






## REVIEW

# Pitfalls and pointers: An accessible guide to marker gene amplicon sequencing in ecological applications

Anita Porath-Krause  | Alexander T. Strauss  | Jeremiah A. Henning  |  
Eric W. Seabloom  | Elizabeth T. Borer 

Department of Ecology, Evolution, and Behavior, University of Minnesota, St. Paul, MN, USA

**Correspondence**

Anita Porath-Krause  
Email: [aporathk@umn.edu](mailto:aporathk@umn.edu)

**Present address**

Alexander T. Strauss, Odum School of Ecology, University of Georgia, Athens, GA, USA

Jeremiah A. Henning, Department of Biology, University of South Alabama, Mobile, AL, USA

**Funding information**

Division of Environmental Biology, Grant/Award Number: 1241895

**Handling Editor:** Kate Quigley

**Abstract**

1. Next-Generation Sequencing (NGS) is a powerful tool that has been rapidly adopted by many ecologists studying microbial communities. Despite the exciting demonstration of NGS technology as a tool for ecological research, cryptic pitfalls inherent to its use can obscure correct interpretation of NGS data. Here, we provide an accessible overview of a NGS process that uses marker gene amplicon sequences (MGAS) that will allow scientists, particularly community ecologists, to make appropriate methodological choices and understand limits on inference about community composition and diversity that can be drawn from MGAS data.
2. We describe the MGAS pipeline, focusing specifically on cryptic sources of variation that have received less emphasis in the ecological literature, but which may substantially impact inference about microbial community diversity and composition. By simulating communities from published microbiome data, we demonstrate how these sources of variation can generate inaccurate or misleading patterns.
3. We specifically highlight sample dilution without researcher awareness and lane-to-lane variability, two cryptic sources of variation arising during the MGAS pipeline. These sources of variation affect estimates of species presence and relative abundance, particularly for species with moderate to low abundances. Each of these sources of bias can lead to errors in the estimation of both absolute and relative abundance within, and turnover among, microbial communities.
4. Awareness and understanding of what happens and, specifically, why it happens during MGAS generation is key to generating a strong dataset and building a robust community matrix. Requesting sample dilution information from the sequencing centre, including technical replicates across sequencing lanes, and understanding how sampling intensity and community taxa distribution patterns shape the measurement of community richness, evenness and diversity are critical for drawing correct ecological inferences using MGAS data.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

**KEY WORDS**

community ecology, community matrix, ecology, lane variability, next-generation sequencing, sample dilution

**1 | INTRODUCTION**

Next-generation sequencing (NGS) is transforming our understanding of biology. In the past few decades, the concept of an 'individual' has shifted from a single organism into a host individual inhabited by a community of dozens to thousands of different species of microbes—the holobiont (Margulis, 1981)—which contribute in significant ways to the function of the 'individual' (Turnbaugh et al., 2007; Turner et al., 2013). By quickly generating millions of data points, NGS has massively accelerated the potential for discovery. In its infancy, NGS primarily identified genes to explain phenotypes of individuals. NGS was a tool to collect and identify whole genomes (mainly from humans and bacteria; van Dijk et al., 2014), or identify and quantify transcripts (Croucher et al., 2009; Gibbons et al., 2009) to ask questions such as, 'How does the genome shape the organism?' The use of NGS methods has moved beyond the original application of genome sequencing and has now enabled fields such as transcriptomics and epigenomics (Davey et al., 2011; Schuster, 2008). NGS technologies complement, and in some instances have replaced, traditional methods used in genetics (Mardis, 2008). Moving at an exponential pace, the introduction of new NGS technologies allows us to address new questions, or long-standing questions in novel ways (Black et al., 2015; Morozova & Marra, 2008).

If they have not done so already, many ecologists are likely to read or review papers and grant proposals that incorporate data employing NGS, and an increasing number of ecologists will be involved in projects using these methods. There are two main approaches used to study microbial communities, marker gene amplicon sequencing (MGAS) and shotgun metagenomic sequencing, the former being more commonly used to describe microbial communities in ecology (Gonzalez et al., 2012). The potential applications of MGAS to community ecology followed on the heels of original applications (Claesson et al., 2010; Shokralla et al., 2012; Wirth et al., 2012), leading to the widespread recognition of the 'microbiome', or diverse microbial communities and their ensemble of activity (Berg et al., 2020) that are important for ecological functioning. For example, microbiomes regulate plant growth and contribute to soil health (Saleem et al., 2019), control sponge tolerance to ocean acidification (Ribes et al., 2016) and modify eelgrass sediment to be a nitrous oxide sink (Nakagawa et al., 2018). The use of MGAS as a tool in ecology is not limited to microbial communities but has become widespread and implemented in research such as biomonitoring (Derocles et al., 2018) and comprehensive spatial-temporal studies (Sanz & Köchling, 2019). Because of the superficial similarity of MGAS data to more traditional ecological approaches—with quantitative data on the number and relative abundance of each taxon or OTU—MGAS data are being increasingly analysed and interpreted with community metrics that were developed using free-living ecological communities, such as species richness, evenness and turnover.

**GLOSSARY**

**Template:** The available RNA or DNA collected from a sample.

**Target:** A specific RNA or DNA region of interest, isolated using primers.

**Barcode:** A unique string of nucleotides that is bound to all sequences within a sample so sample identity can be tracked. This allows multiple samples to be run simultaneously.

**Adapter:** This commonly refers to a short set of nucleotides that contains both the barcode and the connection point for the sequence primer.

**Library:** A grouping of RNA or DNA fragments, typically an aggregation of multiple samples, to which specific adapters have been added so that the fragments can adhere to the sequencing machine.

**Multiplexing:** Combining multiple libraries containing bar-coded sequences onto a single sequencing run.

**Sequencing platform:** The wet-lab technical approach to produce next-generation sequences—typically involves a specific set of reagents and mechanically unique machinery.

**Depth:** The number of sequences with nucleotide bases aligned to the region of interest.

**Cluster:** A group of identical sequence fragments, cloned from a single template, that produce a sequence read.

Despite the exciting demonstration of MGAS technology as a tool for generating new insights in ecological research, there are pitfalls inherent in the use of MGAS to measure community composition and diversity. These pitfalls, if unknown to the user, could bias or even undermine our understanding of the factors shaping ecological communities. At the core of many community ecology analyses is a community matrix quantifying the abundance or presence of different taxa in different samples (Pielou, 1984). As with any measurement technique, MGAS is inherently non-random and may produce biased estimates (Medinger et al., 2010) of the underlying distribution of abundances that form this community matrix, such as exclusion of rare species or over-representation of certain clades (Hugerth & Andersson, 2017; Kelly et al., 2019; McLaren et al., 2019; Taylor et al., 2002; Ye et al., 2019). However, because many of the potential biases occur during sample processing, often at a separate sequencing facility, these biases may be unknown to investigators who are trained in ecology but who may be less knowledgeable about the details of MGAS. Because general standardization is lacking in most MGAS methods (Gohl, 2017), and like any sampling method (e.g. insect pitfall traps, intertidal settlement plates

or satellite imagery, to name a few), a basic understanding of the strengths and weaknesses of the sampling technique is critical for understanding how to analyse and interpret the resulting data.

The abundance and distribution of species are of fundamental interest in ecology (Hutchinson, 1961; Vellend, 2010). For microbes, understanding how communities are assembled and which forces lead to species turnover also promise exciting insights into the role of microbial diversity and composition in host health and ecosystem functioning (Borer et al., 2013; Caporaso et al., 2011; Christian et al., 2015; Fierer & Jackson, 2006; Van Der Heijden et al., 2008). While absolute abundance of species is often of great interest, estimates of relative abundance are often more immediately achievable and still promise substantial insights (e.g. changes in relative abundance along an environmental gradient). However, quantifying species diversity patterns for any group has significant challenges. Strong inference about these processes only arises from identical sampling across imposed and observed environmental gradients. This inference is weakened or fully undermined when differences or errors in sampling or processing methods cause the data to reflect anything other than a consistent, directly comparable measure of a community in space or time (Chase & Knight, 2013; Ma et al., 2019; Morgan et al., 2013; Schirmer et al., 2015; Smith & Peay, 2014).

In spite of the importance of understanding the strengths and weaknesses of MGAS for advancing ecological knowledge, much of the MGAS methodological literature has been published in technical journals (e.g. Buehler et al., 2010; Manley et al., 2016; Quail et al., 2008; Robin et al., 2016) but see (Caporaso et al., 2012; Lindahl et al., 2013; Smith & Peay, 2014). While this technical focus is critical for rapid advances in MGAS technology, this literature can be impenetrable for researchers outside of this area who seek to use this powerful tool. Our goal is to provide an accessible overview of the MGAS process that will allow scientists, particularly community ecologists, to make appropriate methodological choices and understand limits on inferences about communities that can be drawn from MGAS data. We start by describing the steps that occur from the submission of a sample to an MGAS facility using the most common *sequencing platform* (i.e. Illumina; van Dijk et al., 2014), and walk through the steps leading to the production of the community matrix. We then discuss how decisions at each step in this pipeline may lead to bias or variation in the community matrix that could lead to incorrect inferences about microbial communities. We concentrate on metrics commonly used to address core questions in community ecology, including community richness, community diversity and community composition or differences in taxa identity between communities. We then sample published NGS data (Seabloom et al., 2019) to demonstrate how technical choices can ultimately introduce bias in the community matrix. Finally, we recommend ways to reduce controllable variation when collecting, processing, analysing and interpreting MGAS data. While we cannot provide detailed coverage of all issues associated with the application of NGS to community ecology, our goal here is to provide an overview of topics important for the use of MGAS data to make the study of microbial community composition more accessible to community ecologists. Thus, we provide a primer so that the reader will be a more informed builder of community

matrices using data collected by MGAS and a more critical reader of scientific work based on these data.

## 2 | A CONCEPTUAL OVERVIEW OF BUILDING THE COMMUNITY MATRIX: FROM SAMPLE TO OTU

For applications that target a single organism, a narrow group of organisms or gene expression profiling, MGAS is an excellent approach. Accordingly, MGAS has nearly supplanted Sanger Sequencing (Sanger et al., 1977), because the MGAS approach sequences in a manner that is massively parallel, meaning multiple samples can be sequenced at the same time, reducing sequencing cost and time to process. Since its inception, MGAS has been thoroughly studied and improved to achieve a high-quality product via reducing sequencing errors, improving DNA quantification, reducing bias and increasing high-throughput reliability (Gohl et al., 2016; Manley et al., 2016; Quail et al., 2008; Robin et al., 2016; van Dijk et al., 2014). However, these improvements are typically implemented by the sequencing facility, and many users may be unaware of the implications of these changes. Users can also make many choices when utilizing MGAS with their particular organism(s) of interest including, but not limited to, *library* preparation, *sequencing platform*, *depth* of reads, sample extraction kits and even bioinformatics decisions (a Glossary is provided with definitions of italicized terms). These important sources of variation have been studied, described and compared in detail to aid the researcher in forming an informed decision (Abel & Duncavage, 2013; D'Amore et al., 2016; Gohl et al., 2016; Gołębiewski & Tretyn, 2019; Robin et al., 2016; Roy et al., 2018; Sims et al., 2014). Knowledge of these sources of variation and how they may impact inference is critical for effective use of MGAS data to compare microbial community matrices. While certain technical considerations causing MGAS variation have been thoroughly reviewed (e.g. *library* preparation, *sequencing platform*, *read depth* and reagent choice (Goodwin et al., 2016; Levy & Myers, 2016; Slatko et al., 2018; van Dijk et al., 2014)), we use this paper to unveil sources of variation that have received less emphasis in past descriptions and which may substantially impact inference about microbial community diversity and composition. Figure 1 highlights the steps involved in the MGAS pipeline, from sample to community matrix, particularly highlighting key steps in the process where MGAS methodology can bias the community matrix.

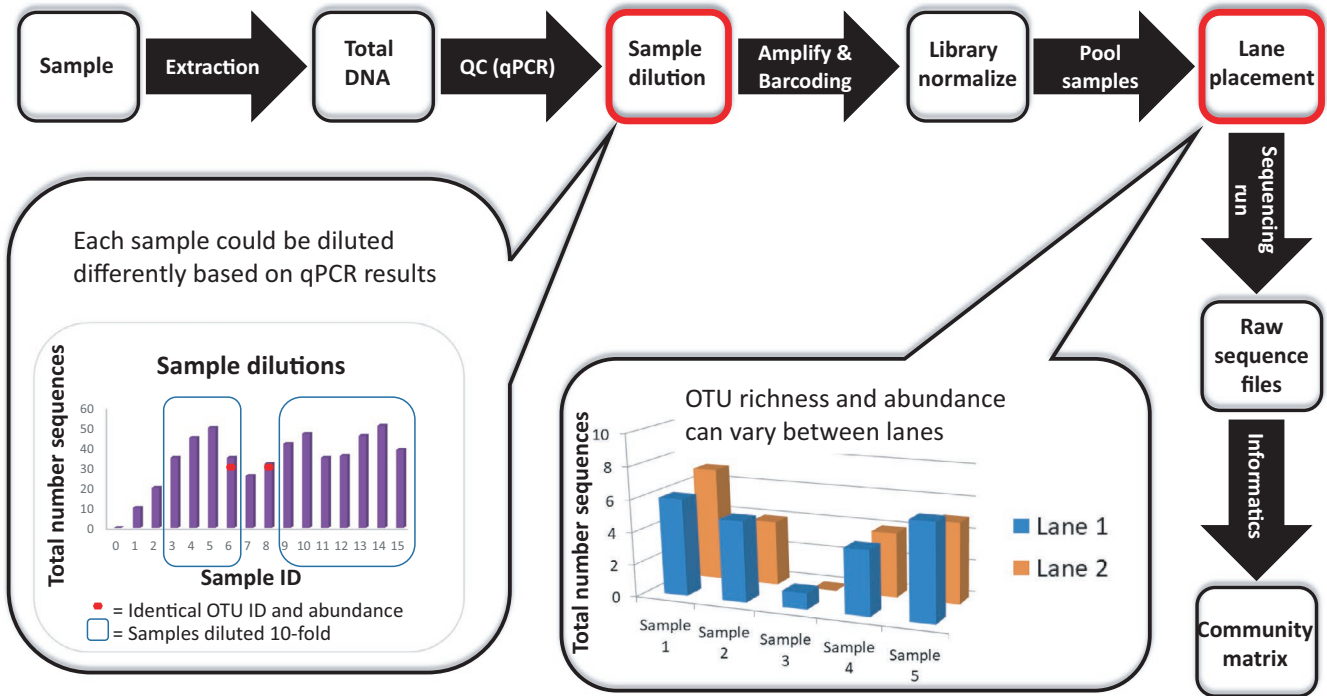
## 3 | SOURCES OF VARIATION

### 3.1 | Dilution of samples based on qPCR results

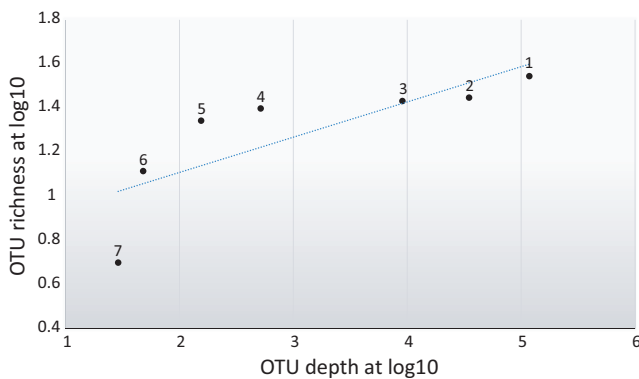
#### 3.1.1 | What happens?

Once DNA is extracted, samples typically undergo a quality control process in which either the DNA concentration is measured using spectrophotometric or fluorescent methods or quantitative PCR

### Marker Gene Amplicon Sequencing Pipeline



**FIGURE 1** Marker Gene Amplicon Sequencing Pipeline. Beginning with the collection of the sample as indicated on the upper left side of figure and moving clockwise to the lower right of the figure to end with the community matrix, the boxes and arrows represent the steps involved in the marker gene amplicon sequences (MGAS) pipeline. Red bolded boxes highlight key points in the process where MGAS methodology may influence the community matrix and are considered areas of concern. Dialogue boxes detail examples of these two areas of concern



**FIGURE 2** Sample dilution non-uniformly decreases depth and richness. 16S sequence depth and richness shown here using a log<sub>10</sub> scale. An identical sample was diluted in 10-fold series which is presented as 1 through 7 representing log<sub>10</sub>(dilution). A linear trendline is represented by the blue line. Data were generated by Gohl et al. (2016) using a mock community and library amplification methods assigned as KAPA-25-HM-276D

(qPCR) is used to determine the initial amount of the *target* present per sample. Many workflows then involve diluting the input sample to normalize the volume, mass of DNA or amount of *target* molecule that is added to the amplification reaction (Figure 1, Area of

Concern). Samples are then *barcoded*, and the sequencing libraries are normalized prior to pooling and simultaneous sequencing (i.e. *multiplexing*). This normalization step allows the even allocation of sequencing reads between the different libraries. Because samples within and among runs can be diluted different amounts (both prior to amplification and during subsequent *library* normalization steps), the resulting data quantifying individual operational taxonomic units (OTUs) may be nonlinear and care should be taken to account for these dilutions when metrics that are sensitive to sampling intensity are being compared (Gohl et al., 2016; Figure 2).

For example, two samples, 6 and 8, as portrayed in Figure 1 in dialogue bubble 1, both contain an identical fungal ITS1 OTU, A (indicated by red dot). While the DNA marker that characterizes A may have a similar number of copies in both samples 6 and 8, the total number of ITS1 OTU sequences (A + B + C) in sample 6 is greater than sample 8, so sample 6 is diluted 10-fold more than sample 8. If the researcher is unaware the samples have been unequally diluted, any estimates of abundance (i.e. copy number) and diversity that do not account for dilution will skew comparisons of abundance among samples. Since it has been shown that dilution can lead to under-sampling of microbial communities (Castle et al., 2018), it is critical that the researcher understands the upstream processing steps, particularly any dilutions that were carried out, which could influence diversity measures (see *Why do we care?*, below).

## 3.2 | Lane-to-lane variability

### 3.2.1 | What happens?

Samples are loaded onto a sequencing machine lane. A typical sequencer has up to eight lanes depending on the *sequencing platform* (Illumina, 2017). The level of sample *multiplexing* is typically determined by the desired read *depth* (or sampling effort), so it is a common practice to place up to 300 samples on a lane. MGAS data are commonly generated with the Illumina MiSeq *sequencing platform* which uses a single lane; however, experiments comprised of large sample sizes (e.g. greater than 300 samples) are divided across more than one MiSeq run, meaning more than one lane. A number of sequencer metrics may be collected but are not commonly reported with MGAS data, including overall read *depth*, sequencing quality score, amount of spike-in control sample and other run-to-run biases (Gohl et al., 2019); we focus on one concern that is relatively cryptic to the typical user. In particular, lane-to-lane variation, meaning identical samples sequenced on different runs (Figure 1, dialogue bubble 2), can vary in *cluster* density or other sequencing parameters between lanes (Elshire et al., 2011; Quail et al., 2008). In addition, samples may be present in unequal amounts in different runs (Gloor et al., 2017). Unequal sample balance within the *library* arises during sample quantification. Quantifying the concentration of available *template* does not necessarily yield the amount of *detectable* material for sequencing. A sample will contain *template* that does not have *adapters*, has only one *adapter* or contains artefacts like *adapter* dimers, causing some of the material in the sample not to be detected for sequencing (Illumina, 2016). It is possible to account for these factors by quantifying libraries with qPCR or ddPCR (Laurie et al., 2013), but these approaches may not be affordable with a large number of samples. Physically manipulating *cluster* density on the lane (flow cell) is another source of variation because accurately quantifying and diluting libraries with the degree of precision required to yield perfectly consistent results is nearly impossible. Libraries are diluted to picomolar concentrations, so any inconsistency with concentration quantification or *library* dilution that yields less than 100% accuracy can potentially create a run with low output, or equally likely, a run that *over-clusters* (Quail et al., 2008). Either possibility causes variation that will impact the final community matrices from each sample.

## 3.3 | How can diluting samples or lane variability change the community profile?

### 3.3.1 | Why do we care?

For microbial sequencing, as for sampling of other communities of species, variation in sample sizes can undermine effective comparison among samples, because more intensive sampling of a community generally increases the total number of species found by increasing the likelihood of discovering rare species (Longino

et al., 2002), analogous to a collector's curve (Mao et al., 2005). The severity of this problem increases with skew of the community's rank abundance distribution (Figure 1). For example, when sequence abundance varies widely among lanes, given the same sampling effort (or *depth* of sequencing), the intensity of sampling will vary inversely with abundance, introducing bias in the characterization of each community's taxonomic richness. Even when a highly diverse community is intensively sampled, rarer species can remain unobserved, resulting in substantial sampling bias because of extreme under-sampling (Lande, 1996). For this reason, in identical samples, lane-to-lane variation in subsampling, analogous to observation error in field studies, can lead to different estimates of species presence and relative abundance, particularly for species with moderate to low abundances. Each of these sources of bias can lead to errors in the estimation of both absolute and relative abundance within, and turnover among, communities (Lande, 1996).

Importantly, different metrics of diversity convey different information about communities. Some metrics of diversity, for example,  $ENS_{PIE}$ , are relatively insensitive to these sources of bias, and are therefore recommended when comparing multiple communities that are sampled at different intensities (Chase & Knight, 2013). On the other hand, the simplest metric of diversity—species richness—is one of the most sensitive to differences in sampling effort. Statistical methods including abundance-based rarefaction can enable comparisons across samples, but this approach sacrifices data (McMurdie & Holmes, 2014). Proportion-based rarefactions are also possible and may be more appropriate when comparing communities of very different sizes (Chao & Jost, 2012), but this approach still suffers from the same limitation (loss of data from the more thoroughly sampled community). All of these metrics of diversity convey unique information about the community.

## 3.4 | OTU delineation methods—An area of interest

Once samples have been through the sequencing process and raw sequence files are generated, the raw sequences are subjected to a bioinformatics workflow. This workflow removes non-*target* sequences (this description is simplified—for detailed information, see Caporaso et al., 2010; Navas-Molina et al., 2013) and then sequences are grouped by similarity to delineate OTUs. To avoid ambiguity, we use the term 'group', but other authors will frequently use the term 'cluster' when referring to OTU delineation based on sequence similarity. Finally, the grouped sequences are matched with a known phylogenetic group to assign a biological identity. While this step in the pipeline is effective in reducing OTU identity error and assigning a biological identity, the specific choices made to delineate and group sequences can create variation invisible to a researcher. For community ecologists comparing individuals across a community, the preferred approach to delineate OTUs is open-reference OTU picking (Rideout et al., 2014). Even though the variation that can arise in this step is well described (Rideout et al., 2014; Schloss, 2016; Zhang et al., 2013), it is still commonly overlooked.

OTU delineation methods and assigning the degree of similarity when grouping sequences is particularly important for ecologists to consider when constructing a community matrix because the matrix affects diversity metric calculations in downstream analyses. For drawing inference about microbial communities, it is critical for ecologists to use an informed strategy to delineate and group the sequences.

A 97% similarity among sequences has become commonplace in most MGAS pipelines for OTU delineation. When using MGAS, bacterial species, for example, are commonly demarcated based on 16s rRNA similarities due to the highly conserved *target* region. To evaluate approaches used to demarcate bacterial species, Stackebrandt and Goebel (1994) compared the original 'gold-standard' method to sequence homology methods. These authors demonstrated that bacterial species with 97.5% or greater 16s rRNA sequence homology will likely have more than 70% DNA similarity and are therefore assumed to be related at the species level. In this foundational paper written 25 years ago, Stackebrandt and Goebel warned, however, that omission of portions of a gene can obscure the reliability of phylogenetic comparisons, necessitating inclusion of the entire gene and all hypervariable regions in species demarcation. Due to cost, time and sequencing *depth* limitations in MGAS, most scientists reduce the 16s rRNA region to just a portion of the gene to demarcate species (Martínez-Porchas et al., 2016; Schloss & Handelsman, 2006; Yang et al., 2016, but see Weirather et al., 2017). Per Stackebrandt and Goebel's recommendation, if a smaller region of the 16s rRNA is required, it is critical to select regions that will yield the same degree of similarity as full-length sequences to create reliable microbial species matrices.

Delineation methods have been proposed that provide alternatives to the 97% similarity approach. For example, it has been argued that an evolutionary criterion (general mixed Yule-coalescence) is more successful than OTU delineation for grouping certain types of organisms (e.g. fungi; Powell et al., 2011); although Lekberg et al. (2014) show there may be no difference between 97% sequence homology and an evolutionary criterion. Some researchers even argue OTU percent sequence grouping may not matter when describing the compositional differences (beta diversity) of microbial communities (Botnen et al., 2018) or broad ecological patterns (Glassman & Martiny, 2018).

Recently, the use of exact sequence variants, also called amplicon sequence variants (ASV), has gained popularity (Callahan et al., 2017). It is argued that by restricting sequence similarity to 100%, the precise nature of ASV improve MGAS reusability across studies and reproducibility in future datasets (Callahan et al., 2017). While it is suggested that the use of ASV captures the true biological presence of an organism, there are concerns that false sequences, generated during sequencing, can be unintentionally incorporated into the dataset and present themselves as rare OTUs, thus artificially inflating species diversity using the ASV method (Huse et al., 2010). Additionally, when comparing ASV data across studies, researchers need to be aware of variation in MGAS protocols

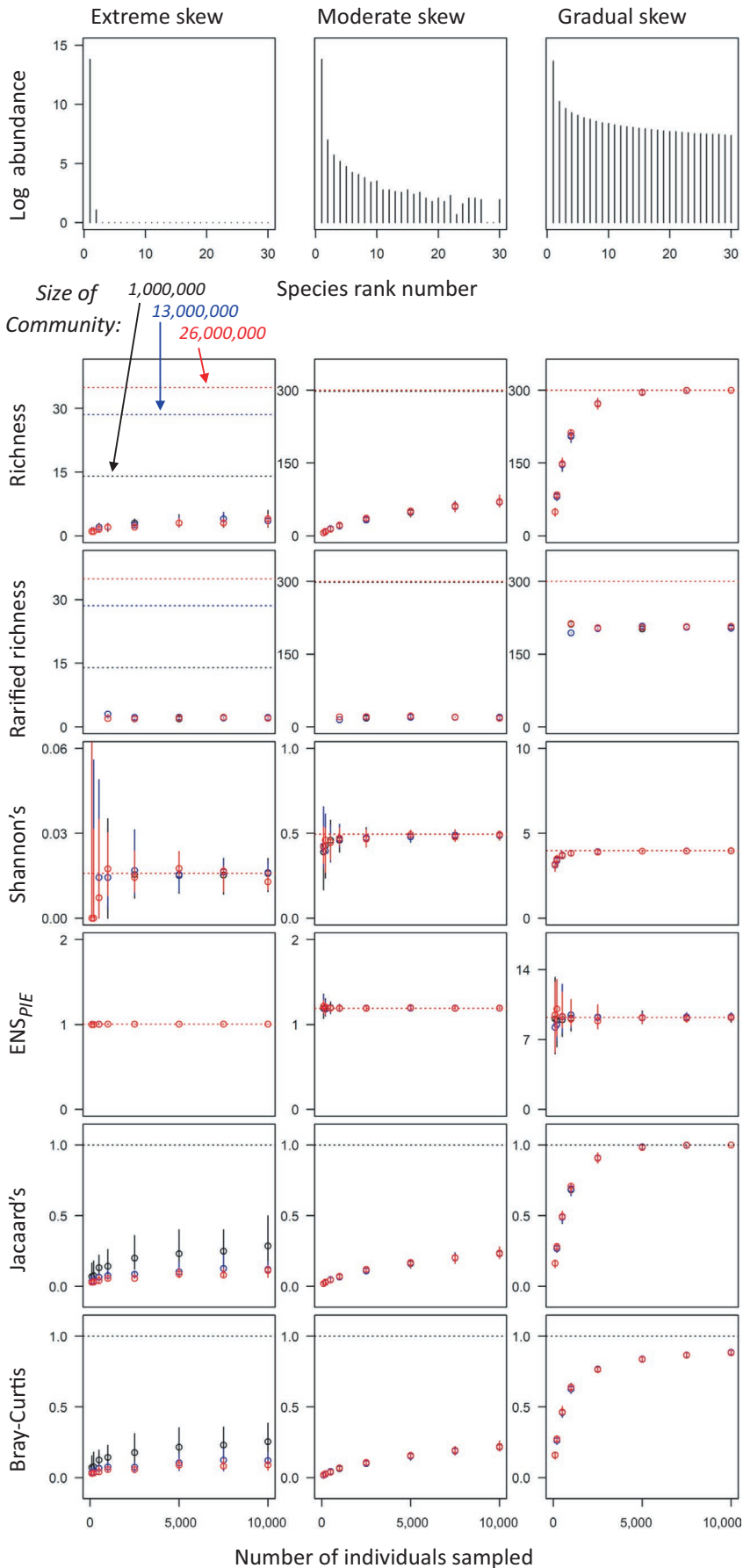
because small differences in the methods can lead to inconsistency in data generation (Gohl et al., 2016).

Delineating species and assigning sequences are currently a somewhat controversial subject but remains an important choice for any researcher seeking to generate and interpret a community matrix using MGAS data (Botnen et al., 2018; Callahan et al., 2017; Huse et al., 2010). Deciding which approach is best for the data and question at hand should be well-informed, utilizing all the resources and arguments made by previous work (including but not limited to the work cited above). Building a robust community matrix is clearly critical for insuring strong inference from these data.

## 4 | SIMULATING DIFFERENT COMMUNITIES: METHODS AND OUTCOMES

Comparisons within and across communities commonly involve measures of diversity. With the awareness that diluting samples or placing samples on different lanes can exclude rare taxa and misrepresent species abundance and presence, building and working with the community matrix generated by MGAS data raises downstream concerns. Here we demonstrate how the community matrix can affect measures of diversity depending on the types of communities (skew of their rank-abundance distributions) and sampling effort (i.e. *depth* of sequencing). Using published MGAS data, we evaluate the bias arising from sampling communities with different evenness (i.e. skew in rank-abundance distribution) at different intensities. Utilizing ITS1 sequence data generated with Illumina MiSeq that represent foliar fungal endophyte community samples from about 120 prairie grass plants across four sites in the north central United States (Seabloom et al., 2019), we fit rank-abundance distributions to the OTU's using the *VEGAN* package (Oksanen et al., 2008) in R version 3.5.2 (R Core Team, 2013). We then simulated communities with a skew to match the empirical data (Figure 3, centre column), doubling the skew parameter (left column) or halving the skew (right column) to generate three versions of skew that are centred around empirical data. These examples (Figure 3) represent typical communities ranging from high evenness, such as dispersed plant litter (Albright & Martiny, 2018), to very low evenness (highly skewed), such as the gut microbiome of human infants (Pannaraj et al., 2017). All simulated communities contained 2,753 species, which was identified as the number of species in the empirical data.

We simulated communities with multinomial distributions of 2,753 categories (representing the number of unique species from the empirical data), with probabilities of each category drawn from the three rank abundance distributions. For each skew scenario, we simulated communities of 1 million, 13 million or 26 million OTUs, to represent a range of microbial community sizes identified across natural systems and in mock communities (from aphids (Jousselin et al., 2016), plants (Seabloom et al., 2019) and mock communities (Bokulich et al., 2013), respectively). Then we subsampled each simulated 'parent' community by randomly drawing 100, 200, 500,



**FIGURE 3** Simulating communities at different sampling intensities. Communities simulated to estimate Richness (second row), Rarefied Richness (third row) and common diversity indices including Shannon's (fourth row),  $ENS_{PIE}$  (Inverse Simpson's; fifth row), Jacaard's (sixth row) and Bray-Curtis (seventh row) at extreme (column 1), moderate (column 2) and gradual (column 3) sampling intensities. A Zipf-Mandelbrot model (Wilson 1991) was used to estimate skew of empirical data (for moderate skew; column 2), with more (column 1) or less skew (column 3) simulated by halving or doubling the skew parameter ( $\gamma$ ). Communities contain 1 million (black circle), 13 million (blue circle) or 26 million (red circle) individuals. Horizontal lines indicate univariate diversity metrics of the full parent communities (rows 2–5) or compositional similarity between the sample and the parent community (rows 6–7; 1 M = black; 13 M = blue; 26 M = red). Error bars are 95% confidence intervals around the median of simulations. All communities contain 2,753 species

1,000, 2,500, 5,000, 7,500 or 10,000 individuals (representing potential sample dilution and the depth of sequencing). For both the simulated parent communities (horizontal dashed lines in Figure 3) and the subsampled communities (points with SE), we calculated richness, abundance-rarefied richness (rarefied to 1,000 individuals) and common measures of alpha diversity including Shannon's diversity index (Shannon, 1948) and inverse Simpson's diversity (Simpson, 1949), also known as the Effective Number of Species based on the Probability of Interspecific Encounter ( $ENS_{PIE}$ ; Chase & Knight, 2013). We also calculated the compositional similarity between the sample and the parent (full) community using abundance-based Bray–Curtis (Bray & Curtis, 1957) and incidence-based Jaccard (Jaccard, 1901) similarities. The simulations were replicated 100x, 20x and 10x for the communities of 1 million, 13 million and 26 million individuals, respectively. Full simulation data and R code can be found on the Dryad Digital Repository under <https://doi.org/10.5061/dryad.fxpnvx0r3>.

Composite measures of alpha diversity that include both richness and evenness (i.e.  $ENS_{PIE}$  and Shannon's metric) were relatively robust to differences in skewness (Figure 3, fourth and fifth rows). Thus, these simulations demonstrate that  $ENS_{PIE}$  and Shannon's metric are relatively robust to even a large arbitrary dilution of a microbial sample. Both measures stabilized at the level of the simulated parent community even with less than 1% of the community sampled. Richness estimates, on the other hand, were highly sensitive to sampling effort (here, *depth* of sequencing), and, over the range of *depth* we considered, only accurately represented richness of the parent community with gradual skew (Figure 3, second row). The sensitivity of richness, but not  $ENS_{PIE}$  or Shannon, to sampling effort is well known in community ecology (Chase & Knight, 2013). Rarefied richness (third row) was relatively consistent across sampling intensities, but chronically underrepresented true richness. We note that diversity metrics, such as Shannon or  $ENS_{PIE}$ , combine information on both the number of taxa present (i.e. richness) and the distribution of abundances among species (i.e. evenness). For example,  $ENS_{PIE}$  is the product of species richness ( $S$ ) and Simpson's Evenness ( $E$ ):  $ENS_{PIE} = S \cdot E$ . Note that  $ENS_{PIE}$  and richness are identical when all species are equally abundant ( $E = 1$ ) but becomes less than richness as evenness declines, because rare species carry less influence (Chase & Knight, 2013). Bray–Curtis and Jaccard metrics suffered from a similar issue as richness, with the sample only converging with the composition of the parent community when the community had a gradual skew (over the range of sampling effort we considered; Figure 3, sixth and seventh rows). Rarefaction can be used to control for some of the biases introduced into richness or compositional estimates by unequal sampling effort (e.g. number of reads). However, rarefaction can introduce its own biases (Chao & Jost, 2012; Lu & Tian, 2017; McMurdie & Holmes, 2014) and careful thought is needed in applying this approach. In the original analysis of the focal dataset we use here, Seabloom et al. (2019) used a rarefied community matrix to calculate community distances and also found that  $ENS_{PIE}$  and rarefied richness were positively correlated ( $r = 0.64$ ). Perhaps most importantly, sample dilution could result

in communities that share little evenness to the parent community from which they were derived.

## 5 | RECOMMENDATIONS TO BUILD A ROBUST COMMUNITY MATRIX

### 5.1 | Know your dilutions

Sample dilution is a standard approach to optimize *target* detection in MGAS. While this approach will increase the *depth* of sequencing, it is also a cryptic source of variation as different samples may be unequally diluted in the sequencing centre (Figure 1). Requesting data on which samples have been diluted and by what degree can inform the researcher about which sample abundances are not directly comparable, thus reducing the possibility of this methodological difference causing misinterpretation of differences among samples (e.g. treatment effects). Some sequencing facilities (e.g. University of Minnesota Genomic Center, Minnesota, USA) will perform qPCR on extracted DNA to obtain a copy number estimate; and then, once sequencing is performed, flag and report any samples where the *target* copy number is less than the intended sequencing *depth* so that researchers are aware of low copy numbers when interpreting the data. Although many sequencing facilities do not automatically provide the information, sample dilution information can and should be requested from the sequencing facility and incorporated into analyses and interpretation of microbial community matrices. More sophisticated statistical methods also have been developed to help account for different read *depths* (McMurdie & Holmes, 2014).

### 5.2 | Include technical replicates across sequencing lanes

Dividing samples across multiple sequencing lanes increases sequencing *depth* but introduces another source of cryptic variation, lane-to-lane variation. This among-lane source of variation introduces differences in estimates of species presence and relative abundance, particularly for species with moderate to low abundances. The best option and our recommendation to reduce lane-to-lane variation is to place all samples that will be compared on a single lane, when possible. However, while ideal, this is unrealistic for many biologists interested in building a well-sampled community structure with hundreds of samples. If a single lane is not possible, an alternative recommendation is to include technical replicates in each lane to compare species presence and abundance across lanes. Technical replicates (i.e. subsamples drawn from the same biological sample and analysed on all lanes) can reveal bias among lanes and underestimation of diversity in the community (Song et al., 2018). In addition, treating lanes as a random effect when analysing the data (e.g. in mixed-effects models, Lindstrom & Bates, 1988) will reduce bias and account for some of the variation that occurs among lanes.



### 5.3 | Understand your unique community structure

Understanding how sampling intensity and community abundance distribution patterns shape the measurement of community richness, evenness and diversity is critical for making ecological inference using MGAS data. Using simulations, we were able to determine that measures of alpha diversity that incorporate evenness are relatively unaffected by these sources of variation. Community distances and richness estimates, however, are likely to be poorer descriptions of the true community because at least half the community needs to be sampled before reaching a point that represents the entire community. The approaches used to simulate these data are familiar in most community ecologists' toolkits and provide examples to emphasize how inference about communities changes based on sampling. However, we encourage the reader to thoroughly screen the literature for research that addresses approaches to analysing compositional data (e.g. Gloor et al., 2017) when analysing MGAS data.

## 6 | CONCLUSION

Marker gene amplicon sequencing (MGAS) is a powerful tool that has opened a new level of scientific exploration accessible in ecology. However, identifying underlying sources of bias is crucial to interpreting and using these data to truly advance knowledge. Here, we shine a light on cryptic sources of variation in MGAS and explain how each source can affect the community profile. While these sources of variation are mostly unavoidable for informed researchers, they are critical to understand for effective use of this powerful tool to answer ecological questions. Awareness of MGAS library preparation and sequencing is key to generating a strong dataset and building a robust community matrix that will help to uncover true ecological patterns and responses in microbial data.

### ACKNOWLEDGEMENTS

Financial support was provided through National Science Foundation Award: DEB1241895 to E.T.B. We would like to thank Daryl M. Gohl for his incredibly insightful and technical comments on the manuscript.

### CONFLICT OF INTEREST

We are unaware of any conflicts of interest associated with this publication.

### AUTHORS' CONTRIBUTIONS

A.P.-K., E.T.B. and E.W.S. conceived the paper; A.P.-K. wrote the paper and all authors contributed to revisions; A.T.S. and J.A.H. built the simulations.

### PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1111/2041-210X.13764>.

### DATA AVAILABILITY STATEMENT

Data deposited in the Dryad Digital Repository <https://doi.org/10.5061/dryad.fxprvx0r3> (Porath-Krause et al., 2021).

### ORCID

Anita Porath-Krause  <https://orcid.org/0000-0002-9119-2372>

Alexander T. Strauss  <https://orcid.org/0000-0003-0633-8443>

Jeremiah A. Henning  <https://orcid.org/0000-0002-2214-4895>

Eric W. Seabloom  <https://orcid.org/0000-0001-6780-9259>

Elizabeth T. Borer  <https://orcid.org/0000-0003-2259-5853>

### REFERENCES

- Abel, H. J., & Duncavage, E. J. (2013). Detection of structural DNA variation from next generation sequencing data: A review of informatic approaches. *Cancer Genetics*, 206(12), 432–440. <https://doi.org/10.1016/j.cancergen.2013.11.002>
- Albright, M. B. N., & Martiny, J. B. H. (2018). Dispersal alters bacterial diversity and composition in a natural community. *ISME Journal*, 12(1), 296–299. <https://doi.org/10.1038/ismej.2017.161>
- Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M. C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H., Kazou, M., Kinkel, L., Lange, L., Lima, N., Loy, A., Macklin, J. A., Maguin, E., Mauchline, T., McClure, R., ... Schloter, M. (2020). Microbiome definition revisited: Old concepts and new challenges. *Microbiome*, 8(1), 1–22. <https://doi.org/10.1186/s40168-020-00875-0>
- Black, J. S., Salto-tellez, M., Mills, K. I., & Catherwood, M. A. (2015). The impact of next generation sequencing technologies on haematological research – A review. *Pathogenesis*, 2, 9–16. <https://doi.org/10.1016/j.pathog.2015.05.004>
- Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., Mills, D. A., & Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods*, 10(1), 57–59. <https://doi.org/10.1038/nmeth.2276>
- Borer, E. T., Kinkel, L. L., May, G., & Seabloom, E. W. (2013). The world within: Quantifying the determinants and outcomes of a host's microbiome. *Basic and Applied Ecology*, 14(7), 533–539. <https://doi.org/10.1016/j.baae.2013.08.009>
- Botnen, S. S., Davey, M. L., Halvorsen, R., & Kausrud, H. (2018). Sequence clustering threshold has little effect on the recovery of microbial community structure. *Molecular Ecology Resources*, 18(5), 1064–1076. <https://doi.org/10.1111/1755-0998.12894>
- Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of Southern Wisconsin. *Ecological Monographs*, 27(4), 325–349. <https://doi.org/10.2307/1942268>
- Buehler, B., Hogrefe, H. H., Scott, G., Ravi, H., Pabón-Peña, C., O'Brien, S., Formosa, R., & Happe, S. (2010). Rapid quantification of DNA libraries for next-generation sequencing. *Methods*, 50(4), 15–18. <https://doi.org/10.1016/j.jymeth.2010.01.004>
- Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker gene data analysis. *ISME Journal*, 11(12), 2639–2643. <https://doi.org/10.1038/ismej.2017.119>
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., ... Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5), 335–336. <https://doi.org/10.1038/nmeth.f.303>
- Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S. M., Betley, J., Fraser, L., Bauer, M., Gormley, N.,

- Gilbert, J. A., Smith, G., & Knight, R. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME Journal*, 6(8), 1621–1624. <https://doi.org/10.1038/ismej.2012.8>
- Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., & Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America*, 108(Suppl 1), 4516–4522. <https://doi.org/10.1073/pnas.1000080107>
- Castle, S. C., Song, Z., Gohl, D. M., Gutknecht, J. L. M., Rosen, C. J., Sadowsky, M. J., Samac, D. A., & Kinkel, L. L. (2018). DNA template dilution impacts amplicon sequencing-based estimates of soil fungal diversity. *Phytophysics Journal*, 2(2), 100–107. <https://doi.org/10.1094/PBIOMES-09-17-0037-R>
- Chao, A., & Jost, L. (2012). Coverage-based rarefaction and extrapolation: Standardizing samples by completeness rather than size. *Ecology*, 93(12), 2533–2547. <https://doi.org/10.1890/11-1952.1>
- Chase, J. M., & Knight, T. M. (2013). Scale-dependent effect sizes of ecological drivers on biodiversity: Why standardised sampling is not enough. *Ecology Letters*, 16, 17–26. <https://doi.org/10.1111/ele.12112>
- Christian, N., Whitaker, B. K., & Clay, K. (2015). Microbiomes: Unifying animal and plant systems through the lens of community ecology theory. *Frontiers in Microbiology*, 6(SEP), 1–15. <https://doi.org/10.3389/fmicb.2015.00869>
- Claesson, M. J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J. R., Ross, R. P., & O'Toole, P. W. (2010). Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research*, 38(22). <https://doi.org/10.1093/nar/gkq873>
- Croucher, N. J., Fookes, M. C., Perkins, T. T., Turner, D. J., Marguerat, S. B., Keane, T., Quail, M. A., He, M., Assefa, S., Bahler, J., Kingsley, R. A., Parkhill, J., Bentley, S. D., Dougan, G., & Thomson, N. R. (2009). A simple method for directional transcriptome sequencing using illumina technology. *Nucleic Acids Research*, 37(22). <https://doi.org/10.1093/nar/gkp811>
- D'Amore, R., Ijaz, U. Z., Schirmer, M., Kenny, J. G., Gregory, R., Darby, A. C., Shakya, M., Podar, M., Quince, C., & Hall, N. (2016). A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics*, 17(1). <https://doi.org/10.1186/s12864-015-2194-9>
- Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12(7), 499–510. <https://doi.org/10.1038/nrg3012>
- Derocles, S. A. P., Bohan, D. A., Dumbrell, A. J., Kitson, J. J. N., Massol, F., Pauvert, C., Plantegenest, M., Vacher, C., & Evans, D. M. (2018). Biomonitoring for the 21st century: Integrating next-generation sequencing into ecological network analysis. *Advances in Ecological Research*, 58, 1–62. <https://doi.org/10.1016/bs.aecr.2017.12.001>
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, 6(5), 1–10. <https://doi.org/10.1371/journal.pone.0019379>
- Fierer, N., & Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America*, 103(3), 626–631. <https://doi.org/10.1073/pnas.0507535103>
- Gibbons, J. G., Janson, E. M., Hittinger, C. T., Johnston, M., Abbot, P., & Rokas, A. (2009). Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. *Molecular Biology and Evolution*, 26(12), 2731–2744. <https://doi.org/10.1093/molbev/msp188>
- Glassman, S. I., & Martiny, J. B. H. (2018). Broad-scale ecological patterns are robust to use of exact. *mSphere*, 3(4), e00148–e218. <https://doi.org/10.1128/mSphere.00148-18>
- Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, 8(NOV), 1–6. <https://doi.org/10.3389/fmicb.2017.02224>
- Gohl, D. M. (2017). The ecological landscape of microbiome science. *Nature Biotechnology*, 35(11), 1047–1049. <https://doi.org/10.1038/nbt.3983>
- Gohl, D. M., Magli, A., Garbe, J., Becker, A., Johnson, D. M., Anderson, S., Auch, B., Billstein, B., Froehling, E., McDevitt, S. L., & Beckman, K. B. (2019). Measuring sequencer size bias using REcount: A novel method for highly accurate Illumina sequencing-based quantification. *Genome Biology*, 20(1), 1–17. <https://doi.org/10.1186/s13059-019-1691-6>
- Gohl, D. M., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A., Gould, T. J., Clayton, J. B., Johnson, T. J., Hunter, R., Knights, D., & Beckman, K. B. (2016). Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nature Biotechnology*, 34(9), 942–949. <https://doi.org/10.1038/nbt.3601>
- Gołębiewski, M., & Tretyn, A. (2019). Generating amplicon reads for microbial community assessment with next-generation sequencing. *Journal of Applied Microbiology*, 128(2), 330–354. <https://doi.org/10.1111/jam.14380>
- Gonzalez, A., King, A., Robeson, M. S., Song, S., Shade, A., Metcalf, J. L., & Knight, R. (2012). Characterizing microbial communities through space and time. *Current Opinion in Biotechnology*, 23(3), 431–436. <https://doi.org/10.1016/j.copbio.2011.11.017>
- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351. <https://doi.org/10.1038/nrg.2016.49>
- Hugerth, L. W., & Andersson, A. F. (2017). Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing. *Frontiers in Microbiology*, 8(SEP), 1–22. <https://doi.org/10.3389/fmicb.2017.01561>
- Huse, S. M., Welch, D. M., Morrison, H. G., & Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology*, 12(7), 1889–1898. <https://doi.org/10.1111/j.1462-2920.2010.02193.x>
- Hutchinson, G. E. (1961). The paradox of the plankton. *The American Naturalist*, 95(882), 137–145. <https://doi.org/10.1086/282171>
- Illumina. (2016). *Optimizing cluster density on Illumina sequencing systems table of contents* (pp. 1–12).
- Illumina. (2017). *An introduction to next-generation sequencing technology* (pp. 1–16). Retrieved from <https://www.illumina.com/science/technology/next-generation-sequencing.html>
- Jaccard, P. (1901). Distribution comparée de la flore alpine dans quelques régions des Alpes occidentales et orientales. *Bulletin de La Société Vaudoise Des Sciences Naturelles*, 37, 241–272.
- Jousselin, E., Clamens, A.-L., Galan, M., Bernard, M., Maman, S., Gschloessl, B., Duport, G., Meseguer, A. S., Calevo, F., & Coeur d'acier, A. (2016). Assessment of a 16S rRNA amplicon Illumina sequencing procedure for studying the microbiome of a symbiont-rich aphid genus. *Molecular Ecology Resources*, 16(3), 628–640. <https://doi.org/10.1111/1755-0998.12478>
- Kelly, R. P., Shelton, A. O., & Gallego, R. (2019). Understanding PCR processes to draw meaningful conclusions from environmental DNA studies. *Scientific Reports*, 9(1), 1–14. <https://doi.org/10.1038/s41598-019-48546-x>
- Lande, R. (1996). Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos*, 76(1), 5–13. Retrieved from <https://www.jstor.org/stable/3545743>
- Laurie, M. T., Bertout, J. A., Taylor, S. D., Burton, J. N., Shendure, J. A., & Bielas, J. H. (2013). Simultaneous digital quantification and

- fluorescence-based size characterization of massively parallel sequencing libraries. *BioTechniques*, 55(2), 61–67. <https://doi.org/10.2144/000114063>
- Lekberg, Y., Gibbons, S. M., & Rosendahl, S. (2014). Will different OTU delineation methods change interpretation of arbuscular mycorrhizal fungal community patterns? *New Phytologist*, 202(4), 1101–1104. <https://doi.org/10.1111/nph.12758>
- Levy, S. E., & Myers, R. M. (2016). Advancements in next-generation sequencing. *Annual Review of Genomics and Human Genetics*, 17, 95–115. <https://doi.org/10.1146/annurev-genom-083115-022413>
- Lindahl, B. D., Nilsson, R. H., Tedersoo, L., Abarenkov, K., Carlsen, T., Kjoller, R., Kõljalg, U., Pennanen, T., Rosendahl, S., Stenlid, J., & Kauserud, H. (2013). Fungal community analysis by high-throughput sequencing of amplified markers – A user's guide. *New Phytologist*, 199(1), 288–299. <https://doi.org/10.1111/nph.12243>
- Lindstrom, M. L., & Bates, D. M. (1988). Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404), 1014–1021. Retrieved from <https://doi.org/10.1080/01621459.1988.10478693>
- Longino, J. T., Coddington, J., & Colwell, R. K. (2002). The ant fauna of a tropical rain forest: Estimating species richness three different ways. *Ecology*, 83(3), 689–702. [https://doi.org/10.1890/0012-9658\(2002\)083\[0689:TAF0AT\]2.0.CO;2](https://doi.org/10.1890/0012-9658(2002)083[0689:TAF0AT]2.0.CO;2)
- Lu, C., & Tian, H. (2017). Global nitrogen and phosphorus fertilizer use for agriculture production in the past half century: Shifted hot spots and nutrient imbalance. *Earth System Science Data*, 9(1), 181–192. <https://doi.org/10.5194/essd-9-181-2017>
- Ma, X., Shao, Y., Tian, L., Flasch, D. A., Mulder, H. L., Edmonson, M. N., Liu, Y. U., Chen, X., Newman, S., Nakitandwe, J., Li, Y., Li, B., Shen, S., Wang, Z., Shurtleff, S., Robison, L. L., Levy, S., Easton, J., & Zhang, J. (2019). Analysis of error profiles in deep next-generation sequencing data. *Genome Biology*, 20(1), 1–15. <https://doi.org/10.1186/s13059-019-1659-6>
- Manley, L. J., Ma, D., & Levine, S. S. (2016). Monitoring error rates in Illumina sequencing. *Journal of Biomolecular Techniques*, 27(4), 125–128. <https://doi.org/10.7171/jbt.16-2704-002>
- Mao, C. X., Colwell, R. K., & Chang, J. (2005). Estimating the species accumulation curve using mixtures. *Biometrics*, 61(2), 433–441. <https://doi.org/10.1111/j.1541-0420.2005.00316.x>
- Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9, 387–402. <https://doi.org/10.1146/annurev.genom.9.081307.164359>
- Margulis, L. (1981). *Symbiosis in cell evolution: Life and its environment on the early earth*. Freeman.
- Martínez-Porchas, M., Villalpando-Canchola, E., & Vargas-Albores, F. (2016). Significant loss of sensitivity and specificity in the taxonomic classification occurs when short 16S rRNA gene sequences are used. *Heliyon*, 2(9). <https://doi.org/10.1016/j.heliyon.2016.e00170>
- McLaren, M. R., Willis, A. D., & Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. *eLife*, 8, 1–31. <https://doi.org/10.7554/eLife.46923>
- McMurdie, P. J., & Holmes, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, 10(4). <https://doi.org/10.1371/journal.pcbi.1003531>
- Medinger, R., Nolte, V., Pandey, R. V., Jost, S., Ottenwälder, B., Schlötterer, C., & Boenigk, J. (2010). Diversity in a hidden world: Potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Molecular Ecology*, 19(Suppl 1), 32–40. <https://doi.org/10.1111/j.1365-294X.2009.04478.x>
- Morgan, M. J., Chariton, A. A., Hartley, D. M., Court, L. N., & Hardy, C. M. (2013). Improved inference of taxonomic richness from environmental DNA. *PLoS ONE*, 8(8). <https://doi.org/10.1371/journal.pone.0071974>
- Morozova, O., & Marra, M. A. (2008). Genomics applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92, 255–264. <https://doi.org/10.1016/j.ygeno.2008.07.001>
- Nakagawa, T., Tsuchiya, Y., Ueda, S., Fukui, M., & Takahashi, R. (2018). Eelgrass sediment microbiome as a nitrous oxide sink in Brackish Lake Akkeshi, Japan. *Microbes and Environments*, 1–10. <https://doi.org/10.1264/jsme2.ME18103>
- Navas-Molina, J. A., Peralta-Sánchez, J. M., González, A., McMurdie, P. J., Vázquez-Baeza, Y., Xu, Z., Ursell, L. K., Lauber, C., Zhou, H., Song, S. J., Huntley, J., Ackermann, G. L., Berg-Lyons, D., Holmes, S., Caporaso, J. G., & Knight, R. (2013). Advancing our understanding of the human microbiome using QIIME. *Methods in Enzymology*, 531. <https://doi.org/10.1016/B978-0-12-407863-5.00019-8>
- Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Simpson, G. L., Solymos, P. M., Stevens, M. H. H., & Wagner, H. (2008). *Vegan: Community Ecology Package*. R Package Version 2.2-0. Retrieved from <http://CRAN.Rproject.org/package=vegan>
- Pannaraj, P. S., Li, F., Cerini, C., Bender, J. M., Yang, S., Rollie, A., Adisetiyo, H., Zabih, S., Lincez, P. J., Bittinger, K., Bailey, A., Bushman, F. D., Sleasman, J. W., & Aldrovandi, G. M. (2017). Association between breast milk bacterial communities and establishment and development of the infant gut microbiome. *JAMA Pediatrics*, 171(7), 647–654. <https://doi.org/10.1001/jamapediatrics.2017.0378>
- Pielou, E. C. (1984). *The interpretation of ecological data: A primer on classification and ordination*. John Wiley & Sons.
- Porath-Krause, A., Strauss, A. T., Henning, J. A., Seabloom, E. W., & Borer, E. T. (2021). Data from: Pitfalls and pointers: An accessible guide to marker gene amplicon sequencing in ecological applications. *Dryad Digital Repository*, <https://doi.org/10.5061/dryad.fxpnvx0r3>
- Powell, J. R., Monaghan, M. T., Öpik, M., & Rillig, M. C. (2011). Evolutionary criteria outperform operational approaches in producing ecologically relevant fungal species inventories. *Molecular Ecology*, 20(3), 655–666. <https://doi.org/10.1111/j.1365-294X.2010.04964.x>
- Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H., & Turner, D. J. (2008). A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, 5(12), 1005–1010. <https://doi.org/10.1038/nmeth.1270>
- R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Retrieved from <http://www.r-project.org/>
- Ribes, M., Calvo, E., Movilla, J., Logares, R., Coma, R., & Pelejero, C. (2016). Restructuring of the sponge microbiome favors tolerance to ocean acidification. *Environmental Microbiology Reports*, 8(4), 536–544. <https://doi.org/10.1111/1758-2229.12430>
- Rideout, J. R., He, Y., Navas-Molina, J. A., Walters, W. A., Ursell, L. K., Gibbons, S. M., Chase, J., McDonald, D., Gonzalez, A., Robbins-Pianka, A., Clemente, J. C., Gilbert, J. A., Huse, S. M., Zhou, H.-W., Knight, R., & Caporaso, J. G. (2014). Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ*, 2014(1), 1–25. <https://doi.org/10.7717/peerj.545>
- Robin, J. D., Ludlow, A. T., La Ranger, R., Wright, W. E., & Shay, J. W. (2016). Comparison of DNA quantification methods for next generation sequencing. *Scientific Reports*, 6, 1–10. <https://doi.org/10.1038/srep24067>
- Roy, S., Coldren, C., Karunamurthy, A., Kip, N. S., Klee, E. W., Lincoln, S. E., Leon, A., Pullambhatla, M., Temple-Smolkin, R. L., Voelkerding, K. V., Wang, C., & Carter, A. B. (2018). Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *Journal of Molecular Diagnostics*, 20(1), 4–27. <https://doi.org/10.1016/j.jmoldx.2017.11.003>

- Saleem, M., Hu, J., & Jousset, A. (2019). More than the sum of its parts: Microbiome biodiversity as a driver of plant growth and soil health. *Annual Review of Ecology, Evolution, and Systematics*, 50(1), 145–168. <https://doi.org/10.1146/annurev-ecolsys-110617-062605>
- Sanger, F., Nicklen, S., & Coulson, A. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), 5463–5467.
- Sanz, J. L., & Köchling, T. (2019). Next-generation sequencing and waste/wastewater treatment: A comprehensive overview. *Reviews in Environmental Science and Bio/Technology*, 18(4), 635–680. <https://doi.org/10.1007/s11157-019-09513-0>
- Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., & Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, 43(6). <https://doi.org/10.1093/nar/gku1341>
- Schloss, P. D. (2016). Application of a database-independent approach to assess the quality of operational taxonomic unit picking methods. *mSystems*, 1(2), 2–5. <https://doi.org/10.1128/msystems.00027-16>
- Schloss, P. D., & Handelsman, J. (2006). Toward a census of bacteria in soil. *PLoS Computational Biology*, 2(7), 0786–0793. <https://doi.org/10.1371/journal.pcbi.0020092>
- Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, 5(1), 16–19. <https://doi.org/10.1038/NMETH1156>
- Seabloom, E. W., Condon, B., Kinkel, L., Komatsu, K. J., Lumibao, C. Y., May, G., McCulley, R. L., & Borer, E. T. (2019). Effects of nutrient supply, herbivory, and host community on fungal endophyte diversity. *Ecology*, 100(9). <https://doi.org/10.1002/ecy.2758>
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423. [https://doi.org/10.1016/s0016-0032\(23\)90506-5](https://doi.org/10.1016/s0016-0032(23)90506-5)
- Shokralla, S., Spall, J. L., Gibson, J. F., & Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, 21(8), 1794–1805. <https://doi.org/10.1111/j.1365-294X.2012.05538.x>
- Simpson, E. H. (1949). Measurement of diversity. *Nature*, 163, 688. <https://doi.org/10.1038/163688a0>
- Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2), 121–132. <https://doi.org/10.1038/nrg3642>
- Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of next-generation sequencing technologies. *Current Protocols in Molecular Biology*, 122(1), 1–11. <https://doi.org/10.1002/cpmb.59>
- Smith, D. P., & Peay, K. G. (2014). Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. *PLoS ONE*, 9(2). <https://doi.org/10.1371/journal.pone.0090234>
- Song, Z., Schlatter, D., Gohl, D. M., & Kinkel, L. L. (2018). Run-to-run sequencing variation can introduce taxon-specific bias in the evaluation of fungal microbiomes. *Phytobiomes Journal*, 2(3), 165–170. <https://doi.org/10.1094/PBIOMES-09-17-0041-R>
- Stackebrandt, E., & Goebel, B. M. (1994). Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic Bacteriology*, 44(4), 846–849.
- Taylor, D. R., Zeyl, C., & Cooke, E. (2002). Conflicting levels of selection in the accumulation of mitochondrial defects in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America*, 99(6), 3690–3694.
- Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature*, 449(7164), 804–810. <https://doi.org/10.1038/nature06244>
- Turner, T. R., James, E. K., & Poole, P. S. (2013). The plant microbiome. *Genome Biology*, 14(6), 1–10. <https://doi.org/10.1186/gb-2013-14-6-209>
- Van Der Heijden, M. G. A., Bardgett, R. D., & Van Straalen, N. M. (2008). The unseen majority: Soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecology Letters*, 11(3), 296–310. <https://doi.org/10.1111/j.1461-0248.2007.01139.x>
- van Dijk, E. L., Auger, H., Jaszczyszyn, J., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics: TIG*, 30(9), 418–426. <https://doi.org/10.1016/j.tig.2014.07.001>
- Vellend, M. (2010). Conceptual synthesis in community ecology. *Quarterly Review of Biology*, 85(2), 183–206. <https://doi.org/10.1086/652373>
- Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., & Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, 6(1), 100. <https://doi.org/10.12688/f1000research.10571.1>
- Wirth, R., Kovács, E., Maráti, G., Bagi, Z., Rákhely, G., & Kovács, K. L. (2012). Characterization of a biogas-producing microbial community by short-read next generation DNA sequencing. *Biotechnology for Biofuels*, 5, 1–16. <https://doi.org/10.1186/1754-6834-5-41>
- Yang, B., Wang, Y., & Qian, P. Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics*, 17(1), 1–8. <https://doi.org/10.1186/s12859-016-0992-y>
- Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic classification. *Cell*, 178(4), 779–794. <https://doi.org/10.1016/j.cell.2019.07.010>
- Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, 29(22), 2869–2876. <https://doi.org/10.1093/bioinformatics/btt499>

**How to cite this article:** Porath-Krause, A., Strauss, A. T., Henning, J. A., Seabloom, E. W., & Borer, E. T. (2022). Pitfalls and pointers: An accessible guide to marker gene amplicon sequencing in ecological applications. *Methods in Ecology and Evolution*, 13, 266–277. <https://doi.org/10.1111/2041-210X.13764>