Gebhard C, Schwarzfischer L, Pham TH, Andreesen R, Mackensen A, Rehli M. 2006. Rapid and sensitive detection of CpG-methylation using methylbinding (MB)-PCR. Nucleic Acids Res 34:e82. https://doi.org/10.1093/nar/gkl437.
113. Yigit E, Hernandez DI, Trujillo JT, Dimalanta E, Bailey CD. 2014. Genome and metagenome sequencing: using the human methyl-binding domain to partition genomic DNA derived from plant tissues. Appl Plant Sci 2:1400064. https://doi.org/10.3732/apps.1400064.
114. Mariac C, Scarcelli N, Pouzadou J, Barnaud A, Billot C, Faye A, Kougbeadjo A, Maillol V, Martin G, Sabot F, Santoni S, Vigouroux Y, Couvreur TLP. 2014. Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. Mol Ecol Resour 14:1103–1113. https://doi.org/10.1111/1755-0998.12258.
115. Thoendel M, Jeraldo PR, Greenwood-Quaintance KE, Yao JZ, Chia N, Hanssen AD, Abdel MP, Patel R. 2016. Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. J Microbiol Methods 127:141–145. https://doi.org/10.1016/j.mimet.2016.05.022.
116. Handschur M, Karlic H, Hertel C, Pfeilstöcker M, Haslberger AG. 2009. Preanalytic removal of human DNA eliminates false signals in general 16S rDNA PCR monitoring of bacterial pathogens in blood. Comp Immunol Microbiol Infect Dis 32:207–219. https://doi.org/10.1016/j.cimid.2007.10.005.
117. Horz H-P, Scheer S, Vianna ME, Conrads G. 2010. New methods for selective isolation of bacterial DNA from human clinical specimens. Anaerobe 16:47–53. https://doi.org/10.1016/j.anaerobe.2009.04.009.
118. Votintseva AA, Pankhurst LJ, Anson LW, Morgan MR, Gascoyne-Binzi D, Walker TM, Quan TP, Wyllie DH, Del Ojo Elias C, Wilcox M, Walker AS, Peto TEA, Crook DW. 2015. Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. J Clin Microbiol 53:1137–1143. https://doi.org/10.1128/JCM.03073-14.
119. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. 2018. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome 6:42. https://doi.org/10.1186/s40168-018-0426-3.
120. Wu Z, Willing B, Bjerketorp J, Jansson JK, Hjort K. 2009. Soft inertial microfluidics for high throughput separation of bacteria from human blood cells. Lab Chip 9:1193–1199. https://doi.org/10.1039/b817611f.
121. Shapiro HM. 2005. Practical flow cytometry. John Wiley & Sons, Hoboken, NJ.
122. Darzynkiewicz Z, Huang X, Zhao H. 2017. Analysis of cellular DNA content by flow cytometry. Curr Protoc Immunol 119:5.7.1–5.7.20. https://doi.org/10.1002/cpim.36.
123. Lebaron P, Servais P, Agogué H, Courties C, Joux F. 2001. Does the high nucleic acid content of individual bacterial cells allow us to discriminate between active cells and inactive cells in aquatic systems? Appl Environ Microbiol 67:1775–1782. https://doi.org/10.1128/AEM.67.4.1775-1782.2001.
124. Rubbens P, Schmidt ML, Props R, Biddanda BA, Boon N, Waegeman W, Denef VJ. 2019. Randomized lasso links microbial taxa with aquatic

functional groups inferred from flow cytometry. mSystems 4:e00093-19. https://doi.org/10.1128/mSystems.00093-19.
125. Utturkar SM, Cude WN, Robeson MS, Yang ZK, Klingeman DM, Land ML, Allman SL, Lu T-YS, Brown SD, Schadt CW, Podar M, Doktycz MJ, Pelletier DA. 2016. Enrichment of root endophytic bacteria from Populus del- toides and single-cell-genomics analysis. Appl Environ Microbiol 82:5698–5708. https://doi.org/10.1128/AEM.01285-16.
126. Ramond J-B, Makhalanyane TP, Tuffin MI, Cowan DA. 2015. Normalization of environmental metagenomic DNA enhances the discovery of under-represented microbial community members. Lett Appl Microbiol 60:359–366. https://doi.org/10.1111/lam.12380.
127. Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, Matz MV, Meleshkevitch E, Moroz LL, Lukyanov SA, Shagin DA. 2004. Simple cDNA normalization using Kamchatka crab duplex-specific nuclease. Nucleic Acids Res 32:e37. https://doi.org/10.1093/nar/gnh031.
128. Shade A, Handelsman J. 2012. Beyond the Venn diagram: the hunt for a core microbiome. Environ Microbiol 14:4–12. https://doi.org/10.1111/j.1462-2920.2011.02585.x.